The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform

Francesco Cremonesi, Vincent Planat, Varvara Kalokyri, Haridimos Kondylakis, Tiziana Sanavia, Victor Miguel Mateos Resinas, Babita Singh, Silvia Uribe

The need for multimodal health data modeling: a practical approach for a federated-learning healthcare platform

Francesco Cremonesi[1,2], Vincent Planat[3], Varvara Kalokyri[4], Haridimos Kondylakis[4], Tiziana Sanavia[5], Victor Miguel Mateos Resinas[6], Babita Singh[7], Silvia Uribe[8]

Abstract

Federated learning initiatives in healthcare are being developed to collaboratively train predictive models without the need to centralize sensitive personal data. GenoMed4All is one such project, with the goal of connecting European clinical and omics data repositories on rare diseases through a federated learning infrastructure. Currently, the consortium faces the challenge of a lack of well-established international datasets and interoperability standards for federated learning applications in healthcare. This paper presents our practical approach to select and implement a Common Data Model (CDM) suitable for the federated training of predictive models within the medical domain. We describe our selection process, composed of identifying the consortium's needs, reviewing our functional and technical architecture specifications, and extracting a list of business requirements. We review the state of the art and evaluate three widely-used approaches (FHIR, OMOP and Phenopackets) based on a checklist of requirements and specifications. We discuss the pros and cons of each approach considering the use cases specific to our consortium as well as the generic issues of implementing a European federated learning healthcare platform. A list of lessons learned from the experience in our consortium is discussed, from the importance of establishing the

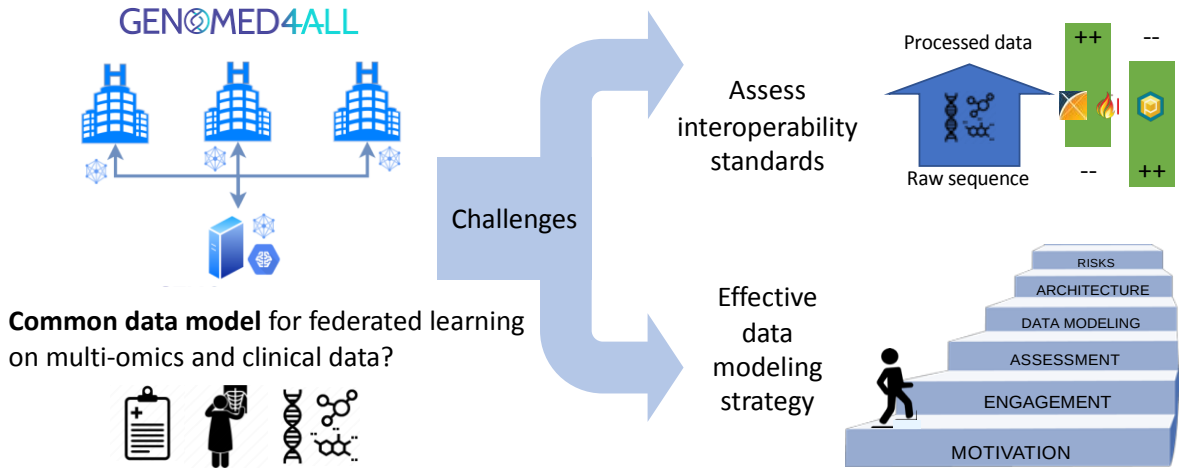
[1] Université Côte d'Azur, Inria Sophia Antipolis-Méditeranée, Epione Research Project, France
[2] Datawizard s.r.l., Rome, Italy
[3] Dedalus, Global Consulting, France
[4] Institute of Computer Science, Foundation for Research and Technology - Hellas, Crete, Greece
[5] Department of Medical Sciences, University of Torino, Torino, Italy
[6] Dedalus Healthcare, Málaga, Spain
[7] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
[8] Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, Universidad Politécnica de Madrid, Madrid, Spain

proper communication channels for all stakeholders to the technical aspects related to -omics data. For large, federated learning projects focused on secondary use of health data for predictive modeling, encompassing multiple data modalities, a phase of data model convergence is sorely needed to gather different data representations developed in the context of medical research, interoperability of clinical care software, imaging and omics analysis into a coherent, unified data model. Our work identifies this need and presents our experience and a list of actionable lessons learned for future work in this direction.

Keywords

Federated learning, data model, healthcare, omics, lessons learned

Graphical Abstract



Statement of Significance

Problem	A lack of well-established and FAIR data modeling approaches for multi-centric, multi-modal federated learning projects in healthcare.
What is already known	Federated learning projects often rely on proprietary data models developed within their own boundaries. However, large-scale international collaborations require a standardized approach, following FAIR principles, able to represent a wide variety of data types (including clinical and –omics data), and designed to support statistical analysis and training of predictive models.

What this paper adds	We identify the need for a common data model in federated learning healthcare projects, relate our experience and methodology for its implementation in the GenoMed4All project, and compile a list of actionable and general lessons learned for similar endeavors.

Introduction

The lack of availability of high-quality datasets comprising large volumes of data represents a significant limitation to the training of machine learning models in the healthcare domain. One solution to this problem, the pooling and integration of datasets from multiple centers, is often met with a strong resistance in the healthcare setting. In addition to the ethical, motivational, and economic barriers preventing data sharing, a review identified legal barriers such as privacy protection, political barriers such as lack of trust, and technical barriers such as a lack of widely adopted sharing solutions **[1]**. These barriers are exacerbated in the context of rare disease research, where datasets are small and thus more easily identifiable, and integration of multiple data modalities (genomic, phenotypical, demographic, medical history, and others) is often required.

Recently the Federated Learning (FL[9]) paradigm has gained popularity as a means to overcome such limitations by providing a scalable, privacy-preserving approach to the joint training of machine learning models across federated health data repositories **[2,3]**. The core idea behind FL is that instead of sharing data, different centers only need to share the parameters of the machine learning model being trained. Despite being very promising, FL may not yet be considered a mature technology in the healthcare domain and international initiatives based on this approach are only now starting to be developed. The GenoMed4All consortium aims to achieve one of the first international implementations of a FL platform for clinical and -omics rare disease data. However, the current challenge that the consortium faces is the systematic lack of standardized protocols for interoperable data and lack of FAIR data principles (Findable, Accessible, Interoperable, Reusable) in the biomedical data community. The multidisciplinary nature of GenoMed4All requires the strict cooperation of different communities, including clinical medicine, data science, engineering, legal, and data

[9] FL: Federated Learning; CDM: Common Data Model; AI: Artificial Intelligence; ML: Machine Learning; SoR: Sources of Records; FAIR: Findable, Accessible, Interoperable, Reusable; EHR: Electronic Health Record; CDSS: Clinical Decision Support System; SCD: Sickle-Cell Disease; MM: Multiple Myeloma; MDS: Myelodysplastic Syndrome.

privacy.

In order to protect and disseminate such valuable information and to mitigate the risk of building a tower of Babel, the GenoMed4All consortium, created in 2021, has put significant preliminary efforts to understand the current state of the art on data interoperability and consented reutilization of data of genetic and phenotypic origin in a FAIR way, utilizing specialized strategies to facilitate the sharing of clinical data and support the FL paradigm. Despite the large volume and heterogeneity of formats and representations available in the literature for healthcare data, we faced the challenge of a lack of state-of-the-art examples of real-world applications of such approaches to large-scale, international, federated learning projects. This process culminated in the definition of the requirements for a project-wide Common Data Model (CDM), an abstract representation of the structure of the data, rather than of the data itself, aimed at explicitly conveying and formalizing the organization of the data and the relationships among data elements via the specification of a machine-readable, codified format. In this article, we present the methodology and the outcomes of our data modeling effort, based on identifying a list of requirements specific to our project which allowed us to study and compare three widely used and highly standardized candidate data models: OMOP, FHIR and Phenopacket. We share the key challenges and lessons learned in the process of designing a model for multimodal healthcare data with the goal of supporting the federated learning paradigm. Ultimately, we hope that our work may be transferable to similar efforts in the future, enabling new initiatives to build on top of our experience.

This paper is structured as follows: in the Materials and Methods section, we present our journey towards the definition of a data model for the GenoMed4All platform through a process of reviewing our implementation plans, eliciting requirements, and selecting a CDM from the available international standards, thus roughly following our chronological order. We begin by establishing the motivation for a CDM in our consortium; then we present our functional and technical platform architecture focusing on how the data flows, Artificial Intelligence (AI) functionalities and technical design choices relate to the CDM; building on this knowledge, we elicit a list of business requirements from all stakeholders; finally, we evaluate the three candidate models based on this list of requirements. In the Results and Discussion section, we identify the pros and cons of each approach and present a series of lessons learned related both to the selection of a specific data model, and also to the generic process of data modeling for healthcare federated learning platforms.

Related work

The last few decades of IT systems research in healthcare have seen the proliferation of a very heterogeneous landscape of proprietary data models for storing and using health data, making it difficult to promote an environment of interoperability, FAIR data sharing and reproducible scientific collaboration **[4]**. The main reason behind such diversity is that these systems have been developed within the confines of specific design spaces, reflecting medical sociology, recording practices, and research needs, as well as serving different communities like clinical practitioners or clinical research. Standard data representations are typically developed by providers of health care, health insurance, or research, based on data formats from multiple sources such as electronic health records (EHR), lab tests, insurance claims, and specialty electronic devices **[5]**. We may distinguish standards:

- for improving the connection between health research and care delivery, such as the HCSRN-VDW **[6]** and the CESR-VDW **[7]** common data model;
- focused on simplifying the sharing of data, for example HL7 clinical documents Architecture (CDA) **[8]**, medical records **[9]**, medical imaging information **[10]**, or for the pharmacovigilance use case **[11,12]**;
- those to facilitate data exchange between different health systems like FHIR **[13]**;
- Designed to map electronic medical records as part of medical insight and research data hubs, and/or for mapping and collaboration of datasets like OHDSI OMOP **[14]**, i2b2 **[15]**, PCORNET **[16]** (patient-reported and payor data for research) or even FHIR (considered as a data warehousing standard);
- Specialized into acquisition, archive and interchange of metadata and data for clinical research studies like CDISC **[17]**.
- Those specialized in genomic data collection, storage, analyzing, and sharing representation and sharing like SAM/BAM, VCF, Phenopacket endorsed by the Global Alliance for Genomics and Health (GA4GH) **[18]** **.**

In spite of the abundance of specialized data formats, reutilization of medical data for clinical research purposes is still a rare occurrence **[19]**, especially in the case of predictive analytics and machine learning **[20]**. Moreover, these different standards are covering distinct functional domains, which raises the problem of a mature CDM covering clinical, imaging and -omics data types.

Out of those models, within the GenoMed4All project we identified three CDM candidates based on our experience and expertise, that have been demonstrated to support predictive analytics including genomic data: FHIR, OMOP-CDM and Phenopackets. The HL7 Fast Healthcare Interoperability Resources (FHIR) is one of the most robust and complete healthcare data persistence and exchange specifications that support full semantic interoperability. FHIR has already been successfully applied to a federated learning approach, for example in the context of the Personal Health Train Project **[21].** The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) **[14]** is one of the most widely used CDMs in the life sciences community and has been demonstrated to facilitate ML-based oncological studies **[22],** clinical predictive modeling **[23],** and other ML-based analysis **[24].** While neither standards were developed with a specific focus on genomics, both have been extended to -omics data **[**ERREUR ! SIGNET NON DEFINI.,ERREUR ! SIGNET NON DEFINI.,ERREUR ! SIGNET NON DEFINI.**]**. Conversely, the Global Alliance for Genomics and Health Steering Committee approved Phenopackets as a format specifically designed to integrate genomic and phenotypic information **[25]**.

Federated learning has been successfully applied in the context of medical research in the past decade (see e.g. **[26,27,28]** for some early applications), and recent reviews confirm a growing trend in terms of number of large-scale FL projects **[2,29,30,31]**. This paper deals with the question of how to best identify and integrate healthcare interoperability standards and a data model within the context of a FL platform. While recent efforts have been made in this regard, for example FL platforms relying on OMOP as a backend data representation and FHIR as a data transfer format have been described **[32,33]**, to our knowledge this question still remains understudied.

Materials and Methods

Federated learning is a distributed machine learning paradigm where multiple participants collaborate in the training of a unique global predictive model. There are multiple flavors of FL depending on the partitioning of the data, the type of participants, and other aspects **[34,35]**. **Erreur ! Source du renvoi introuvable.** broadly shows the main workflow and architecture of the horizontal, cross-silo, model-centric FL approach chosen by Genomed4All. In this paradigm, each local node (called edge node in FL notation) trains a

local ML model based on the features extracted from the local existing data. Once these models are prepared, only their parameters are sent to the central node, thus avoiding the sharing of personal data. In the central server, the parameters coming from the different edges are aggregated to obtain the general model, whose parameters are also sent back to the nodes to locally run the main model. This process is executed iteratively multiple times until a satisfactory level of convergence has been reached. The main feature of this approach is that the only information being exchanged are the models' parameters.

The purpose of the GenoMed4All project is to build a platform enabling data scientists to conduct FL experiments leveraging the network of medical Sources of Records (SoR) within the consortium. In what follows, we present the outcomes of our preliminary requirements gathering, identification of design principles, and technical infrastructure proposal, which are guiding the initial implementation attempts currently underway. Several details, especially regarding the technical implementation, have not yet been fully fleshed out.

The implementation of the GenoMed4All platform will be carried out in a multi-phase approach. The first phase is focused on the deployment of a centralized version of the AI models based on a subset of the data collected in a centralized feature store, after the necessary pseudonymization and other privacy measures have been applied. The purpose of this phase is to enable a slight degree of data exploration and model debugging to data scientists, with the ultimate goal of improving the definition of the model architecture and hyperparameters. The second phase will put in place the actual federated training and deployment within the platform, allowing an improvement of the models developed in the first phase by means of recruiting additional participant sites, thus leading to more data and hopefully better generalization of the models.

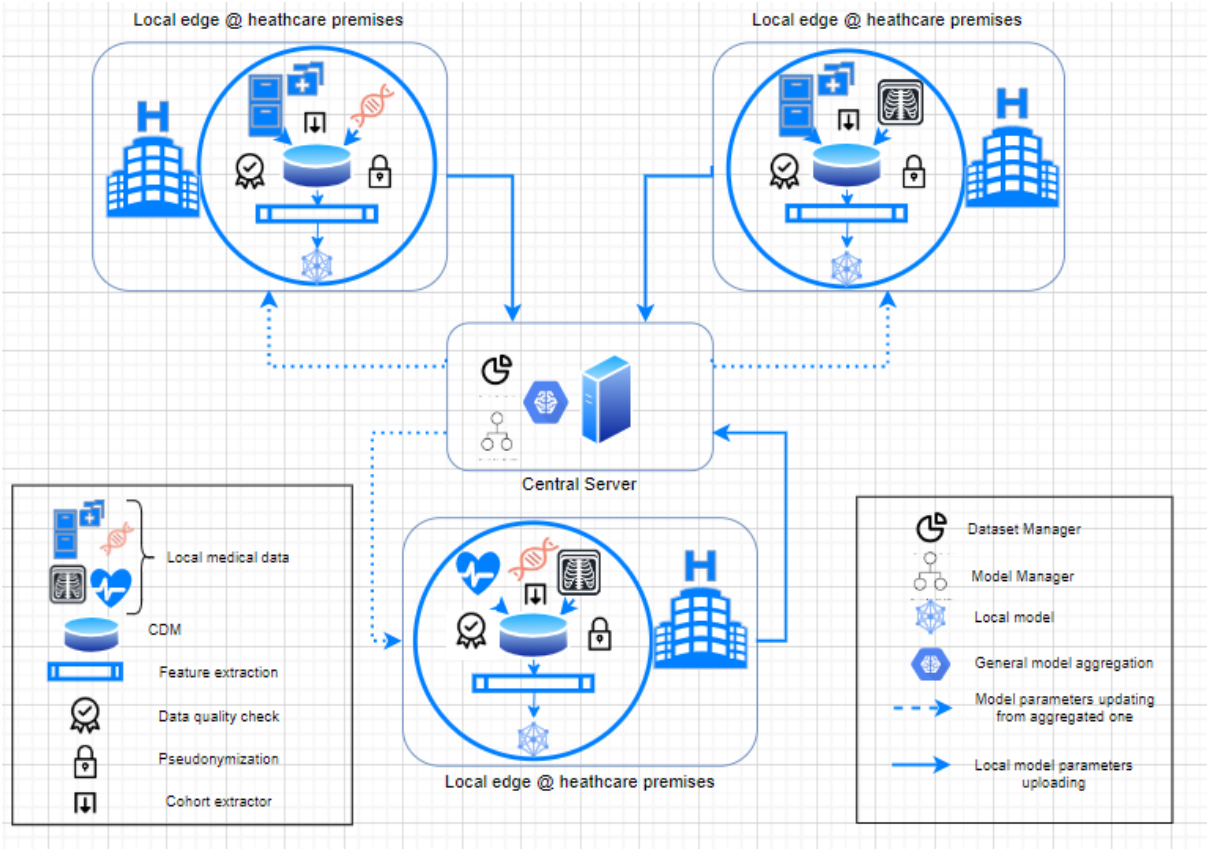


Figure 1 Federated learning platform architecture for GenomedAll. The central server is in charge of registering models along with the associated dataset characteristics (described in the model descriptor). Training monitoring and dataset exploration (based on metadata) are also envisioned from this location. The multiple edge nodes run the CDM preceded by the data integration pipeline ingesting and transforming the different sources of records. Out of the CDM the data is prepared and normalized for model training.

Motivation for a common data model

The large heterogeneity in data representations and semantics across SoR leads to unnecessary complexity in the analysis, reuse, and interpretation of data in health research. A well-established approach to overcome this issue relies upon a Common Data Model (CDM), a collection of rules which standardizes both the structure and the semantics of disparate datasets through the use of ontologies, coding systems, and formal documentation.

During the preliminary design phase of the functional and technical architecture for GenoMed4All, we realized how much even the simple issue of justifying the need for a CDM for federated learning platform requires careful analysis and a tight interdisciplinary collaboration. Using a CDM does not come for free: it requires training software developers and data analysts, significant levels of expertise and domain knowledge, and a considerable initial development effort. However, there's a lot to be gained by this investment.

In a federated approach, a CDM adopted by all SoR is crucial to avoid the combinatorial complexity of maintaining separate data preparation and feature extraction pipelines for each combination of SoR and ML algorithm. This reduces the burden on data analysts and allows clinical and research data to be appropriately merged and compared across institutions. Additionally, following CDM specifications, such as e.g. adopting specific terminologies and ontologies, can greatly improve semantic consistency across SoR, ultimately leading to better data quality. Finally, such a standardized approach makes the raw data auditable and searchable through automatics means [**36**].

The long-term maintenance and scalability of the platform and of the ML models may also be impacted by the adoption of a CDM. For example, when onboarding new members into an experiment a data quality assessment is mandatory to avoid a degradation of the overall federated training performance. And although data quality is essential for both centralized and federated solutions, in a centralized solution data quality checking and data update are easier to implement as data are located centrally and can be examined at will, whereas federated learning imposes additional difficulties as data are are distributed, and the modelers do not have access to the data [**37**]. This data governance step may be greatly facilitated if a CDM is enforced between the sources of data and the data preparation pipeline, allowing predefined scripts and metrics to be extracted before engaging into the training process.

Finally, the adoption of a CDM may also benefit health care providers directly by simplifying the deployment of other data services like data exploration, decision support or alerting systems, thus further contributing to offsetting the initial data management costs associated with a CDM. Especially for DSS, the heterogeneity of systems with varying data types and structures has been identified as an important factor that hinders CDSS implementation in a real clinical setting [**38**].

Data modeling for federated training of AI models

The interaction between data modeling and the training of learning algorithms is an evolving subject of exploration as it must align two communities with their own practices and tools: AI researchers and IT platform integrators. To the best of our knowledge this topic has not been extensively analyzed in the healthcare domain (for example the IHE AI interoperability in Imaging white paper **[39]** provides limited indication for data/AI model interaction). In

Genomed4All project this topic has been addressed by the Federated learning platform both during the model development and training phases.

Model development phase

Data scientists in the consortium have identified a need for a preliminary model development phase in which a subset of the data is centralized an made available to them. A dataset anonymization and preliminary analysis is usually performed in this phase prior to entering the research phase. At this stage, the datasets delivered to data scientists are very often not aligned to a CDM.

After the data is made available to the data scientists, there are more challenges to be solved in order to develop an ML-based model and provide an automatic application of the whole process, such as e.g. the choice of the ML algorithm, which strongly depends on the data, and therefore on the data model. We expect that higher heterogeneity of the features considered in the datasets corresponds to more complex relationships among these features, fostering the usage of ML-based solutions. In addition, the model's validation should be performed on a set of patients that are not used to learn the normal parameters, in order to avoid overfitting issues and to generalize well when applied to new patients. **2** Different data representations can impact the model development steps described above, for example by inducing different training and validation data splits, introducing numerical errors, or affecting the feature selection process.

At the end of the development and validation phase the research team delivers a *model descriptor*, an object whose goal is to provide a description of all the information and resources needed by the FL edge system to perform its job. Indeed, in most cases the model is not coded to support a CDM data format as input and so a data transformation pipeline must be described for preparing the data for training. The *training descriptor* includes: a cohort script to extract data from the CDM into the format requested by the model; imaging and genomic pre-processing pipeline descriptor; data quality check metrics; the model and its training plan; the minimum system resources capabilities (CPU, Memory, Storage) to perform the training.

Federated platform for model training

Following common design patterns **[18]** we split the architecture in two distinct components, matching the Federated learning paradigm presented above. A *central server* whose primary objective is the registration, training monitoring and aggregation of models; and an *edge server* in charge of driving the training plan and pushing model updates once training is finished. These two sides of the architecture handle different components and resources that are closely relying on the data stores implementing the CDM. The components of the architecture that interact with the common Data model are presented in **Erreur ! Source du renvoi introuvable.** and Figure 2. A short description of their interaction with the CDM is given in Table 1.

Table 1 Platform components having significant interactions with the CDM.

Component	**Role and interaction with CDM**
Pseudonymization	(Optional) Depends on hospital data privacy governance rules as the CDM may be used by other research teams to explore & create cohorts.
cohort extractor and dataset	Extraction of the dataset from the CDM that will serve the model training at the edge (see Figure 2)
Quality check metrics	Computed on the dataset (mandatory fields, validity & completeness) to asset the quality & exclude any data that would pollute the training phase (see Figure 2)
Dataset manager	To reference, as a catalog (based on metadata aligned with the CDM format to ease the analysis) the datasets already deployed in the different edge hospitals (see **Erreur ! Source du renvoi introuvable.**)
Model manager	to reference, as a catalog, all the models that are currently handled by the federated training framework. The registration references all the resources associated with the model (cohort extraction script, data quality check metrics, pseudonymization

The model training data flow

The diagram in Figure 2 presents the data flow orchestrated by the FL platform at the edge. The training descriptor, described in the previous paragraph, is used by the integration pipeline to configure the data flow execution. It is also used by the FL edge model trainer to

control and perform the model training. In this flow, the CDM ingests data from the different SoR after feature extraction for clinical images and genomic data, alignment on a single terminology and formatting into the CDM, and finally pseudonymization. Then the cohort extractor retrieves a dataset from the CDM and applies data normalization using a set of dictionaries to convert categorical data into the numerical data format for input into the AI model. A data quality check based on the training descriptor metrics is then applied, and if successful the model training plan can start. When it ends, and if the quality reaches the expected level, the model update is sent back to the central platform for aggregation. The integration of the algorithm into a Clinical Decision Support system used by clinicians requires further validations that are out of the scope of this data flow.

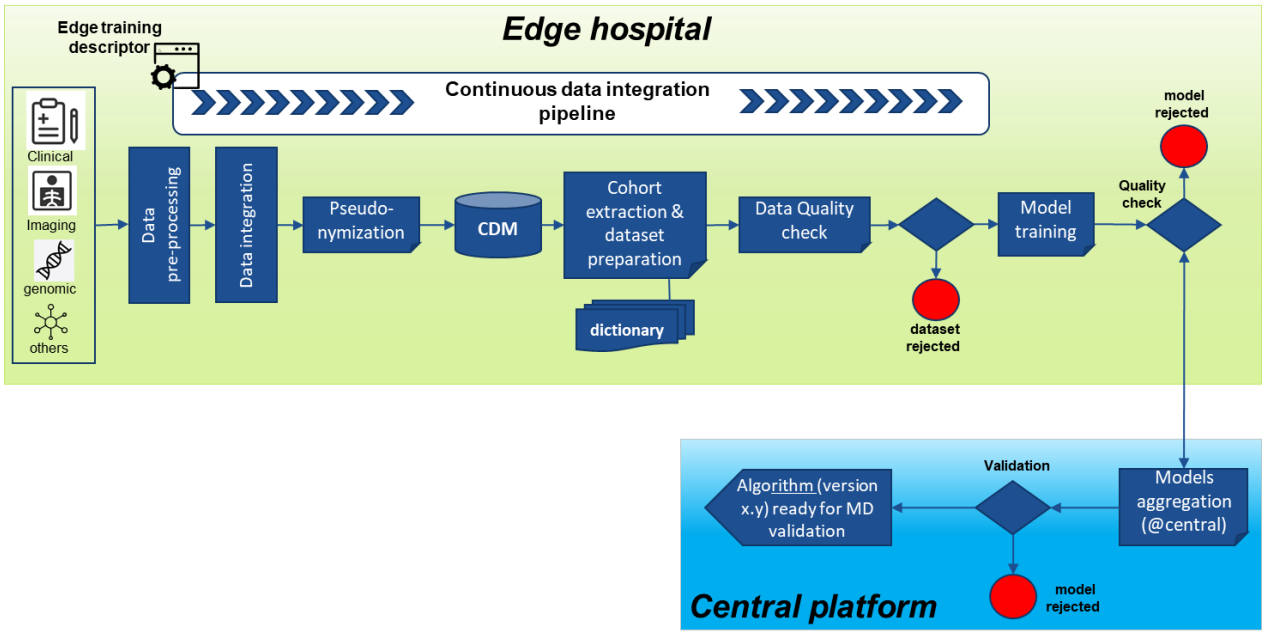


Figure 2 Conceptual Data flow for model training: from the source of record to the deployed algorithm in a decision support system. The pipeline execution and model training are configured based on the model descriptor which provides: i) the cohort extraction definition along with the dictionary for data normalization; ii) the minimum data quality metrics to accept data for training; iii) the training plan. Once the training is finished the model is sent back to the central platform, which is responsible for detecting potential data shifts and aggregating a new resulting model. This new version enters then the Decision support validation phase (which may include a clinical trial step).

Data modeling business requirements

We present an extensive list of requirements for the GenoMed4All federated learning platform in Table 2, with the goal of providing a template and inspiration for similar projects, and extrapolate generalizable lessons in the discussions section. A first version of these

requirements was obtained by analyzing the proposed technical infrastructure, data flows, AI model development strategy described in the sections above. Furthermore, a list of high-level data elements present at each SoR in the consortium was gathered to aid in identifying whether the shortlisted CDMs could indeed adequately represent the data (see Appendix "Genomed4All: clinical demonstrators and available data"). The requirements were discussed in the context of group and individual interviews with the different domain experts in the consortium, including AI experts, clinical data providers, and software engineers, following a requirement engineering approach **[40]**. After that, the requirements were refined and the levels of priority, as defined by the MoSCoW rules **[41],** were assigned iteratively and collaboratively in the context of the bi-weekly consortium meetings as well as during the course of two consortium-wide workshops.

Specifically for the data transformation requirements, for a federated learning setting to be able to work the data used for training in the various hospitals should use the same data model, format and terminologies. However, as this is rarely the case, it is common that transformations should be performed before the data are actually usable for the any federated learning model. Although neither the FAIR principles nor federated learning have a prerequisite the data transformation steps this is commonly required to be performed in practice **[ref].**

Table 2. Data model requirements for the GenoMed4All project. We collected the main expectations and constraints from different stakeholders regarding the data model to be used in the GenoMed4All project. We summarize the most commonly expressed points, categorized into different general topics.

Input by clinical data providers	The raw clinical data provided as input may or may not respect interoperability standards, and it may be in a proprietary format. The platform SHOULD support as many input formats as possible.
	The data model SHOULD support the use of the same terminologies and ontologies as the raw source data.
	The data model SHOULD be customizable & extensible allowing adaptation to data format specifications brought the project.
	The data model SHOULD be agnostic to any terminology.
	The data model SHOULD be unique across all data providers' edge nodes.
	The effort required from clinical data managers to prepare the data for federated training through the platform SHOULD be minimized.

	The data model MUST support both cross-sectional and longitudinal data.
	The data model COULD support updates and changes to the data structure (slowly changing dimensions).
Data privacy	Data relating to the same patient across multiple centers/datasets SHOULD be recognizable as belonging to the same subject.
	The data model MUST support all GDPR and national privacy laws.
	The data model MUST support the possibility to cancel data at the granularity of individual patients, in respect with GDPR.
	The data model MUST support any pseudonymization strategy.
	The data model must be implemented by at least one technology that offers access granting capabilities to meta data allowing the care institution to control its data exposure level to other actors.
Data model design	The data model MUST support the representation of all target data types at their intended level of detail, including clinical, demographic, genomic, radiomic, and laboratory data.
	The data model MUST support additional information extracted from the raw data, possibly through automated AI algorithms, such as e.g., radiomics and genomics features
	The data model MUST respect FAIR principles.
	The data model MUST be agnostic w.r.t. the origin of the data.
	The data model COULD retain information about the origin of the data.
	The data model SHOULD be patient-centric.
	The data model COULD be highly normalized to minimize duplication of information.
	The data model MUST support both storage and transfer semantics and operations.
Data transformations	The data model MUST provide a uniform interface for accessing and querying data for the purposes of training federated AI models.
	The information contained in the data model after transformation MUST be at least equivalent to the information contained in the raw source data.
	Data transformations from the raw source to the AI training dataset SHOULD be as simple as possible.
	Data transformations from the data model to the training dataset COULD be disease-specific, algorithm-specific, and training plan-specific.
Data consumers and federated	The quality of the data exposed by the CDM for training predictive models MUST be at least as high as the quality of the raw input data.

training	Given any training plan for an AI algorithm, the data model SHOULD be structured in such a way that simplifies the retrieval of all the records in the training dataset.
	In-depth knowledge of the data model SHOULD NOT be required from data scientists to manipulate the training dataset.
	The data model MUST be sufficiently documented to enable the engagement of additional clinical data providers.
	The data model MUST provide extensive metadata to enable basic data exploration and simple aggregated queries.

Given the high level of detail in Table 2 and the specificity of some of the requirements to the GenoMed4All project, we also extracted a more generic data architecture checklist in Table 3 to enable future projects to build on our experience. The checklist covers all phases of the data flows – from the input of raw clinical data to the generation of training data for the ML algorithms – and is structured as a short list of ten questions and related recommended actions. In this checklist we aimed to capture the main difficulties, challenges, and points of friction that we encountered in the process of designing the data model for the project, with the goal of providing a simple yet effective way to pragmatically approach the design of similar systems in the future.

Table 3 Data architecture checklist. We distilled the main points of discussion from our experience into a checklist of questions and recommended actions, to be used for future implementations of a data architecture based on a CDM in a federated learning platform in the healthcare domain.

Phase	**Question**	**Recommended action**
Input by clinical data providers	What are the clinical aims and the research question?	*An iterative research design process involving clinical experts (to define the research question, describe available data, and evaluate proposed analysis) and ML experts (to assess available data and define ML approach).*
	What data and what type of statistical or ML analysis can be used to answer such question?	
	What format is the raw data available in? Which standardized nomenclatures and ontologies have been used in the raw data?	*Survey among clinical data providers.*

Data privacy	How to ensure full compliance with GDPR and national data privacy laws?	*Identify GDPR roles, pseudonymization strategies, and assess data flows. Consider carrying out a full DPIA.*
Data model design	Does the data model accurately cover all the data types, with straightforward mapping from the raw data?	*Gather metadata about all available data types and carry out a preliminary mapping exercise for all candidate data models.*
	Does the data model support both input by clinical data providers and extraction by data scientists? How might the data be modified/updated in the future?	*Preliminary analysis of possible strategies for data input, modification, and extraction.*
	What kind of auditing and governance metadata is supported by the data model?	*Preliminary assessment of metadata needs and exploratory mapping to data model.*
Data transformations	How to ensure proper data governance and how to retain semantic content and the same level of information throughout the data transformation, wrangling and curation processes?	*Establish two detailed mapping documents: the first maps the raw data to the data model, the second maps the data model to the training data format.*
Data consumers and federated training	Does the data architecture support all intended use cases for data consumption?	*Identify all use cases (with support of clinical, AI and software engineering experts) and assess support.*
	Is the data quality sufficient to support federated training of ML models?	*Identify data quality needs and evaluate putting in place quality checks (automatic or manual).*
	What are the challenges and biases related to the (potentially uneven) distribution of data across SoRs?	*Survey data providers to identify biases in patient demographics' distributions and data volumes. Conduct statistical analysis prior to the training of predictive models.*

Mapping data models to the requirements

One of the main technical challenges that we encountered in the process of selection and design of the data model was to identify which available international standards best suited our needs. We present here a preliminary mapping exercise trying to identify whether the three data models under investigation are adequate in covering all data aspects that should be modeled within GenoMed4All. We expand on the "Data model design" row from the checklist in Table 3 and our prior work on gathering the requirements expressed in Table 2 to identify a set of dimensions to evaluate three commonly used standards for clinical and genomics data: FHIR, OMOP, and Phenopackets.

Our first concern was to look at the expressivity of each model with the goal of making sure that all types of data may be adequately represented in the platform. We summarize the results of this exercise in Table 4 for the clinical data. Given the importance of genomics and –omics data for GenoMed4All, and the relative lack of maturity of well-established standards for these data types, we also present a detailed analysis for genomics data in Table 5, and Table 6 for other - omics data. Given the still-preliminary nature of our investigation, we did not yet have access to the full list of data elements that will be made available for federated training. Instead, these tables provide a first cursory analysis on possible mapping strategies that would satisfy our project requirements for data macro-categories, with the caveat that details about each data type may lead to different approaches when the full list of data elements is finally made available. Additional details about this mapping exercise can be found in the Appendix "Additional details for mapping exercise".

Table 4 Mappings between the three candidate common data models and the GenoMed4All clinical information.

Clinical Information			
	FHIR	**OMOP-CDM**	**Phenopacket**
Demographic	Patient, Observation	Person, Observation, Measurement	Individual
Treatments	Procedure, ServiceRequest	Drug_Exposure, Procedure_Occurrence	Treatment
Diagnosis	Observation, DiagnosticReport	Condition_Occurrence	Disease
Conditions/Clinical Manifestations	Condition, Observation	Condition_Occurrence, Observation	Phenotypic features, Evidence
Laboratory	Observation,	Measurement	Biosample, Measurement

	Measure, MeasureReport		
Longitudinal data (History)	Most FHIR resources of interest have a date field, e.g. Observation.effectiveIssued, Procedure.performedDateTime	Condition_Occurrence linked with Observation	Individual elements such as Individual, PhenotypicFeature, Biosample, Medical action, combined with building blocks such as, Time element to create longitudinal structure
Terminologies/ Ontologies	CodeSystem ValueSet, ConceptMap	Concept, Vocabulary, Domain, Concept_class, Concept_relationship, Concept_synonym, Concept_ancestor	Ontologies are well supported in phenopackets

In the specific case of genomics, we found this mapping exercise to be quite challenging even at the relatively coarse level of detail that we are considering here. For example, while our proposed approach for mapping the information of which variants are significant in FHIR relies on the interpretation field within the Observation resource, there is debate whether this is the best approach, as other valid alternatives can be considered such as e.g. the Variant confidence status [**42**]. Similarly, it remains unclear to us whether the Variant_annotation table represents the best mapping in the OMOP-CDM.

*Table 5 Mappings between the three candidate common data models and the GenoMed4All **genomics** information.*

Genomics Information			
	FHIR	**OMOP-CDM**	**Phenopacket**
Presence/absence of each mutation	DiagnosticReport-genetics, Observation-genetics, Molecular_sequence	Variant_Occurrence	VariationDescriptor
Genomic location	DiagnosticReport-genetics, Observation-genetics	Genomic_Test, Target_Gene, Variant_Occurrence	VariationDescriptor
Variant type	DiagnosticReport-genetics,	Variant_occurrence	VariationDescriptor

	Observation-genetics		
Genotype information	DiagnosticReport-genetics, Observation-genetics	Variant_occurrence	VariationDescriptor
Annotation of which variants are significant	Using interpretation field in Observation	Variant_annotation	VariantInterpretation

*Table 6 Mappings between the three candidate common data models and the GENOMED4ALL **–omics** and other data types information.*

Other data types and –omics data			
	FHIR	**OMOP-CDM**	**Phenopacket**
Imaging data	Media, ImagingStudy, DiagnosticReport	Radiology_Image Radiology_Occurrence	N.A.
Oxygenscan data	Observation, Measure, MeasureReport	Procedure_occurrence, Measurement	N.A.
Metabolomics	Observation, Measure, MeasureReport, DiagnosticReport	Measurement (with LOINC and SNOMED)	N.A.

Results and Discussion

Comparison of data models

Based on our coarse-grained preliminary analysis, all three data models were deemed generally capable of supporting all the clinical and genomics data requirements foreseen for the GenoMed4All project, as well as all the necessary processes for querying and transforming information required by the AI algorithms. We found more overlap in the capabilities offered by FHIR and OMOP, whereas Phenopackets, which has better expressivity for genomic data, has the disadvantage of not being able to adequately represent some common data types such as imaging or other –omics data. The complementarity of the three data models is particularly obvious in the specific case of representation of genomics

data, where we highlighted some difficulties in mapping specific concepts to FHIR and OMOP. We report in **Erreur ! Source du renvoi introuvable.** the hierarchy of genomics data needs for GenoMed4All [43]: while FHIR and OMOP are better suited for highly processed genomics data and linking with other clinical data types, Phenopackets retains the ability to natively express raw sequencing data and linking it to specific phenotypical manifestations.

In general, all analyzed data models serve well as a layer of standardization for clinical research data within one's own research network. However, in case one wants to reuse and integrate a set of datasets in broader clinical research communities across different research networks, this requires a global data model as a reference standard to facilitate not only data model harmonization and data integration, but also easy data transfer and exchange. Towards this end, there has been an effort recently to combine several data standards in a single global common data model that will allow researchers to pull data from multiple sources and compile it in the same structure without degradation of the information. We believe that a combination of data models, as well as the implementation of Phenopackets within the FHIR or OMOP standard, will facilitate in better developing and improving the interoperability and standardization in genomics data. This effort may be carried out in the context of integrating with Phenopackets-based platforms such as RD-Connect, or in the context of leveraging the protobuf layer of Phenopackets as a data interface for data scientists.

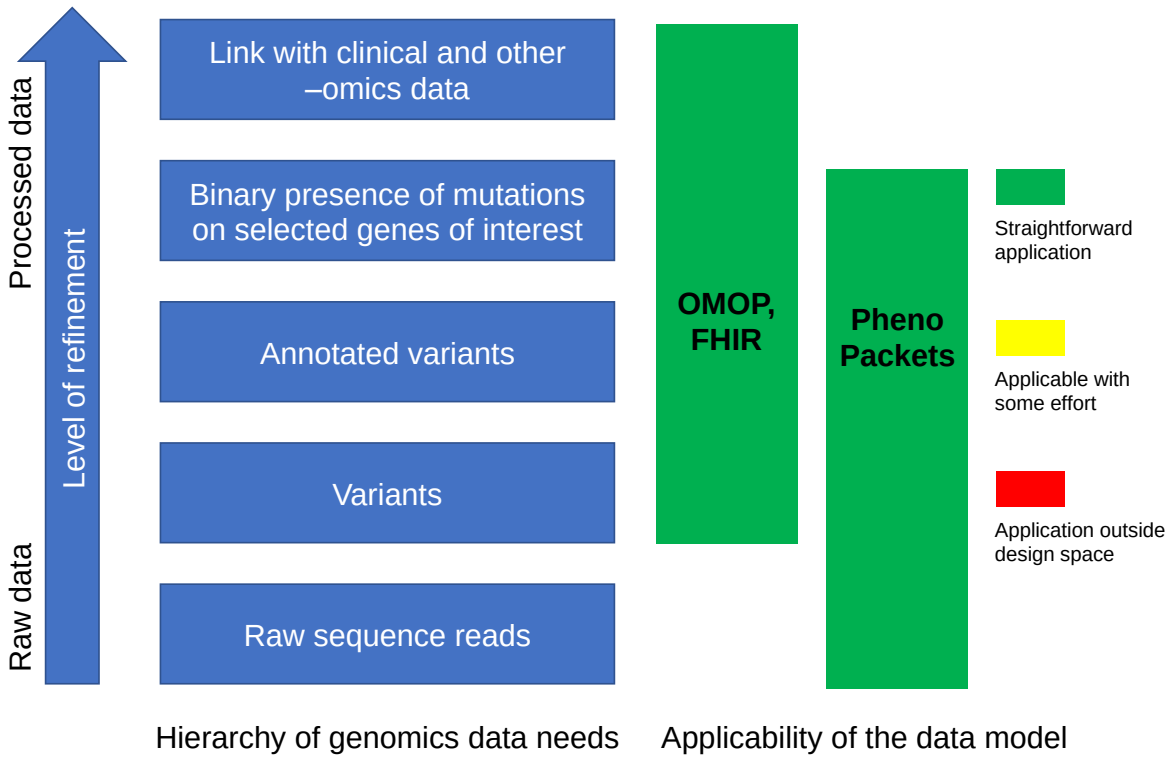


Figure 3 Hierarchy of genomics data needs in the GenoMed4All project and level of applicability for the FHIR, OMOP and Phenopacket data models. The analysis and ML model training within the GenoMed4All is expected to cover different levels of genomics data refinement, spanning from raw sequencing data to linking genomics information with other clinical data. While all the analyzed standards are able to natively represent data at a medium level of processing, we find that OMOP and FHIR are more easily applied to highly processed genomics data (green bars) but do not support representing raw sequencing data (red bars), while the situation is reversed for Phenopackets.

Data modeling for a federated-learning healthcare platform: lessons learned

The journey towards the implementation of a CDM for the GenoMed4All project has been a challenging experience that required multiple steps, engagement of different types of stakeholders, and the solution to technical and organizational issues. We summarize in Figure 4 the main steps in this journey, starting from identifying the need for a CDM and ending with the actual implementation in the data flows and processes of the project. In what follows, we relay the main lessons that we learned from our experience during this journey and extract generalizable insights for future projects.

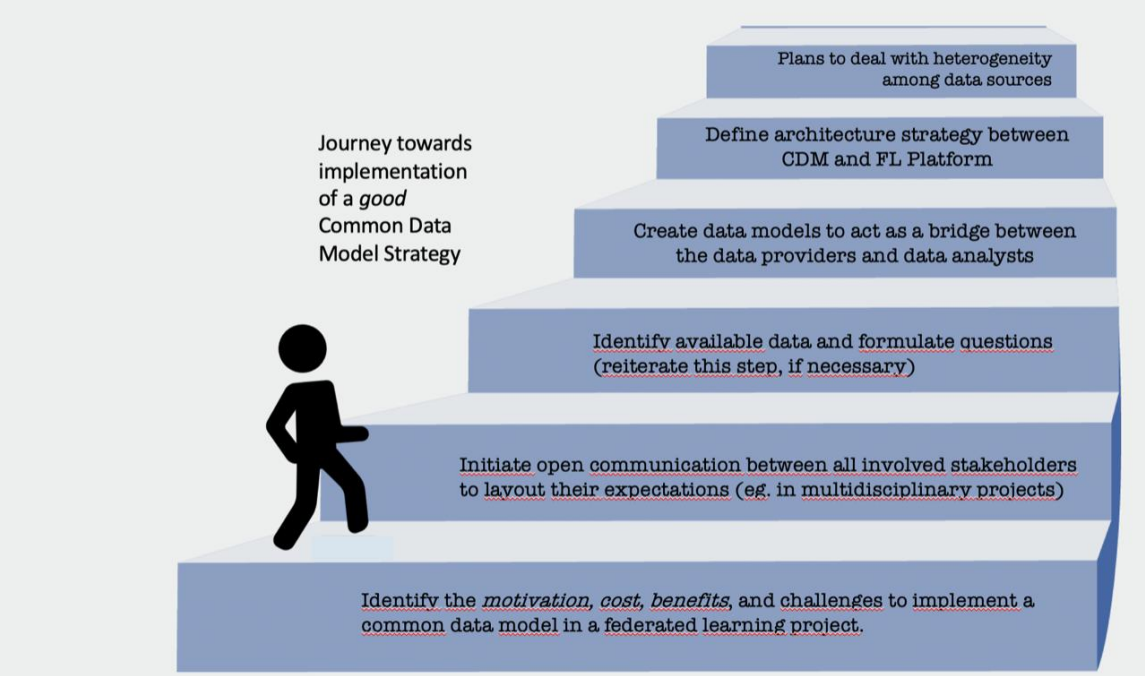


Figure 4 Journey towards the implementation of a good CDM strategy. First, the need and motivation for a CDM must be established, and a cost-benefits analysis carried out taking into account the specific context of the project. In order to be successful, open communication and engagement with all involved stakeholders (platform engineers, clinicians and ML experts) is necessary. Through open discourse and collaboration, the research questions must be formulated and their data requirements expressed. Finally, a data model may be created and/or selected, and a common strategy defined between data provides (clinical institutions) and consumers (FL platform and data scientists). Measures to mitigate risk, such as plans for dealing with data heterogeneity, must be implemented as a last step.

Lesson 1: The need for a common data model in a federated learning project

One of the main results of the data modeling exercise in the GenoMed4All project has been to identify the precise motivation, cost, benefits, and challenges related to implementing a CDM in a federated learning platform for medical research. A CDM contract between data providers and consumers enables homogeneity, consistency and standardization for both semantic and syntactic aspects, thus leading to improved data quality and reusability. Furthermore, it makes the data more readily auditable and searchable, thus also improving its findability. In a federated setting, a CDM is also crucial to support scalability, by simplifying the recruitement of new SoR, and future-proofing of the platform. However, enforcing such a contract from the beginning can be met with some resistance by clinical data providers as it shifts some of the data management burden to them – especially in the cases where they lack

the in-house expertise to perform complex dataset mappings – as well as by data scientists who may need to divert time and effort from the development and training of ML models to learning the technicalities of the CDM.

Lesson 2: Implementing a CDM in a federated network is not only a technical challenge

In our experience, the implementation of a CDM within a federated network of clinical data providers proved to be not only a technical challenge, but also a complex management issue requiring open communication between all stakeholders involved. We found that the three communities involved, i.e., the clinical, AI and engineering teams, all had slightly different expectations which needed to be balanced out. The clinical team was mostly worried about the additional effort and technical difficulty of mapping the data to yet another terminology and structure, an operation that may seem redundant at first as it does not necessarily add any more medical information to the data. The AI team was worried about ease of adoption and quick manipulation in the context of developing and testing new algorithms. The engineering team was worried about duplication of work and pushed to enforce standardization wherever possible. Similarly, we experienced a contrast between a strive for generality in designing the data model and the narrowly focused nature of ML algorithms, which represent an ad-hoc solution by their very nature.

Lesson 3: Breaking the cycle of requirements

A common stalling point in our discussions about data requirements for the project was met when we hit the cycle of requirements: to identify the available data we need to define in detail a specific research question, but to define a research question we need to know what kind of data is available. We often encountered this deadlock situation during interdisciplinary meetings, where the AI experts would ask the clinical team to identify interesting research questions, but the clinical team would answer back that this was difficult to say because they didn't know which kinds of questions could be answered with data, and when the AI team would ask what kind of data is available, the reply was that it depends on the type of question to be asked. Breaking this cycle required a long and iterative work of educating each side of the discussion to the other side's specific needs, requirements, and sometimes even language. This lessons, which applies to any AI project in the medical field,

is especially relevant for FL scenarios where the data scientists never have access to the full dataset.

Lesson 4: Convergence of data standards from the -omics and clinical domains is needed

Our comparative analysis has identified that, in the context of modeling -omics and clinical data for ML applications, there is a lack for an industry-standard common approach able to fully cover both the phenotype and the -omics information. This may partly be due to the origin of common standards, which have typically been developed within a more restricted design space. However, convergence of standards such as OMOP, FHIR and Phenopackets is sorely needed to enable the next generation of AI models fulfill the goal of precision medicine.

Lesson 5: The CDM strategy and FL platform are closely related: their integration must be anticipated

The analysis and identification of the FL platform requirements (and the choice of the underlying technology) quickly revealed a close interrelationship with the CDM. Three key elements of this platform, previously described in this paper, have exposed interfaces with the CDM: the central server dataset manager, the cohort extractor, and the client data integration pipeline. This close inter-relationship highlighted a major architecture risk if it was not well anticipated upstream in the design.

Lesson 6: Heterogeneity among stakeholder communities, and within data sources, represents a significant risk to the project

We found that discussions on the data model represented the first moment where the three different stakeholder communities (clinical, AI, and engineering) were required to concretely discuss technical details and produce actionable outcomes in our project. Among such diverse communities, communication was hampered by the lack of a shared common language and point of view, ultimately resulting in a slower start of data-related activities and, in extreme cases, loss of engagement. Terms that may seem trivial to one group were completely obscure for another: for example, commonly used words such as features and labels in the ML community were not understood by the clinical team.

In our experience, the need for structured and curated data represented the biggest challenge in integrating data from external sources into the GenoMed4All platform, especially considering that not all clinical data providers may have the expertise or the computational infrastructure to provide the necessary data flows and transformations. All the datasets used within the first phase of the project required significant manual work to reach an interoperable standardized format, directly impacting the project's execution as well as the final quality of the data. While this lesson applies to any ML project in healthcare, it is especially relevant for FL platforms due to their distributed design and explicit goal of including multiple heterogeneous data sources.

Conclusion

This work presented a case study for selecting and implementing a common data model (CDM) in a federated learning healthcare platform, based on the experience of the GenoMed4All consortium. We identified a lack of well-established implementations of medical data standards in federated learning projects at large scale, able to cover multiple data modalities and –omics data in particular, and to support the federated training of predictive models. In our case, the need for a CDM arose as the solution to challenges related to the heterogeneity in data representations and semantics across sources of records, interoperability, reutilization of secondary data, auditing, quality, long term sustainability and scalability of the platform.

In order to select the most appropriate approach for our specific use case, we first conducted a design exercise to understand how the CDM should fit in the architecture of the platform. We quickly realized that a collaborative effort from all stakeholders involved, not just the systems architecture experts, was required in order to correctly position the CDM in the full end-to-end process, from the upload of pseudonymized data by clinical data managers to the training of predictive models by data scientists. In this process we reviewed the actors, the functionalities that should be supported by the platform, and the main processes and data flows to identify all the critical points where a CDM could have an impact. This exercise culminated in a set of business requirements that must be satisfied by the chosen data modeling approach to fully satisfy all stakeholder needs. Even though the exercise was conducted with our specific use case in mind, we believe that most of the challenges, solutions and lessons learned may be generalized to other initiatives dedicated to federated

learning in healthcare, hence we share some details about our technical architecture, medical and AI approaches, as well as the list of requirements.

We then surveyed the most commonly used data models in healthcare, and preselected FHIR, OMOP and Phenopackets for further exploration due to their widespread adoption, usability for predictive analytics, and familiarity for our consortium members. Building on our list of requirements, we share the in-depth comparison and the insights derived from the analysis of these standards and their applicability to our needs. We found that all three standards are generally capable of covering all the basic requirements, however FHIR and OMOP are supported by a larger community, a wider adoption, and a better coverage of the specifics of common medical data types, while Phenopackets is better suited for the representation of genomic data and the linking with phenotypical information. Finally, we share a list of lessons learned that we believe could be applicable to other initiatives with a similar goal of supporting the federated training of predictive models applied to the medical domain, especially those with a focus on rare disease and –omics data. Namely, we reiterate the need for and importance of a CDM in such initiatives, the importance of collaboration among all stakeholder communities, the role of the CDM as a *fil rouge* in the end-to-end flows of the process, and the challenges related to heterogeneity across data sources.

Acknowledgments

GenoMed4All has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101017549. The authors would like to thank all members of the GenoMed4All for their helpful contributions and insightful discussions.

Appendix

Genomed4All: clinical demonstrators and available data

The applicability and utility of GenoMed4All will be shown with clinical demonstrators focused on specific disease areas: an oncological use case including Myelodysplastic Syndromes (MDS) and Multiple Myeloma (MM), and a non-oncological use case including Sickle Cell Disease (SCD). As a first step in the data modeling process, the clinical team worked closely with the AI experts to define a research plan. This involved finding the right

balance between identifying clinically relevant research questions and the feasibility of their analysis through ML methods. Building on the research questions and their intended AI analysis we elicited the data requirements, leveraging the expertise of the engineering team in the consortium to define a plan for a scalable and interoperable platform.

Each clinical demonstrator has identified a set of research questions to be answered with predictive modeling tools and AI, leveraging data from several repositories in a federated learning setting. The research questions range both unsupervised and supervised ML approaches, addressing both cross-sectional and longitudinal analysis of the patients' cohorts for the three use cases, and can be divided into broad categories according to the corresponding ML technique: clustering, classification, and survival analysis.

Upon discussion, it quickly became clear that the data needs vary depending not only on the disease area of focus, but also on the specific research question. We asked the clinical teams from each disease area in GenoMed4All to share detailed information about the currently available datasets to be analyzed within the scope of the project. In the interest of respecting interoperability standards, we endeavored to identify which standardized nomenclatures and ontologies were already being used. We found a highly heterogeneous situation, comprising a mix of well-established data types with their related common terminologies, as well as niche terminologies (such as the rare-disease specific ORPHA codes **[44]**) and uncommon data types generated by cutting-edge techniques such as Oxygenscan **[45]**. We summarize our findings for GenoMed4All in Table 7, where we list for all data modalities the associated technical data type and the standardized terminologies and ontologies being used by data providers participating in the project.

Table 7 Data types with related terminologies and ontologies for the GenoMed4All disease use cases.

Data modality	Examples	Type	Terminologies and ontologies
demographics	*Objective characteristics about the patient: age, sex, height, etc..*	Free text, numeric, category	
treatments	*type of*	Coded term, category	ATC;

	treatment received, when and how, including both drugs and procedures.		CPT
diagnosis	*Diagnosed disease, date of diagnosis.*	Coded term	ORPHA WHO-2016
conditions and clinical manifestations	*Characterizations of the clinical status of a patient, excluding diagnosis*	Coded term, free text, category	SCDO; ORPHA; HPO; SNOMED- CT
laboratory	*Results from tests and assays not conducted in the clinic. Includes hematological data.*	Numeric, coded term	LOINC
genomics	*Presence/absence of mutation/alteration; genomic location; type of variation; genotype information; annotation of variants significance*	Coded term, string of coded terms	HGNC; HGVS HGNC
cytogenetics	*Chromosomal abnormalities,*	Coded term, string of coded terms	ISCN string

	deletions, duplications, inversions, etc…		
imaging	*MRI or PET*	Bytes	DICOM format
oxygenscan	*Lorrca oxygenscan technique*	Numeric	None exists

Additional details for mapping exercise

Building on other requirements, we also tried to assess the ease of interacting with the data in each format, summarized in Table 8 and Table 9, as well as the possibility to embed governance and auditing information within the data model for quality purposes, as shown in Table 10.

Table 8 How to query information in the three candidate common data models.

How to query information			
	FHIR	**OMOP-CDM**	**Phenopacket**
All the data of a patient	FHIR resources are saved in document stores (e.g., ElasticSearch, MongoDB.) Using elastic API, there are the following options1. Get each document from each resource filtered by any field (i.e. patientID) with a POST request; 2. Using _msearch, it is possible to perform a single query and retrieve the information needed.	SQL query joining all OMOP tables based on the person_id. (All tables have person_id as FK)	Phenopackets is a protobuf file, so as a file it cannot be queried. A user has to store the information needed in a database and then the database will provide a way of querying it.
All the data of all the patients in a disease case study	Creating nested queries is one approach. Another approach is to create an index with all the info needed, and the user can directly make a request to this index containing all the info required.	SELECT query from the Person table joined with Condition_Occurrence, filtering on the condition values referring to the disease.	
All the data associated with	This data will be post-processed in async mode to be easily and quickly consumed.	Write the proper SQL query for getting the required information	

the training of a given AI algorithm			

Table 9 Proposed strategies for extracting information from clinical sites to the three candidate common data models, and outputting it to AI compliant formats.

I/O of data to/from CDM			
	FHIR	**OMOP-CDM**	**Phenopacket**
Extracting EHR data as input to the CDM	Python ETL tool with specific mappers - will depend on each specific data provider.	ETL tools: **whiterabbit** (summary reports informing the design of the ETL pipeline), **rabbit in a hat** (syntactic mappings), **usagi** (mappings between vocabularies), **Achilles** (assessing mapping quality),	By using **FHIR - phenopackets**, if the hospital EHR system are already FHIR compliant.
Extracting data from other clinical databases as input to the CDM	Python ETL tool with specific mapperrs - will depend on each specific data provider	**whiterabbit** (summary reports informing the design of the ETL pipeline), **rabbit in a hat** (syntactic mappings), **usagi** (mappings between vocabularies), **Achilles** (assessing mapping quality)	
Outputting data in a format suitable for AI algorithms	Python tool with Spark that will generate csv files	SQL queries for accessing the required information, and then map it to the input format required by AI algorithms (e.g., csv)	Using **protobuf** in python and getting a python object where one can access the value. It would be needed a mapper to input for AI algorithms.

Table 10 Proposed strategies for representing information about data governance and auditing.

Representing information about data governance and auditing			
	FHIR	**OMOP-CDM**	**Phenopacket**
Data	Data for specific	Add specific metadata	Storing query returning the

corresponding to specific AI algorithms	algorithms should match some filters, that will identify the data set used for each algorithm.	attributes to the Metadata OMOP-CDM table.	dataset in a text or some other format. However, this might be too ambitious and creating an additional table for each AI algorithm with the IDs of each patient and the feature sets might be more feasible, in the short-middle term.
Auditing information	FHIR contains metadata that can contain some of this information.	Metadata table for storing metadata information about a dataset, and Cdm_Source for storing detail about the source database and the process used to transform the data into the OMOP-CDM.	This should be handled by the platform, so that it logs all the update operations.
Contributing institution for each data sample	Each resource has an identifier field, that is a list of ids. Each identifier has a reference to the Organization.	Care_Site (for institutions), Provider (for individuals).	There are some description metadata fields which can store this kind of information.

---

1 Van Panhuis, Willem G., et al. "A systematic review of barriers to data sharing in public health." *BMC public health* 14.1 (2014): 1-9.

2 Rieke, Nicola, et al. "The future of digital health with federated learning." NPJ digital medicine 3.1 (2020): 1-7.

3 Xu, Jie, et al. "Federated learning for healthcare informatics." Journal of Healthcare Informatics Research 5.1 (2021): 1-19.

4 Dugas, Martin, et al. "Portal of medical data models: information infrastructure for medical research and healthcare." Database 2016 (2016).

5 Weeks, John, and Roy Pardee. "Learning to share health care data: a brief timeline of influential common data models and distributed health data networks in US health care research." EGEMs 7.1 (2019).

6 Ross, Tyler R., et al. "The HMO research network virtual data warehouse: a public data model to support collaboration." Egems 2.1 (2014).

7 Bachman, Donald, Pierre-Andre La Chance, and Mark Hornbrook. "Ps1-28: Kaiser permanente center for effectiveness and safety research." Clinical medicine & research 8.3-4 (2010): 207-207.

8 Dolin, Robert H., et al. "HL7 clinical document architecture, release 2." Journal of the American Medical Informatics Association 13.1 (2006): 30-39.

9 Kohl, Christian D., Sebastian Garde, and Petra Knaup. "Facilitating secondary use of medical data by using openEHR archetypes." MEDINFO 2010. IOS Press, 2010. 1117-1121.

10 Mustra, Mario, Kresimir Delac, and Mislav Grgic. "Overview of the DICOM standard." 2008 50th International Symposium ELMAR. Vol. 1. IEEE, 2008.

11 Curtis, Lesley H., et al. "Design considerations, architecture, and use of the Mini-Sentinel distributed data system." Pharmacoepidemiology and drug safety 21 (2012): 23-31.

12 International Organization for Standardization. “Identification of medicinal products”. ISO 11615 (2017).

13 Welcome to FHIR. HL7 International. URL: https://www.hl7.org/fhir/ [accessed 2022-06-10]

14 Overhage, J. Marc, et al. "Validation of a common data model for active safety surveillance research." Journal of the American Medical Informatics Association 19.1 (2012): 54-60.

15 Murphy, Shawn N., et al. "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)." Journal of the American Medical Informatics Association 17.2 (2010): 124-130.

16 Fleurence, Rachael L., et al. "Launching PCORnet, a national patient-centered clinical research network." *Journal of the American Medical Informatics Association* 21.4 (2014): 578-582.

17 Kuchinke, Wolfgang, et al. "CDISC standard-based electronic archiving of clinical trials." *Methods of information in medicine* 48.05 (2009): 408-413.

18 Lo, Sin Kit, et al. "Architectural patterns for the design of federated learning systems." arXiv preprint arXiv:2101.02373 (2021).

19 Prokosch, Hans-Ulrich, and Thomas Ganslandt. "Perspectives for medical informatics." Methods of information in medicine 48.01 (2009): 38-44.

20 Shinozaki, Ayaka. "Electronic medical records and machine learning in approaches to drug development." Artificial intelligence in Oncology drug discovery and development. IntechOpen, 2020.

21 Choudhury, Ananya, et al. "Personal health train on fhir: A privacy preserving federated approach for analyzing fair data in healthcare." *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*. Springer, Singapore, 2020.

22 Ahmadi, Najia, et al. "OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review." International journal of molecular sciences 23.19 (2022): 11834.

23 Khalilia, Mohammed, et al. "Clinical predictive modeling development and deployment through FHIR web services." *AMIA Annual Symposium Proceedings*. Vol. 2015. American Medical Informatics Association, 2015.

24 BATHELT, Franziska. "The usage of OHDSI OMOP–a scoping review." *Proceedings of the German Medical Data Sciences (GMDS)* (2021): 95-95.

25 Jacobsen, Julius OB, et al. "The GA4GH Phenopacket schema: A computable representation of clinical data for precision medicine." medRxiv (2021).

26 Jochems, A., et al. “Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept”. Radiotherapy and Oncology, 121(3), 459–467. (2016). https://doi.org/10.1016/j.radonc.2016.10.002

27 Deist, Timo M., et al. “Infrastructure and Distributed Learning Methodology for Privacy-Preserving Multi-Centric Rapid Learning Health Care: EuroCAT.” Clinical and Translational Radiation Oncology, vol. 4, June (2017), pp. 24–31, https://doi.org/10.1016/j.ctro.2016.12.004.

28 Damiani, Andrea, et al. “Distributed Learning to Protect Privacy in Multi-Centric Clinical Studies.” Artificial Intelligence in Medicine, edited by John H. Holmes et al., Springer International Publishing, 2015, pp. 65–75, https://doi.org/10.1007/978-3-319-19551-3_8.

29 Friedman, Charles P., Adam K. Wong, and David Blumenthal. "Achieving a nationwide learning health system." Science translational medicine 2.57 (2010).

30 Crowson, Matthew G., et al. "A systematic review of federated learning applications for biomedical data." PLOS Digital Health 1.5 (2022).

31 Antunes, Rodolfo Stoffel, et al. "Federated Learning for Healthcare: Systematic Review and Architecture Proposal." ACM Transactions on Intelligent Systems and Technology (TIST) 13.4 (2022): 1-23.

32 Gruendner, Julian, et al. “KETOS: Clinical Decision Support and Machine Learning as a Service – A Training and Deployment Platform Based on Docker, OMOP-CDM, and FHIR Web Services.” *PLoS ONE*, vol. 14, no. 10, Oct. 2019, p. e0223010, https://doi.org/10.1371/journal.pone.0223010.

33 Khalilia, Mohammed, et al. “Clinical Predictive Modeling Development and Deployment through FHIR Web Services.” *AMIA Annual Symposium Proceedings*, vol. 2015, Nov. 2015, pp. 717–26, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765683/.

34 Lo, Sin Kit, et al. “A Systematic Literature Review on Federated Machine Learning: From a Software Engineering Perspective.” ACM Computing Surveys, vol. 54, no. 5, (2021), pp. 1–39, https://doi.org/10.1145/3450288.

35 Kairouz, Peter, et al. “Advances and Open Problems in Federated Learning.” *ArXiv:1912.04977 [Cs, Stat]*, Mar. 2021, http://arxiv.org/abs/1912.04977.

36 Huser V, Kahn MG, Brown JS, Gouripeddi R. Methods for examining data quality in healthcare integrated data repositories. Pac Symp Biocomput. 2018;23:628–33.

37 Zhou, Zirui, et al. "Towards fair federated learning." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021.

38 Yoo, Junsang, et al. "Development of an Interoperable and Easily Transferable Clinical Decision Support System Deployment Platform: System Design and Development Study." *Journal of Medical Internet Research* 24.7 (2022): e37928.

39 Genereaux, Brad, et al. " IHE Radiology White Paper - AI Interoperability in Imaging " White Paper. (2021)

40 Vogelsang, Andreas, and Markus Borg. "Requirements engineering for machine learning: Perspectives from data scientists." *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2019.

41 Bittner, Kurt, and Ian Spence. *Use case modeling*. Addison-Wesley Professional, 2003.

42 http://build.fhir.org/ig/HL7/genomics-reporting/sequencing.html#variant-confidence. Accessed: 17 November 2022

43 Alterovitz, G, et al. “Enabling clinical genomics for precision medicine via HL7 fast healthcare interoperability resources.” Sync for Genes report for the Office of the National Coordinator for Health Information Technology (2017). https://www.healthit.gov/sites/default/files/sync_for_genes_report_november_2017.pdf

44 Rath, Ana, et al. "Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users." Human mutation 33.5 (2012): 803-808.

45 Rab, Minke AE, et al. "Rapid and reproducible characterization of sickling during automated deoxygenation in sickle cell disease patients." American journal of hematology 94.5 (2019): 575-584.