Implementing precision medicine with the Liver Cancer Collaborative

A blueprint for delivery using the Aridhia DRE

Authors: Winteringham L, Donellan A, Russell S, Carter K

Executive Summary

Current State and Vision

Liver cancer remains one of the most challenging cancers worldwide, with a high mortality rate due to late diagnosis and limited treatment options. The Liver Cancer Collaborative (LCC)[1] is committed to transforming liver cancer research and treatment through a clinician-led biorepository and an integrated digital research environment (DRE). By harnessing cutting-edge technologies and comprehensive data integration, the LCC aims to accelerate the development of new treatments and improve patient outcomes globally.

Key Objectives and Impact

The DRE will enable the LCC to:

- Integrate clinical, genomic, and research data to facilitate precision medicine.
- Provide a secure platform for data sharing and collaboration.
- Enhance research capabilities through innovative technologies such as multi-OMICS and patient-derived models.
- Ultimately improve liver cancer treatment outcomes and patient care.

Introduction

The Liver Cancer Collaborative (LCC) includes clinicians and researchers from multiple institutions working together to address the increasing global burden of liver cancer. The goal of the Collaborative is to deliver a clinician-led liver cancer biorepository with integrated clinical, genomic and research data to facilitate discovery research and accelerate the development of new treatments for liver cancer. The LCC has established an innovative research pipeline, combining coordinated and centralised tissue/serum biobanking, integrated, cutting edge “multi-OMICS” technologies, as well as patient-derived organoid (PDO) and patient derived xenograft (PDX) screening platforms. The liver cancer biorepository is contained within the Perkins Cancer Biobank, leveraging and building on an established structure that includes approved ethics and governance, an experienced team of clinical and laboratory staff, and a comprehensive laboratory management system.

A significant number of datasets have been produced including whole exome sequencing (231), bulk RNA sequencing (180), single-cell or nucleus RNA sequencing (123) and spatial transcriptomics (14). Medium throughput drug screening data, on a subset of PDOs, is also underway. The LCC are continuing to expand the data types, currently undertaking methylome and proteome analysis on

[1] https://www.livercancercollaborative.au/

selected participants as well as developing protocols for the inclusion of radiomics. In addition to research data detailed clinical information is being collated by the LCC clinical team.

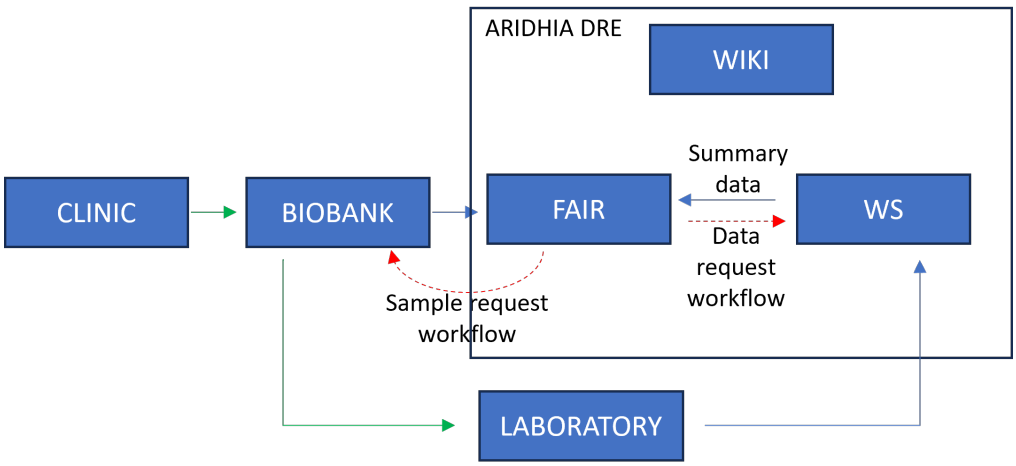
Despite considerable effort, applying precision medicine in a clinical setting remains an elusive goal. The extremely siloed nature of clinical and research data is a major challenge that needs to be addressed before progress can be made. This paper outlines the architecture and use-cases of the secure online Digital Research Environment, developed in conjunction with Aridhia Informatics. This platform contains both clinical and research data for a cohort of nearly 300 liver cancer patients including return visits. These data will enable the development a precision medicine approach to treating liver cancer.

Background

Liver cancer is the 3rd leading cause of cancer deaths worldwide[2]. This high mortality rate is largely due to the late diagnosis of liver cancer as well as disease recurrence and metastasis. Patients diagnosed with early-stage liver cancer may undergo a partial hepatectomy or other surgical procedures, however, for patients diagnosed with late-stage disease there are less options. Moreover, the lack of access to intermediate and late stage biospecimens has significantly impacted the ability to undertake the research needed to develop better treatments for this cohort of patients.

The LCC is a collaboration between research scientists and clinicians across all three tertiary hospitals and multiple research institutions in Perth, Western Australia. Together with Aridhia, the LCC have developed a secure, fully audited, digital environment to facilitate collaboration and innovation. The DRE connects the unique partnership of hospital clinics, cancer biobank and research laboratories to enable cutting-edge research (Figure 1). As the platform develops, researchers have access to a growing collection of research and clinical data that will underpin vital research discoveries as well as enable personalised and precision medicine.

Figure 1 Liver Cancer Collaborative clinical and research environment



There is an urgent need for a redefined, enhanced, and revolutionary approach to make meaningful change to outcomes for liver cancer patients. The DRE offers a practical, and high-impact opportunity for patients and medical professionals to collaboratively harness data and knowledge to identify new treatments and establish a new standard of care. The integration of precision medicine into clinical

[2] 2005-2024 American Society of Clinical Oncology (ASCO). Cancer.Net. Accessed 26/4/24 https://www.cancer.net/

care will empower healthcare providers to make more informed treatment decisions and as the platform and data collection evolve, the quality of care will continually improve. Importantly, we see this innovative approach as a gateway to a ground-breaking liver cancer treatment hub.

Configuration of the LCC DRE

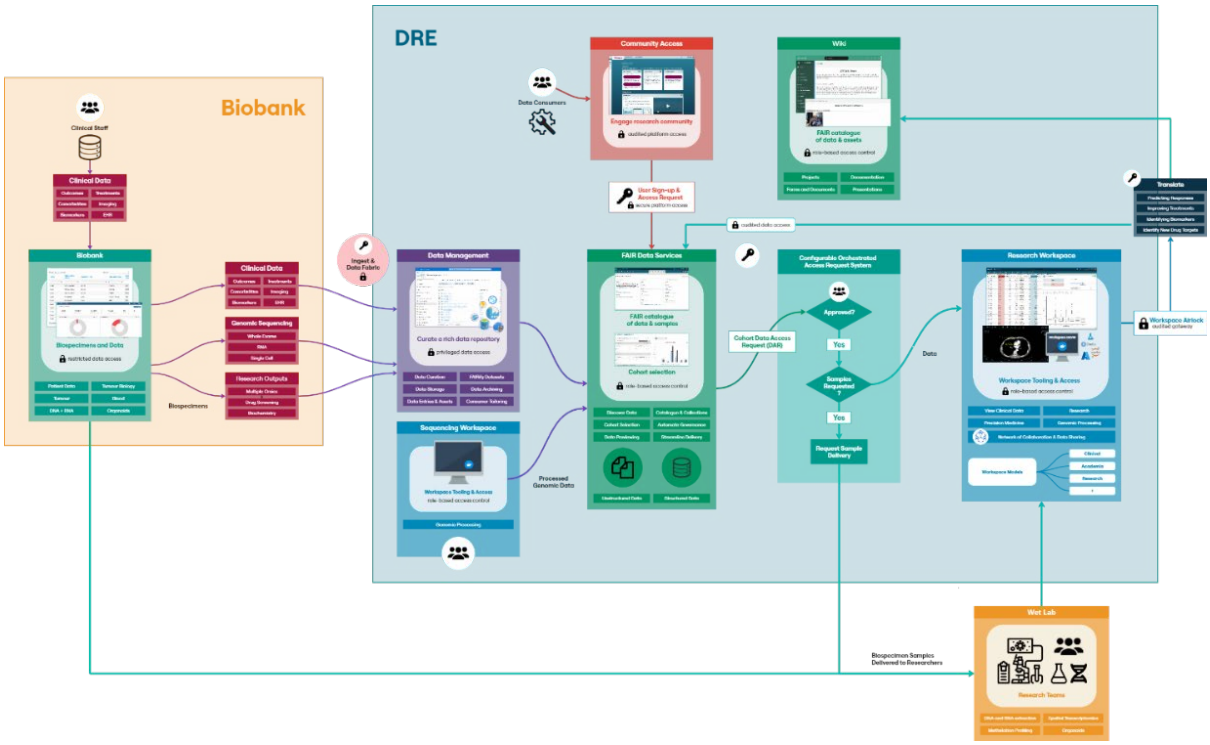
The LCC DRE provides a single 'shop front' for multi-mode data including structured clinical data in an OMOP-like format, 'omics data for each patient in a variety of raw and processed formats, and numerous biospecimen data on each patient. Having a consistent dataset catalogue and metadata dictionary allows researchers and clinicians to explore and understand available data prior to requesting access. However, that data must first be extracted from host systems, transformed into a format compatible with the FAIR Data Services component of the DRE and then loaded into the system. This process is known as 'Extract, Transform, Load' (ETL); a key process in data integration.

The following diagram (Figure 2) provides a high-level architecture of the LCC DRE platform showing the flow of data and interaction from loading data into the DRE from the Biobank, to requesting access with that data, working with the data in a secure environment and then requesting airlock approval of data. The DRE provides a secure area for ETL pipeline processing and data management. Findable, Accessible, Interoperable, and Reusable (FAIR) APIs allow for automation of data upload and refresh. The FAIR Data Services component allows for users to find data of interest, understand if it is relevant to their study and request through an automated, orchestrated data access request (DAR) system. A DAR is either approved or rejected by the LCC approval committee. In some cases, requests will be for a combination of clinical data and bio-samples. Bio-samples will be physically delivered to a laboratory while clinical data will be automatically transferred to a secure, collaborative DRE Workspace. The Workspace comes pre-loaded with in-built tooling to assist with data analysis, including RStudio, Jupyter Notebook and further built in tabular data analysis tools. Each workspace has a dedicated, isolated Postgres Database for SQL interaction. Virtual machines can provide more flexible compute, with users, upon approval, having the ability to bring their own code and further data into a virtual machine. Users of virtual machines benefit from a dedicated Gitea instance for source code control and version management of code, models, data and documentation.

Workspaces are gated by inbound and outbound airlocks. No data can leave a workspace without approval from an appropriate administrator (data owner/governance controller), and all data activity within a workspace is audited in a comprehensive audit file available to administrators and workspace managers.

An internal wiki-based project catalogue is available to all LCC DRE users for sharing information on projects and data and to further encourage collaboration.

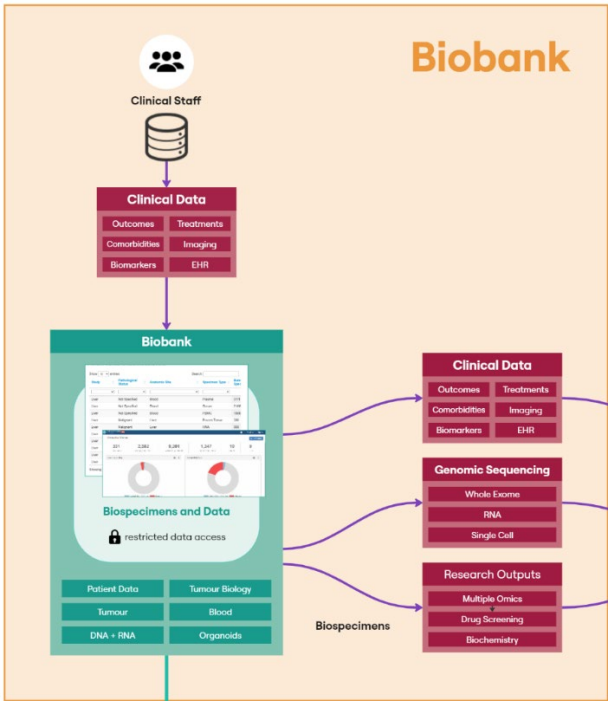
Figure 2 Configuration of the LCC DRE



Technical Implementation

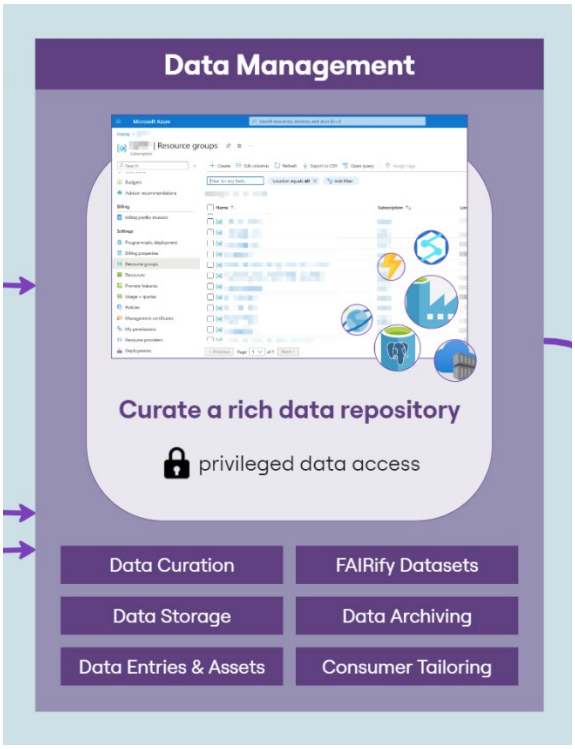
The following sections the specific components of the diagram:

Biobank



The Perkins Cancer Biobank, which underpins the LCC, uses OpenSpecimen as its biobank laboratory information management system to track the collection, storage and utilisation of each patient specimen. Open Specimen sits within the University of Western Australia network and access is restricted to data managers and clinical staff. A publicly available dashboard has been published on the Perkins external website to facilitate sample sharing. Forms configured within OpenSpecimen are used to collate an array of clinical measurements associated with a unique PPID.

OpenSpecimen provides an Application Programming Interface (API) that allows for users with the appropriate credentials to programmatically extract data from the system as part of an ETL pipeline. LCC have defined two ETL processes, one for the specimen data and one for the clinical measurements.

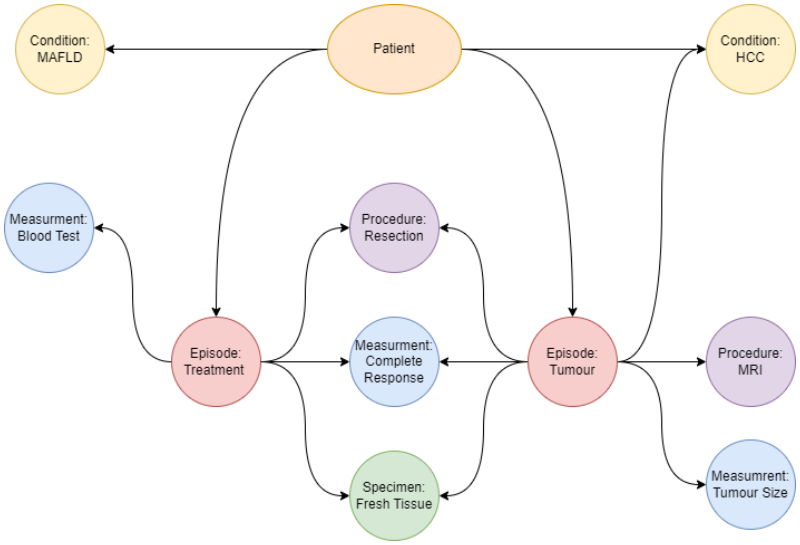


This is not a one-off process as new concepts continue to be added to the OpenSpecimen source, an automated, reliable and recurring pipeline, using the FAIR API to update existing datasets with new entries to lookup tables as new concepts are added.

The specimen dataset contains high-level information on each patient such as treatment and aetiology to help researchers browse specimens available in the Perkins Cancer Biobank. FAIR provides a 'Cohort Builder' tool, that provides the facility to understand the shape of a dataset and whether or not it meets the requirements of a study.

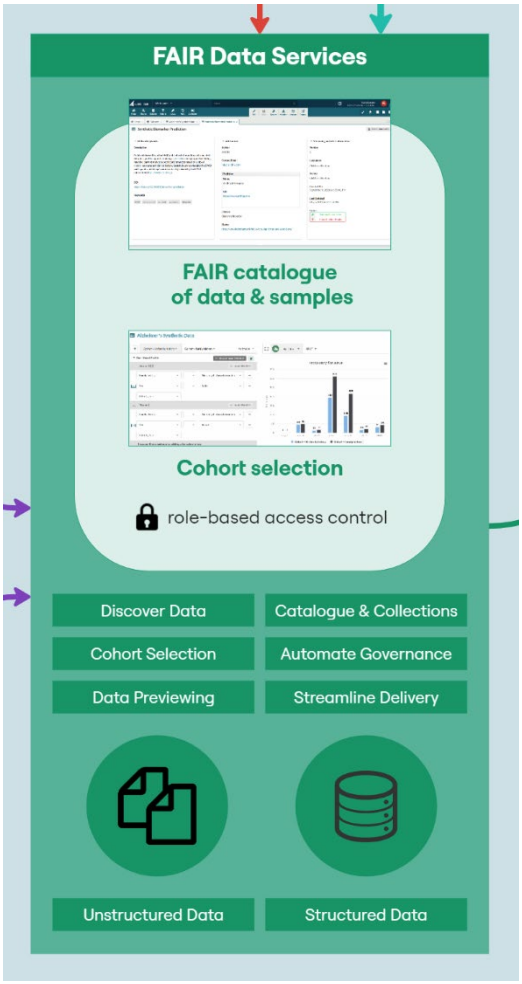
Clinical Data

Figure 3. Clinical data relationships



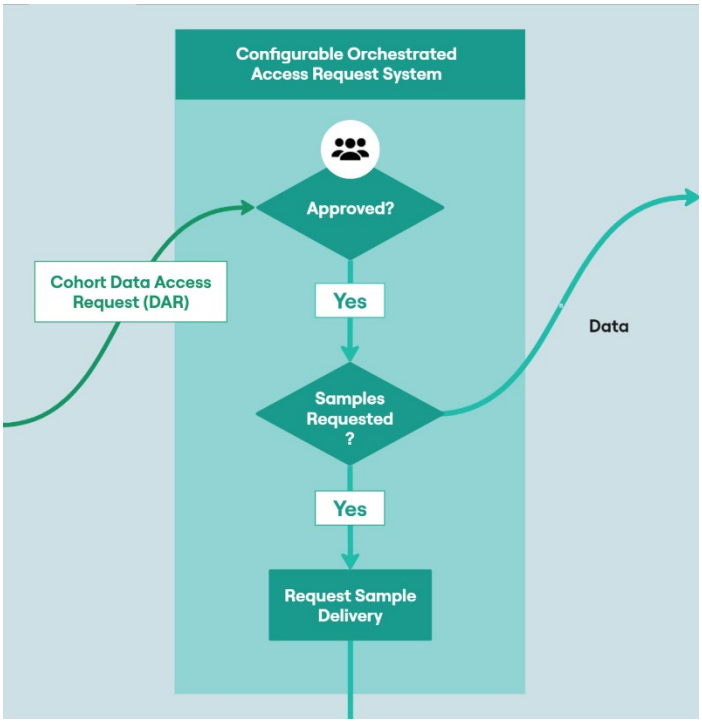
Clinical measurement data captures 40+ datapoints around each patient's cancer journey (Figure 3). The data collected helps us to build a detailed picture of the patient's treatments, responses, comorbidities, biomarkers and details around their cancer and individual tumours. This information provides researchers with a highly detailed background for each sample and a snapshot of the patient when the sample was collected.

The clinical team reviews clinical measurement data in the data collection forms in OpenSpecimen. This data then undergoes an ETL process before being made available in FAIR. The ETL process puts the data into a common set of tables and builds up a model representing the relationships between the data. The data model used for the clinical data is based on the OMOP data model and makes use of their work representing oncology data. Individual records associated with each PPID are stored in the measurement, specimen, observation, condition, drug and procedure tables. These records are aggregated using the episode tables to represent a higher-level view of the different stages of the patients' journey. For example, a treatment episode will be linked to the procedure used to treat the patients as well as any specimens collected from the procedure, blood tests and imaging from before and after, and details about the tumours. Individual tumours are also represented as episodes. Relationships between individual records are also captured. This usually entails associating measurements with what the measurement is from.



Particularly important in this use-case is the identification of a subset of biospecimens for delivery to other wet laboratories. Users can interact with Cohort Builder to identify and request specific samples for their research. The high-level patient data (aetiology and treatment) is also added to other datasets like the RNA and WES sequencing datasets.

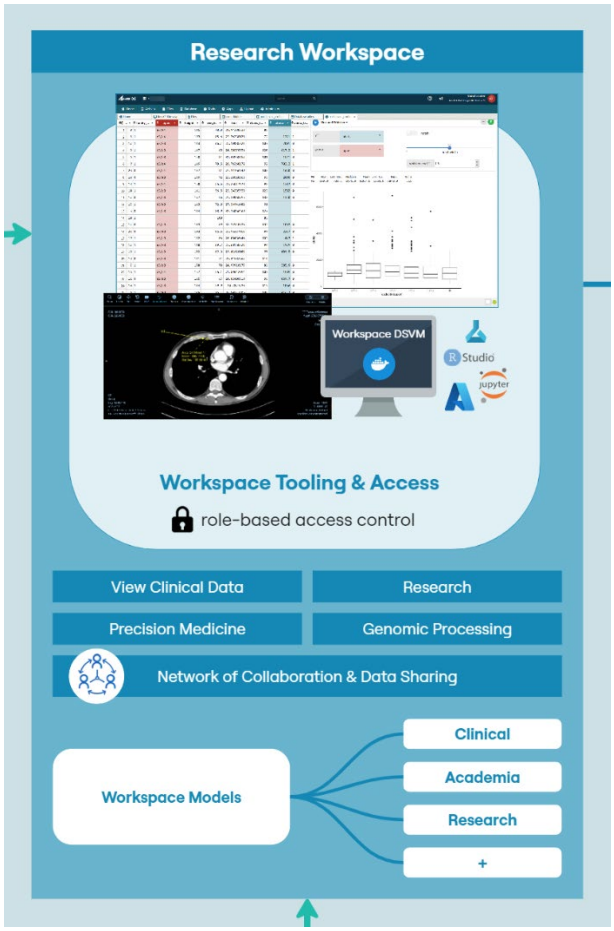
The LCC DRE provides two datasets for requesting specimens. The biospecimen data dataset which has high-level data on all samples is freely available to researchers withing the DRE. Researchers can build a cohort of potential samples using FAIR before refining their selection and identifying individual samples in the workspace. Samples are requested through the biospecimen request dataset. This dataset does not have any data associated with it but has a custom form where researchers can specify the samples requested and how they intend on utilising the samples.



Facilitating data governance best practices

The DRE provides the facility to orchestrate and automate data access requests (DARs). Each DAR is fully configurable, allowing individuals who are not members of the DRE (such as data owners, or members of a review committee) to be involved in the Data Access Request process and approve or deny access. Users can track progress of their access request and are notified upon approval or rejection. Sample requests are reviewed by the LCC committee and if approved, processed through OpenSpecimen. These samples are then physically delivered to the requestor where analysis can take place, along with data analysis of corresponding clinical and research data. Information on the request, such as purpose is tracked in OpenSpecimen, associated with the requested samples and fed back into FAIR as part of the high-level patient data. This positive feedback loop allows researchers to build cohorts based on research that has previously been carried out and enriches the available data.

When requesting data from the FAIR component, the user must have a valid and active workspace. Upon approval of a dataset that includes clinical or genomic data, the data is made accessible from within a workspace. The workspace is a secure, collaborative, web-based research area, providing dedicated out of the box and LCC-specific tooling to facilitate detailed statistical analysis of clinical data and omics data.



The workspace has several components to facilitate secure analysis while providing audit, traceability and control over access (Figure 4). Each workspace must have at least one administrator. That administrator can be a workspace member directly, or an LCC administrator. Only administrators have the facility to approve or reject file egress via an airlock mechanism. This ensures that only those users approved by LCC have the facility to allow data to enter or leave a workspace.

All transfers to the workspace must first pass through an 'Inbox'. The inbox gives an administrator the opportunity to inspect incoming files and also ensure that they have been scanned for viruses and other malware. Only administrators can view and approve files within the inbox. Once incoming files are accepted, they are transferred to a shared file system for use by all researchers.

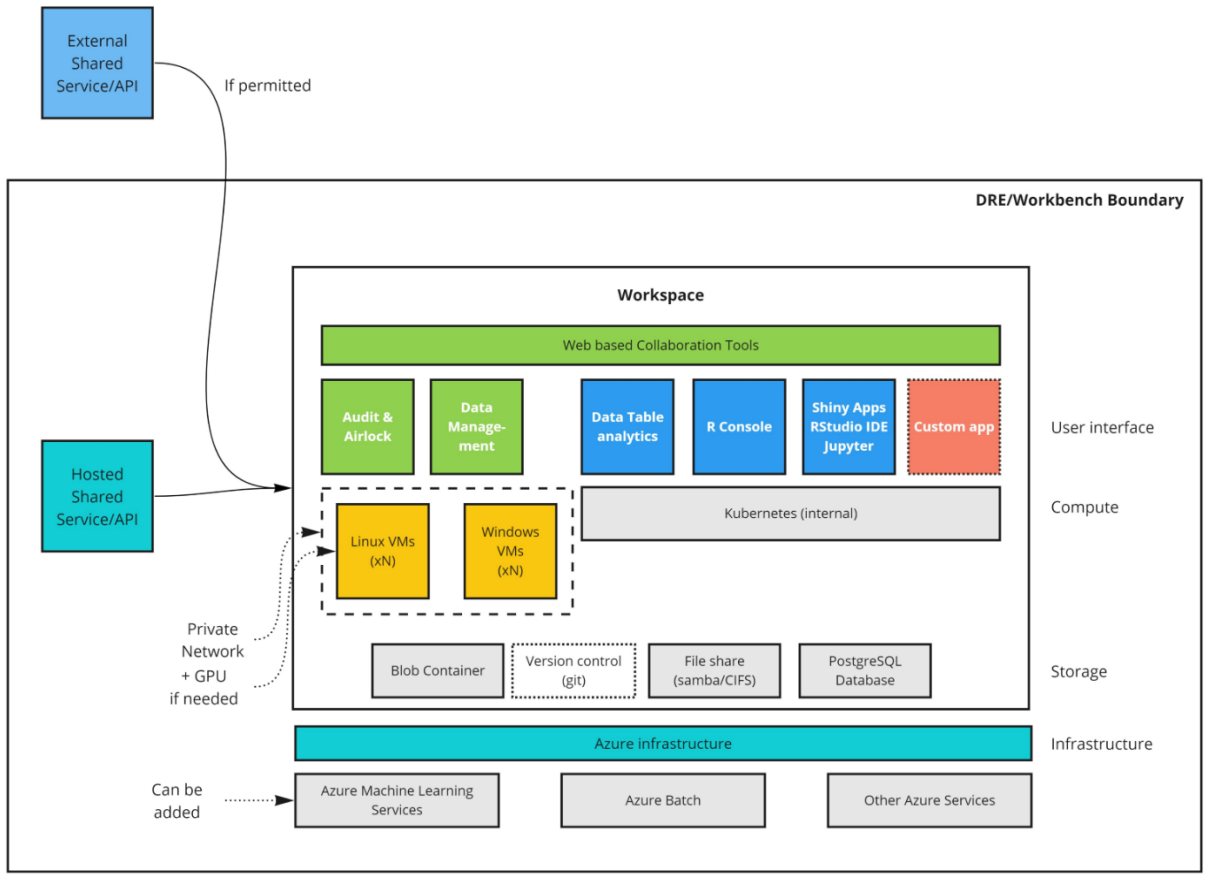
A workspace typically contains a few tools to provide the researcher with the capabilities they require to work with and analyze data securely while collaborating with other users within the same environment.

Common components of a workspace would include:

- A shared file system for structured and unstructured data to allow workspaces members to work on a common set of data and code files
- A relational database that can be used to contain original data or build novel datasets and provide SQL access from workspace tools
- A Kubernetes cluster for hosting customized containerized applications.

- A secure token-based file upload mechanism to allow users to bring their own code and data into the workspace
- An inbound 'airlock' to scan incoming files for viruses and provide a means for administrators to review incoming files prior to general access
- R and Python coding environments to build apps and models
- Data exploration tools such as table sorting, filtering and pre-defined biomedical analysis modules
- Git, to provide the ability to collaborate and version code and data within a workspace
- Access to virtual machine computers, including GPU backed machines for large-scale modelling purposes.

Figure 4. Anatomy of a workspace

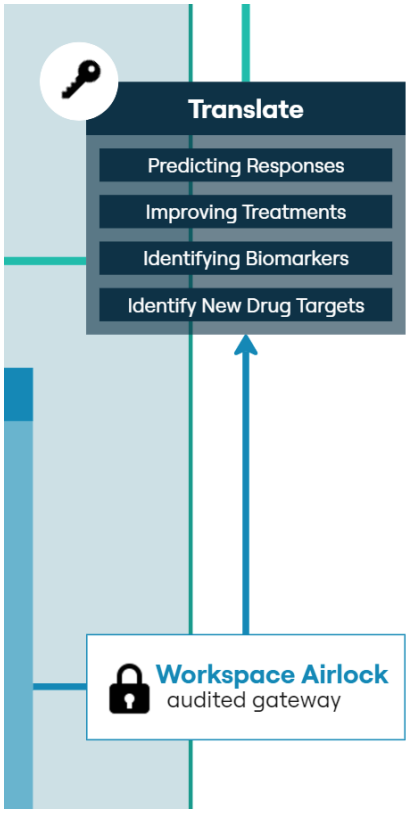


It is common for researchers to bring their own tools to a workspace, but there is a growing requirement to pre-provision workspaces with research and analysis use-case specific tooling. LCC are developing a suite of dedicated tooling tailored to the clinical and sample data to provide insight into patient journeys and treatment pathways. These tools will be available to all users with approved access to the data, allowing for rich visualisation of complex overlapping, longitudinal patient data.

Advanced researchers have access to R and Python coding environments, along with version control capabilities. Using these tools, developers can contribute modelling code or dedicated tooling that can be re-used in other workspaces. Workspaces ensure that teams are self-contained with no possibility of cross pollination of data across projects. Audit trails and version history can chart the progress of development.

Upon completion, models can be exported via a secure airlock process. This airlock ensures that administrators can review content prior to export. The export can be either direct from the cloud to the local machine or to another workspace. A use-case being that these models develop and are versioned during the drug development journey, with each workspace maintaining a versioned and audited digital archive of work and individuals who contributed.

Expected Outcomes



The LCC has established a multi-component pipeline that enables efficient and auditable biobanking, provision of samples for research data generation, and data capture for processing, analysis, sharing and collaboration. The security credentials of Aridhia's DRE allow the integration of detailed clinical data with these research outputs, providing an unprecedented opportunity to identify critical events associated with each response or outcome across a patient's cancer journey.

The DRE is currently supporting multiple research projects across several identified areas of need in liver cancer treatment and management. We are using several approaches to develop multi-omics-based biomarkers to predict clinically relevant events such as response to treatment, disease remission and recurrence. We have programs focused on identifying novel drug targets and biologically important pathways with the end goal to develop new therapeutics. The DRE is also supporting more clinically driven programs including the assessment of new imaging modalities and the establishment of a radiomics pipeline.

Importantly, as research projects progress new data will be added to FAIR becoming available to further inform new research questions insuring a dynamic and constantly developing rich resource for collaborative research. The wiki will facilitate new internal collaborations building upon previous findings and successes. Once published, the data in the FAIR component of the DRE become publicly available contributing to global research programs and accelerating the progress of new treatments and improved patient care models.

Looking ahead, our current infrastructure positions us well to conduct clinical trials for liver cancer. Specifically, we are exploring the prospect of establishing platform trials. A platform trial is generally disease-focused and adaptive offering agility and flexibility, enabling protocols to be quickly amended or halted based on emerging treatments, efficacy, or safety data. The architecture of Aridhia's DRE is particularly suited for managing these trials, providing securely controlled and rigorously audited environments where data, models, code, and results can be shared among contributing centres or trial arms while maintaining each centre's autonomy over its own data.[3]

Additionally, we will be adding and configuring a Federated Node to the DRE later this year to build an international network of liver cancer collaborations that are privacy preserving, ensuring data sovereignty for all network participants. [4]

The overall goal is to achieve better outcomes for people with liver cancer. Liver cancer continues to have a significant global impact; there have been no effective treatments, and the 5-year survival rate remains abysmal. We believe this collaborative and comprehensive approach to generating, analysing and translating multiple integrated datasets will positively impact the lives of those diagnosed with liver cancer.

Summary and Next Steps

We have established a multi-component pipeline for efficient and auditable biobanking, sample provision, and data capture for processing, analysis, and sharing. Using Aridhia's DRE we can integrate detailed clinical data with research outputs, enabling the identification of critical events across a patient's cancer journey.

The DRE supports various liver cancer research projects including discovery of multi-omics-based biomarkers to predict treatment responses, disease remission, and recurrence, identifying novel drug targets and pathways, assessing new imaging modalities and more recently building a radiomics model to enhance routine imaging capability.

As research progresses, new data is added to FAIR, creating a dynamic resource for collaborative research. The wiki facilitates internal collaborations and builds on previous findings. Once published, FAIR data becomes publicly available, contributing to global research and accelerating new treatments and improved patient care.

Next Steps

To advance our mission, securing additional funding is crucial. We aim to attract industry support by showcasing the DRE's capabilities and potential for impactful research. Connecting with global researchers and integrating other datasets through our platform will further enhance our collaborative efforts and drive innovation in liver cancer treatment.

Our infrastructure is well-positioned for conducting clinical trials, including platform trials. These adaptive trials offer flexibility, enabling protocol adjustments based on emerging data. Aridhia's DRE architecture supports these trials by providing secure, audited environments for data sharing while maintaining each centre's autonomy.

Our goal is to achieve better outcomes for liver cancer patients. With no effective treatments and a low 5-year survival rate, our comprehensive approach to integrated data aims to positively impact the

[3] Contemp Clin Trials. 2024 May 5;142:107559, doi: 10.1016/j.cct.2024.107559

[4] How the Aridhia DRE is Enabling Federated Analysis in the PHEMS Consortium | Trusted Data Sharing Network | Digital Research Environment

lives of those diagnosed with liver cancer. By expanding our funding, industry partnerships, and global collaborations, we can accelerate the discovery of new treatments and improve patient care worldwide.

Engagement and Collaboration

We invite researchers, clinicians, and industry partners to join us in the fight against liver cancer. Together, we can harness the power of data and innovation to improve patient outcomes and advance the field of liver cancer research. For more information on how to get involved, please visit our website or contact us directly.