GUARDIANS CCIA ZERO Data Commons – Initial Data Model Report

Contents

1 Purpose

This document outlines the selection and scope of the initial data model for CCIA's ZERO Data Commons project delivered under the GUARDIANS initiative. It summarises the evaluation of potential models, rationale for the chosen approach, and how the model aligns with project requirements, user stories, and open standards.

2 Data model selection

2.1 Models considered

As part of the data modelling process, we reviewed several models to assess their suitability for the project's goals, with a focus on interoperability, alignment with open standards and FAIR principles, and practical implementation needs.

Model considered	Description	Comments
FHIR (data model)	Data exchange and interoperability Optimized for real-time data exchange, enabling interoperability across healthcare delivery environments by supporting the integration of various Electronic Health Record (EHR) systems, mobile apps, and health IT solutions.	• Large community • Comprehensive documentation • Large ecosystem of auxiliary tools and applications • Base model accounts for the most important parts of our internal data model • Extensibility is built in and accounted for
OMOP (data model)	Analytics and research A standardized framework for storing and analysing observational health data, facilitating long-term data preservation, research, and cross-institutional collaboration.	• Smaller than FHIR, but less flexible/extensible • Heavy emphasis on ETL scripts • More complicated deployment procedures

Model considered	Description	Comments
CBioPortal (data model + web portal)	Analytical platform cBioPortal's primary goal is to make complex cancer genomic data accessible and interpretable for cancer biologists and clinicians. We transform multimodal data into interactive visualizations that facilitate biological discovery and clinical decision-making.	• Includes a fully developed frontend application • Data model and API are not GA4GH compliant
Various ontologies	Disease ontologies to map our data to, including: • MONDO • NCIT • WHO	• Early stages, will have a stronger consideration further into the project
Data Repository Service (DRS) (standard/API)	Standardised data retrieval Provides a standardised set of data retrieval methods.	• Provides a GA4GH-compliant API for retrieving file-based data • May be useful in the later stages of GUARDIANS, after successful integration of REMS and ELSA
Variation Representation Specification (VRS) (standard)	Standardised variant annotation Variant representation and federated identification	• May be useful in the second stage of data model validation, when the focus is on the ETL of molecular data

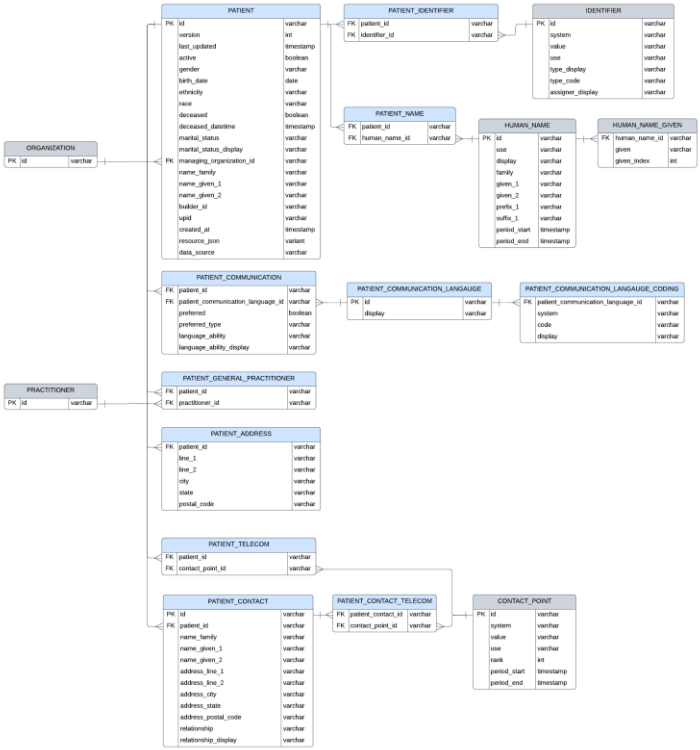
2.2 Rationale for FHIR model selection

All models considered place a heavy emphasis on sophisticated ETL scripts that must be custom-built. FHIR has been chosen as the preliminary data model for the following reasons:

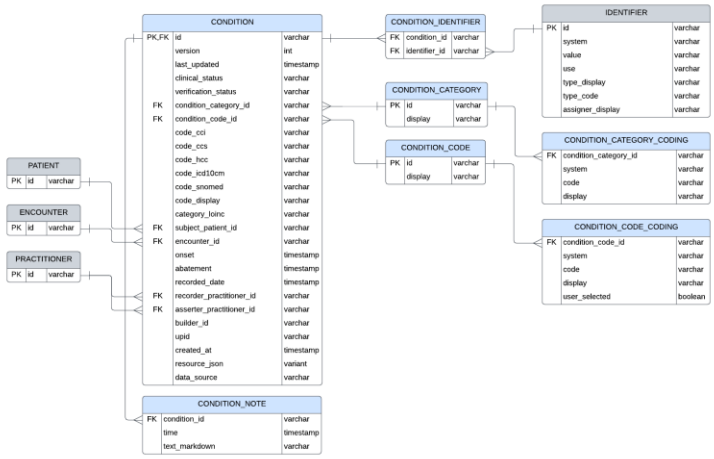
- Mature data model and tooling with an established community
- Flexibility and extensibility are major design considerations
- Most of our core tables can be easily mapped to specific resources

While the FHIR specification details many resource types, preliminary adoption of the data model will involve only the following resources:

- Patient to map our participants

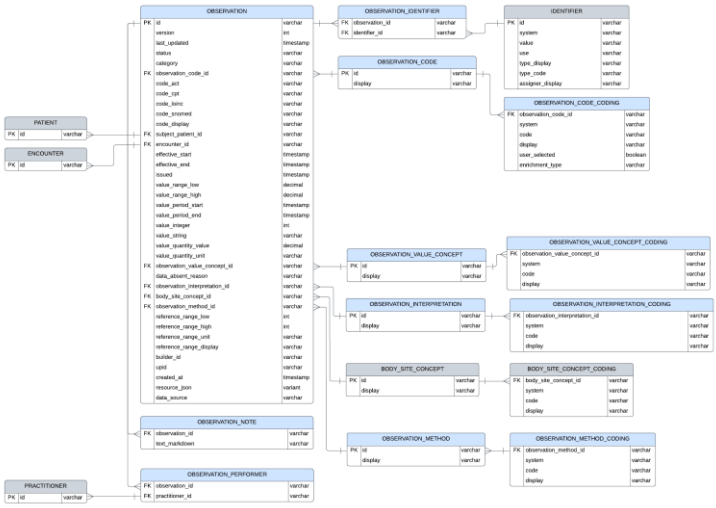


- Condition to map the diagnosis of the participant



- Specimen to map the biosamples of the participant

- Observation to map various metadata, including vital status, age at diagnosis and minimal biomarker information.



These resources will help to address numerous user stories - documented here - by exposing important data points from CCI's internal databases.

In addition to the reasons listed above, FHIR was also chosen for its extensibility – one of its foundational tenets. Throughout our continued evaluation and refinement of the data model, we may utilise the Extension element to accommodate for esoteric information not covered by the base model. FHIR acknowledges the need for extensibility and so has integrated Extension elements into the design of the data model itself. We foresee this to be particularly useful in integrating our detailed molecular data or captured metadata specific to paediatric cancer.

2.3 Risks and implications

While the adoption of FHIR offers strong alignment with open standards and interoperability goals, it also introduces several considerations that will need to be managed as the project progresses.

Risks/implication	Impact	Comments
Data model doesn't reflect every aspect of our internal data model	Low	FHIR was chosen specifically for its extensibility. More specific data points can be accommodated for via the Extension element.
Data model will require many manual customisations	Medium	Given the specificity of paediatric cancer data (on both the metadata and molecular data level), certain generic resources in FHIR will need to be extended several times to fully capture the richness of our data. This can be partially mitigated by streamlining the process of customising resources.
Data model is not performant at large scale	Low	The FHIR database was designed with large datasets in mind; as such, performance has been taken into consideration by the developers of the

Risks/implication	Impact	Comments
		data model and reference implementation.
Complexity of the data model will hinder future work/customisation	Medium	Tooling exists within the community to mitigate the complexity of the data model. Resources exist to automate certain difficulties associated with integrating FHIR with other data models, such as OMOP.

3 Next steps

The initial data model provides a core foundation for the data sharing platform and will continue to evolve over the course of the project. The table below outlines the next steps and planned stages related to the data model. Web interface development will continue in parallel, with each stage informed by user feedback and testing.

Stage	Target date	Description
Initial data model	July 2025 (Completed)	As described in this document
Data model refinement	November 2025	- FHIR test deployment: Deploy test instance of FHIR, populate with de-identified data subset, and utilise it in the prototype via standardised APIs - Refined data model and harmonisation strategy: updated model incorporating feedback and strategically selected ontologies, aligned with open standards (e.g. GA4GH) - Evaluate integration of OMOP data model to complement FHIR (more info here) - New version of data model validated and published - Defined API requirements: core use cases, acceptance criteria, and stub integration tests where applicable
API development	May 2026	- Secure API development: Implementation and testing of APIs for basic data access and sharing, compliant with FAIR principles - API documentation and release: Open-source publication of fully documented APIs - Semi-automated DAC system: Deployment of data access request system, validated against user stories and access governance needs
Production release	November 2026	- Finalised web interface, data model, APIs, and DAC system delivered as a fully integrated platform.