

I-ETL: an interoperability-aware health (meta)data pipeline to enable federated analyses

Nelly Barret¹, Anna Bernasconi¹, Boris Bikbov¹, Pietro Pinoli^{1*}

¹Department of Electronics, Information and Bioengineering – Politecnico di Milano, Italy.

*Corresponding author(s). E-mail(s): pietro.pinoli@polimi.it; Contributing authors: nelly.barret@polimi.it; anna.bernasconi@polimi.it; boris.bikbov@polimi.it;

Abstract

Background: Clinicians are interested in better understanding complex diseases, such as cancer or rare diseases, so they need to produce and exchange data to mutualize sources and join forces. To do so and ensure privacy, a natural way consists in using a decentralized architecture and Federated Learning algorithms. This ensures that data stays in the organization in which it has been collected, but requires data to be collected in similar settings and similar models. In practice, this is often not the case because healthcare institutions work individually with different representations and raw data; they do not have means to normalize their data, and even less to do so across centers. For instance, clinicians have at hand phenotypic, clinical, imaging and genomic data (each individually collected) and want to better understand some diseases by analyzing them together. This example highlights the needs and challenges for a cooperative use of this wealth of information.

Methods: We designed and implemented a framework, named I-ETL, for integrating highly heterogeneous healthcare datasets of hospitals in interoperable databases. Our proposal is twofold: (*i*) we devise two general and extensible conceptual models for modeling both data and metadata and (*ii*) we propose an Extract-Transform-Load (ETL) pipeline ensuring and assessing interoperability from the start.

Results: By conducting experiments on open-source datasets, we show that I-ETL succeeds in representing various health datasets in a unified way thanks to our two general conceptual models. Next, we demonstrate the importance of blending interoperability as a first-class citizen in integration pipelines, ensuring possible collaboration between different centers.

Conclusions: As a framework, I-ETL contributes to integrate and improve interoperability between healthcare institutions. When used in a decentralized federated platform, it eases the federated analysis of the different hospital databases and helps clinicians to obtain insights and knowledge on medical conditions of interest.

Keywords: heterogeneous data, healthcare, data model, FAIR principles, federated learning

1 Background

The overall recent digitization of the healthcare sector has led to new opportunities for researchers and clinicians to access, integrate, analyze, share, and reuse medical data [1]. An example of this is the emergence of self-tracking tools, which are source of plenty of healthcare data for various profiles of citizens [2]. Plenty of other initiatives have been conducted and deployed at different levels, from city- and national-, to European- and world-wide platforms, and on diverse healthcare major open issues (access to healthcare, cancer, genetic rare diseases, etc). For instance, the World Health Organization (WHO) collects world-wide data in their data hub [3] on various healthcare domains, including COVID-19, mortality and access to healthcare, and computes yearly reports [4] after integrating and analyzing it. Another example is the European Union, which heavily invests in rare diseases investigation [5] since 2007, notably to develop new tools to decipher them, as they remain largely poorly understood. As of today, around 6,000 rare diseases are known and 80% of them are of genetic origin, most of which have no effective treatment or allow for easy diagnosis. Conducting data-driven research in the context of rare diseases is evidently challenging, as assembling a sufficiently large dataset would require aggregating medical and genetic data from patients across numerous geographically distributed clinical institutions; however, current regulations, such as the General Data Protection Regulation (GDPR), impede the aggregation of sensitive data in a centralized repository. In response to this, the scientific community is developing tools and solutions to analyze medical and genetic data in a federated and secure manner. These tools do not require the exchange of real data but solely aggregate extracted information (e.g., statistics or partial models) computed from local datasets. Performing federated analyses allows the study of larger sets of patients, potentially with heterogeneous types of data, thus obtaining more accurate results and insights.

However, from a data perspective, federated analyses require a significant effort in standardization and harmonization to ensure interoperability among distributed datasets. It is clear that manual curation and cleaning are not feasible or scalable. Therefore, collaborating hospitals crucially need tools to automatically integrate heterogeneous and sensitive datasets with interoperability as a first-class citizen.

1.1 Conceptual models for healthcare data

Several large projects have proposed instruments for modeling and enforcing interoperability of distributed heterogeneous healthcare datasets, intending to facilitate federated analyses. Notably, the EHDEN project [6, 7] safely integrates hundreds of relational (tabular) datasets across Europe to provide clinicians a single endpoint where they can select datasets of interest for observational studies, define protocols and run analyses. For unifying the different concepts across tables, they used the *OMOP* [8] (Observational Medical Outcomes Partnership) conceptual model as a Common Data Model (CDM). Similarly, [9] converts European healthcare data to the FHIR [10] data model in order to define Deep Learning models for early-diagnosis of children and teenagers regarding non-communicable chronic diseases. There exist many widely-adopted CDMs with the same intent of *OMOP* and *FHIR*; the most significant are i2b2 [11] for precision medicine, *GCM* [12, 13] for genomic data (developed within the GeCo ERC AdG project [14]) and the *Human Cell Atlas* [15] for human cells, but also general ones such as *openEHR* [16], and *SMART* [17].

To fit the input data into a CDM of choice, it is common to rely on an ETL pipeline (Extract-Transform-Load). This is a three-step process where data is extracted from input sources, transformed/cleaned, and loaded into a target data container, usually a database, whose schema is the CDM. D-ETL [18] is a dynamic ETL pipeline partially automating the process by providing data harmonization techniques and simplifying the transformation process. In turn, experts specify “ETL structured rules” for mapping the actual input data to the CDM. Then, these rules are transformed into SQL statements and data is loaded into the target database. For the more general FHIR standard, [19] proposes a 6-step workflow to help healthcare scientists to model their data with FHIR entities – this is a more conceptual approach of the ETL process.

1.2 Metadata for better interoperability

To allow interoperability among datasets situated at different institutions, it is crucial to accurately describe each piece of information using unambiguous metadata. **Metadata** is supplementary information that allows the assignment of meaning to both the type of data and its associated value. While metadata can be of multiple forms, in clinical and biomedical settings it is a best practice to (re)use existing ontologies that describe data concepts. Many specialized ontologies exist, such as SNOMED-CT [20] for general-purpose healthcare terms, LOINC [21, 22] for clinical measurements, or OrphaNet [23] for disease classification. By using ontologies, data points can be mapped to unique ontology resources, thus enabling interoperability among datasets of interest.

Moreover, metadata are essential to interoperability, which is part of the “I” prescription of the FAIR principles [24], a set of guidelines to make data and its accompanying metadata Findable, Accessible, Interoperable, and Reusable. In general, FAIR principles encourage IT experts to identify their resources with unique and reusable identifiers (e.g., URIs), use widely used standards and protocols, and provide rich metadata by reusing existing ontologies. They are, by definition, general enough to allow their adoption by very heterogeneous projects. Therefore, many healthcare

integration systems have been designed with FAIR principles in mind; this is the case of UMG-MeDIC [25] and Scaleus-FD [26].

Although the principles of FAIRness are applied during dataset processing, it is also important to assess whether the dataset (or other digital object, such as a database) itself meets the FAIR criteria. According to a recent survey [27], existing assessment tools are often tied to a given context, may involve manual assessment, and often focus solely on data FAIRness, not considering metadata and semantics FAIRness. A possible solution to overcome these limitations is to integrate FAIRness assessment from the start and provide explainable metrics to users, to let them improve the quality (and FAIRness) of their data and metadata.

1.3 Limitations and contributions

As illustrated before, many approaches have been proposed to model healthcare datasets that are very heterogeneous (in their model and content) and to enforce interoperability. Nevertheless, they exhibit some limitations.

First, many of them are tied to a single data model (e.g., EHDEN [6, 7] only integrates tabular data) or leverage a CDM tied to an healthcare domain (e.g., OMOP [8] models observational data only and is hardly extensible to other domains like genomic or imaging data). Second, they all exhibit entities with specific attributes (see, for instance, the FHIR *Observation* entity with its 24 attributes, including `reason`, `status`, `subject`, and `value`). In turn, these models lead to hand-made integration workflows (to map data concepts with those in the model and to adapt to specific attributes) or very abstract workflows. As an example, Dynamic-ETL [18] (a data integration pipeline) requires experts to write rules expressing how to match source models to OMOP. On the contrary, [19] proposes a FHIR-based workflow to transform existing medical data to FHIR. However, it does not include practical steps to realize the data integration part. The above reasons and examples emphasize why existing models are not convenient for the transformation of existing data to those models (while we note that they are suited for collecting new data). The main limitations include the attribute-based models leading to specific pipelines that are hardly reusable, and the technical barriers (e.g., map input and target models, and write corresponding data rules) that experts may not be able to surpass. Finally, FAIR metrics must be computed along ETL pipelines to ensure high findability, accessibility, interoperability, and reusability, while most works focus on assessing the interoperability on the obtained data [27].

In this work, we propose **I-ETL**, a novel framework to enforce interoperability among heterogeneous distributed healthcare datasets. It ensures privacy, requires moderate input from clinical experts, and computes a holistic interoperability assessment. Our main contributions are:

1. Two tightly linked, extensible, **conceptual models for both metadata and data**, based on experts' knowledge, for achieving data interoperability within and across hospitals' data stores.
2. An **ETL pipeline** in which interoperability is a first-class citizen, producing a target database to be used in federated and distributed analytics contexts.

3. Guarantee of **interoperability** using a set of metrics that are progressively assessed during I-ETL.

2 Methods

2.1 I-ETL approach and concepts

The proposed I-ETL approach is a 5-step data science pipeline (see Figure 1). Starting from the left side, medical experts decide collaboratively on a specific topic to study, e.g., pediatric intelligence disability or kidney cancer, and *select a set of datasets* relevant to their chosen topic. A **dataset** is any sort of file containing data about patients; it may be, for instance, a CSV file containing the clinical measures obtained from blood samples, a DICOM file obtained from an eye MRI scan, a VCF file comprising patients' genomic variants, etc.

Next, after inspecting the selected datasets, practitioners *define a set of relevant features* F in those datasets. In this work, a **feature** is a specific attribute of data, e.g., the birth date of a patient, the size of the dark regions in MRI eye scans, etc.

Keeping F in mind, practitioners *fill the metadata* M of the chosen datasets. We define **metadata** as any information providing the context to understand and interpret a feature. Metadata typically includes the feature name, its data type, possibly an ontology resource that could be associated with it, etc. (details on our metadata model are given in Section 2.2).

The fourth step in the global pipeline is the ETL process (later described in Section 2.4). In a nutshell, this aims at transforming the input datasets into a target database whose schema is our general conceptual model for healthcare datasets (presented in Section 2.3), enabling interoperability between the ingested datasets, but also with the other databases used in the federated analysis task. During the pipeline, interoperability metrics are collected (see Section 2.5) and are then reported to practitioners.

2.2 Metadata model

After picking relevant datasets for a use case of interest, practitioners have to select or extract a set of relevant features in their datasets. However, simply selecting (or extracting) a set of relevant features is not sufficient because individual datasets, or hospitals, may encode similar features with different names, thus prohibiting interoperability. To overcome this semantic issue, we ask experts to specify which ontology resource may be used to represent each feature. This mapping of features to existing ontology resources, done during the metadata creation, improves interoperability and alignment between hospitals. Unmapped features can still be referred to by their original names, even though this considerably hinders interoperability.

The following list depicts our one-entity **metadata model** for describing datasets with interoperability as a first-class citizen:

- **Name:** the name of the feature;
- **Ontology:** the name of the ontology chosen to represent the feature concept where ontologies can be selected in well-known portals like BioPortal [28];

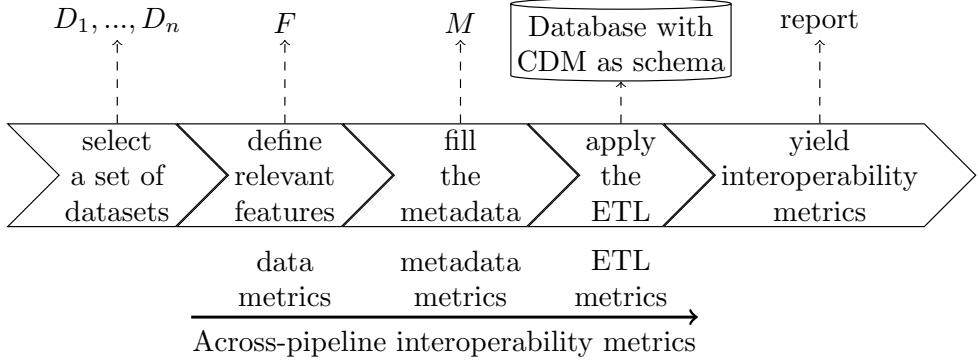


Fig. 1 I-ETL, the framework building interoperable databases for federated analyses from heterogeneous healthcare data. Large arrows represent steps in the framework; dashed edges connect a step to its corresponding (intermediate) result. Interoperability metrics are shown below the task during which they are computed.

- **Code**: the code of the resource in the selected ontology that represents the feature;
- **Kind**: the type of feature, i.e., phenotypic, clinical, genomic, image, etc;
- **DataType**: the expected value type (among `string`, `integer`, `numeric`, `boolean`, `category`, `date`, or `datetime`);
- **Unit**: the unit to interpret the values when the feature data type is `integer` or `numeric`;
- **Categories**: when the feature is categorical, each value is paired to an ontology resource (a pair of the ontology name and a code).
- **Visibility**: whether the values for that feature can be shown publicly (`public`), after anonymization (`anonymized`), or cannot be shown at all (`private`), depending on their sensitivity.

Example instances. Assume we have two phenotypic features selected from datasets for the kidney disease scenario, namely age and sex:

- The first feature is instantiated with the following values: `name` is “`calc_age`”, `ontology` and `code` are SNOMED-CT and “`397669002`”, `dataType` is `integer`, `unit` is “`years`” while `categories` is null (the feature is not categorical), and the `visibility` is `anonymized` to prevent the access to the original patient age (the age could be deduced if the patient is younger or older than the average).
- The second feature, about the sex, is instantiated with `name` “`sex`”, `ontology` is SNOMED-CT with the `code` “`734000001`”. The `dataType` is `category`, there will be no `unit` and the `visibility` would be `public`. For `categories`, we have two pairs: (`Female`, (SNOMED-CT, `248152002`)) and (`Male`, (SNOMED-CT, `248153007`)).

2.3 Common data model

After describing features with metadata, we run the ETL pipeline to transform the input datasets in a homogenized, interoperable target database. This relies on a **common data model** (CDM in short), which is a conceptual model for representing homogeneously and making all the selected data interoperable. Each database included in a federated scenario, implements the same CDM so that all databases can cooperate, be accessed, and queried in a unified way, regardless of the heterogeneous datasets they carried in origin.

Our CDM is illustrated in Figure 2; it promotes two important characteristics. First, it isolates medical knowledge and concepts specified in the metadata from the actual data instances by distinguishing two concepts: **features and records**. These two concepts are abstractions of how any dataset can be represented and make our model a common data model (as opposed to a project-based data model). As previously defined, a FEATURE is an attribute of the data, e.g., the birth date of a patient. Instead, a RECORD is the actual value that a patient has for a certain feature, as produced by an hospital. For instance, according to hospital H1, for patient P1 and a feature whose *name* is “birth_date” the value is “01/01/2000”. Second, it describes six different kinds of data (phenotypic, clinical, medicine, diagnosis, genomic, and imaging), while being easily extensible to new types of data, e.g., administrative data or patient history. This also contributes to make our model general enough to be considered as a common data model, possibly reused by many other diverse projects whose goals are to model very heterogeneous healthcare datasets in various settings.

In our conceptual model (Figure 2), we follow the typical notation of Entity-Relationship diagrams [29] established in [30]. The central entity is the RECORD; each record has a unique *identifier*, a *value*, and the name of the *dataset* to which it belongs. The *value* is of type *any* as records may capture not only atomic values, such as integer or boolean, but also complex ones, e.g., categories. The *value* can be further anonymized depending on its sensitivity (see Section 2.4). The attribute *dataset* allows to keep track of the provenance, i.e., from where individual values come from.

Since clinicians need various kinds of data when studying a research matter, the RECORD entity is specialized in 6 entities, namely phenotypic, clinical, medicine, diagnosis, genomic, or imaging, each of which inherits the RECORD attributes. Note that this set of data kinds can be easily extended or modified for different scenarios.

When a patient has an appointment, practitioners first collect phenotypic data, i.e., information about the patient, the environment and habits, each value leading to a PHENOTYPICRECORD. Next, patients usually go through a series of tests in laboratories, including blood tests, breathing tests, skin tests, etc. For each test, captured values become CLINICALRECORD instances. If the clinical record is associated with a sample, the sample identifier is stored in the optional attribute *baseId*. Genome sequences of patients are aligned on a default genome to obtain genomic variants. Those variants and their associated information, such as the chromosome on which it appears or the confidence, are represented with the GENOMICRECORD entity. The genomic record also saves the VCF file path from which the value has been extracted with the attribute *vcf*. Moreover, some patients are also asked to go for MRI scans to better visualize areas of interest. Each MRI scan is an image, from which a set

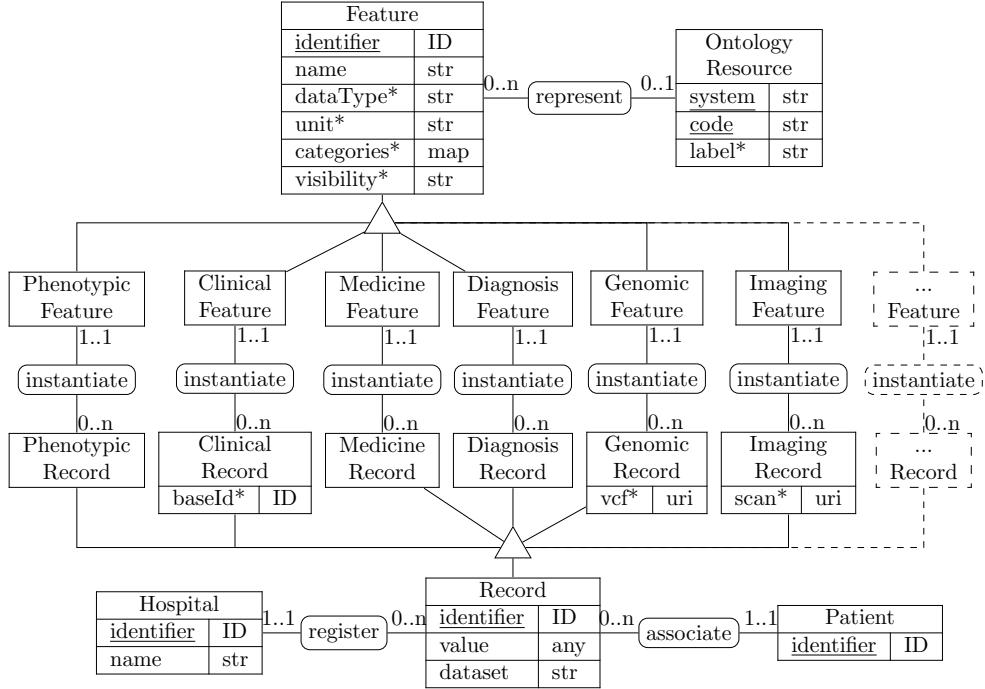


Fig. 2 The interoperable conceptual model instantiated at each medical center database. Rectangles are entities, rounded boxes are relationships and triangles are specializations. Primary keys are underlined, and optional attributes are marked with a * (star). Our cardinalities adopt the notation in [29], e.g., a Record instantiates exactly one Feature, is associated with exactly one Patient, and is registered by exactly one Hospital. Features can be instantiated in 0 to n Record entities.

of features are extracted and each value is stored as an IMAGINGRECORD. As for genomic records, imaging records save the image file path from which the value has been extracted using the attribute *scan*. For patients taking medicines, each information in a prescription is seen as a MEDICINERECORD. Finally, patients may obtain a diagnosis, leading to a DIAGNOSISRECORD.

Each RECORD is registered by exactly one HOSPITAL, i.e., a clinical institution contributing to the federated analysis task. It has an *identifier* and a *name*. A RECORD has for subject exactly one PATIENT, i.e., an individual included in the studied cohort. Patients only have an *identifier* because they are totally anonymized for privacy, thus, no further personal information can be included.

Each RECORD entity instantiates exactly one FEATURE of the same kind, e.g., a phenotypic record instantiates a phenotypic feature. Those features are specializations of the FEATURE entity, carrying all the attributes. Each feature has an *identifier*; other attributes come from the metadata model. In turn, each feature has a *name*, a *dataType*, a *unit*, a set of *categories*, and a *visibility*. The last four attributes are optional because: the *dataType* and the *visibility* may not be specified by medical

experts (*visibility* defaults to `private` for privacy reasons), the *unit* exists only for numerical features, and the list of *categories* only exists for categorical features.

A `FEATURE` is represented by zero or one `ONTOLOGYRESOURCE`. Sometimes, features cannot be formalized in any ontology (hence the cardinality $0..1$). It contains a *system*, i.e., the endpoint URL to access the ontology, a *code*, i.e., a unique identifier for the represented concept in that ontology, and a *label*, a human-friendly name for that concept. All of them are strings, only *system* and *code* are mandatory (in order to identify the represented concept).

A `PHENOTYPICFEATURE` captures any variable about the factors that may affect the patient, such as the environment, daily routine, habits, etc. A `CLINICALFEATURE` may be about any chemical or clinical measurement. A `GENOMICFEATURE` typically concerns the chromosomes carrying variants, the exact position of the variant, the type of variation (addition, deletion, mutation), etc. `IMAGINGFEATURE` instances can represent diverse features depending on the patient’s disease. For instance, if the patient is affected by a brain tumor, the features may include the coordinates of the tumor in the MRI scan, the darkness of the tumor in the scan, etc. Additional features about the scanner and software can be included too, e.g., the software version, the scanner name, etc. A `MEDICINEFEATURE` captures variables such as the name of the medication, the start and length of the prescription, whether the patient took the medicine, etc. Finally, a `DIAGNOSISFEATURE` captures a characteristic of the diagnosis attributed to a patient. Diagnosis characteristics include the diagnosis name, but also the affected gene, whether the patient is a carrier or affected, etc.

2.4 ETL pipeline and target database

Our three-step ETL pipeline leverages the CDM for integrating the input data in the target database, as shown in Figure 3. It takes as input the selected datasets D_1, \dots, D_n and their accompanying metadata M . Next, the three steps Extract, Transform and Load are performed, as detailed in the following. While performing them, a number of metrics for assessing interoperability are computed to keep track of interoperability from the start to the very end of the process.

Extract step. The input data, i.e., the metadata M specified by clinical experts and the datasets D_1, \dots, D_n , is read and stored in memory. No normalization is applied at this stage (postponed to the Transform step).

Transform step. This step creates new objects (instances) that will fit the CDM, i.e., the conceptual model of the target database. While creating the objects, it also ensures that they are interoperable with each other, e.g., by applying interoperability implementation techniques to values. We proceed as follows:

1. A set of `PATIENT` instances is created as the union of all the patients in the input datasets. Each patient is anonymized by assigning an identifier with the scheme `<HospitalName:counter>`, where `HospitalName` is the hospital name and `counter` is an auto-incremented number. No further personal information is stored for patients in the CDM for privacy reasons.

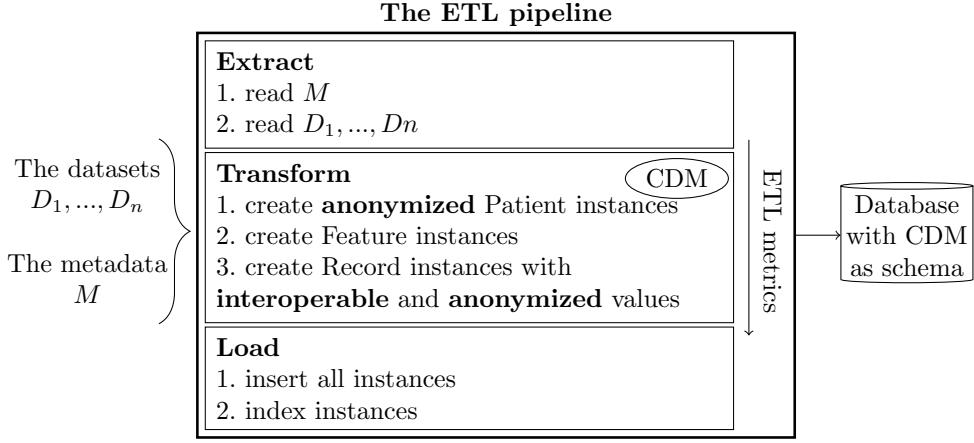


Fig. 3 The ETL pipeline for integrating diverse and heterogeneous datasets, keeping interoperability as a first-class citizen.

2. Each feature $f \in F$ is created based on the available information in the metadata M and is specialized based on its kind. An identifier of the form $\langle \text{Feature} : \text{counter} \rangle$ is assigned to each of them. If a feature presents both ontology name and code in its metadata, then an ONTOLOGYRESOURCE is created, respectively with the ontology name and code in attributes *system* and *code*. A query asking for the ontology resource label is sent to the ontology. If a non-empty result is returned, it is stored in the attribute *label*, otherwise the attribute remains empty. Next, the attributes *dataType*, *unit* and *visibility* are directly obtained from the metadata. Finally, the attribute *categories* is computed as a map containing pairs of a categorical value (a string) and the associated ONTOLOGYRESOURCE instance (if it exists).
3. RECORD instances are created out of the input datasets D_1, \dots, D_n . For each patient having a non-empty value v for a given feature of kind k , a k RECORD is created, and the attribute *value* stores v .

The *dataset* attribute is set to the examined dataset D_i . A unique identifier is assigned to the new instance, namely an identifier of the form $\langle \text{Record} : \text{counter} \rangle$. The created record also includes three foreign keys: *register*, *associate*, and *instantiate*, which are identifiers of existing HOSPITAL, PATIENT, and FEATURE instances. Specific attributes (e.g., the *baseId* for CLINICALRECORD instances) are extracted from the dataset while creating the records.

Finally, before setting the *value* with v , we make the value as **interoperable** and **secure** as possible by proceeding as follows.

Interoperable. Each value v undergoes transformations to **enhance its interoperability**, as detailed in Table 1. Each technique depends on v 's data type; when not applicable, the related functions return the initial value (e.g., string “tru” cannot be cast to a boolean, thus the function yields “tru”).

Secure. Then, we **secure** the interoperable (or initial) value based on the feature *visibility* and *dataType* attributes. When the visibility equals **anonymized**, **date**

Data type	Function to enable interoperability
<code>string</code>	Normalize v (trim spaces and lower case v)
<code>category</code>	Yield the associated OntologyResource in <code>categories</code>
<code>date</code>	Cast v to an ISO format
<code>datetime</code>	Cast v to an ISO format
<code>boolean</code>	Cast v to <code>True</code> or <code>False</code>
<code>integer/numeric</code> and $f_{unit} = v_{unit} = \emptyset$	Cast v to an <code>integer</code> , respectively a <code>numeric</code>
<code>integer/numeric</code> and $f_{unit} = v_{unit}$	Remove the unit and cast v (to <code>int.</code> , resp. <code>num.</code>)
<code>integer/numeric</code> and $f_{unit} \neq v_{unit}$	Yield v

Table 1 Functions used on Record values to enhance their interoperability. If the technique does not succeed on the input value, the function returns the value as is.

and `datetime` values are deprived from their day, respectively minutes and seconds. Otherwise (the `visibility` is `private` or `public`, or the `dataType` is not `date` nor `datetime`), the value v is returned as is.

Load step. The last step of the ETL process is to load the instances created in memory into the target database. This includes: the HOSPITAL, PATIENT, FEATURE, and RECORD instances. After insertion, instances are indexed to ensure fast access and querying.

2.5 Interoperability assessment

The I-ETL framework finally reports a set of interoperability metrics that are computed across the integration pipeline. Concerning databases that are created using I-ETL, interoperability can be measured at three levels: the database is compared with itself (we call this *self-interoperability*); with other databases in the same institution (termed *intra-institution interoperability*); and with the databases of other institutions (termed *inter-institution interoperability*). For databases generated through I-ETL that received high metrics scores, we ensure interoperability on all three levels. These three levels of interoperability can be ensured both from a syntactic and semantic point of view. All data providers use the same data model (the one presented in Section 2.3) and this guarantees syntactic interoperability. For semantic interoperability, we favor it by allowing experts to map the features of interest to well-known and widely adopted ontologies.

We derived a set of **interoperability metrics**, which -respectively- target the selected data, the created metadata, and the ETL process. Specifically, **data metrics** are computed to assess the completeness of datasets selected for a specific federated scenario; **metadata metrics** assess the coherence and completeness of the metadata created for the selected datasets; and **ETL metrics** assess to what extent interoperability and anonymization have been achieved during the ETL process, as well as the coherence of the target database.

It is crucial that FAIR principles (specifically, interoperability) are put from the beginning of the process that achieves the database creation. Table 2 lists our metrics, which are computed from the start to the end to evaluate self-, intra-institution and inter-institution interoperability. The next sections detail the three sets of metrics.

Step	Metric
Data	(A1) Ratio of selected features (A2) Ratio of datasets that do not require dedicated extraction
Metadata	(M1) Features with both non-empty ontology <i>name</i> and <i>code</i> (M2) Features with non-empty <i>dataType</i> (M3) Features with non-empty <i>visibility</i> (M4) Categorical features with non-empty set of <i>categories</i> (M5) Numerical features with non-empty <i>unit</i>
ETL	(E1) Presence of non-empty <i>label</i> in Ontology Resource (E2) Values for which interoperability implementation has succeeded (E3) Correspondence of numerical values <i>unit</i> and Feature <i>unit</i> (E4) Presence of categorical value in the Feature <i>categories</i> (E5) Records with known Hospital references (E6) Records with known Patient references (E7) Records with known Feature references

Table 2 The set of metrics recorded through I-ETL to assess interoperability across the pipeline.

2.5.1 Data metrics

Data metrics are computed on the selected features F for datasets D_1, \dots, D_n ; we defined two. (A1) represents how many features have been retained for the selected datasets, providing insight of the extent to which the datasets are mapped to metadata. Small subsets of (well-crafted) features are generally preferable in federated scenarios, even if possible information loss must be taken into account. (A2) represents the number of datasets that do not require extraction using dedicated tools. Dedicated data extraction, defined as the process of identifying and retrieving information using external tools, is often needed with complex data types such as images and genomic files. For instance, MRI scans are images for which dedicated extraction is needed in order to obtain data from the image itself. Examples of data extracted from lung MRI scans include the size of dark regions in the scan, whether abnormalities can be seen, etc. In general, using dedicated tools may lead to an overall decrease in interoperability, e.g., when different versions or pipelines are used and/or if images are of low quality.

2.5.2 Metadata metrics

For metadata metrics, we evaluate the features F described by the clinicians in the metadata M (Figure 1, step 3). For all features, we count how many have (M1) both *ontology* and *code*, (M2) non-empty *dataType*, and (M3) non-empty *visibility*. While these three attributes are not mandatory, they are important to achieve high intra- and inter-institution interoperability, crucial for the federated analysis of the datasets within an institution and with other institutions.

Next, only for categorical features, we count how many of them have at least one associated category (M4). Values that cannot be mapped to the specified categories decrease interoperability. Finally, we count numerical features that are equipped with

a valid *unit* (M5); note that features without this information hinder the interpretation of values (consider, as an example, the ambiguity in determining whether age is expressed in weeks or months).

2.5.3 ETL metrics

ETL metrics are computed during the Transform step while making features and values as interoperable as possible. We compute (E1) as the number of ontology resources (whether they represent a Feature or a categorical value) having a non-empty *label*. (E1) captures the interoperability of both features and categorical values because it is of the same importance if one of them carries an empty label (both hinder interoperability). The label is empty if the ontology is not provided with a query interface or if the request sent to the ontology fails for any reason. For instance, the LOINC resource, whose code is LL4034-6, does exist at URL <https://loinc.org/LL4034-6>, but the query asking for information about this resource returns an empty result, because the LOINC query interface does not access the latest release of the ontology. At this point of the ETL, failing queries affect only the OntologyResources' *label*; however, if one wants to retrieve more information for that specific resource, it will not be possible, thus limiting interoperability.

Next, (E2) counts the ratio of Record values for which interoperability implementation has succeeded (recall Table 1); (E3) counts the ratio of numerical Record values having the same unit as the Feature they instantiate; with (E4), we assess the quality of categorical values by checking how many of them were declared in the Feature *categories* list that they instantiate. Overall, (E2), (E3), and (E4) measure the coherence between metadata and data.

Finally, (E5), (E6), and (E7) ensure that all created instances refer to (other) objects existing in the database. This is achieved by counting how many references to hospitals, patients, and features point to instances in the database, preventing broken references.

2.5.4 Anchoring metrics to FAIR principles

To legitimate our metrics w.r.t. the FAIR principles, we recall the three sub-principles of Interoperability and explain how our proposed pipeline and metrics conform to them.

I1. Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

Our conceptual common data model is designed in a way such that it can be implemented within any type of database (relational, NoSQL, etc). Moreover, metadata can be easily specified using a tabular file, such as an Excel or CSV file, while following our metadata model. The selected datasets can be of any machine-readable format and dedicated extraction is applied for datasets necessitating pre-processing.

I2. Data and metadata use vocabularies that follow FAIR principles.

While specifying metadata, experts are asked to assign an ontology resource to each feature of interest. Despite not mandatory, it is highly recommended in order to achieve

good interoperability. Experts are also advised to use well-known and recognized ontologies in their domain. At best, the ontology is publicly accessible; otherwise, it may need to be accessed under authentication, but should not be hospital-dependent. The ETL pipeline will reuse these ontologies to encode categorical values, thus increasing interoperability both for metadata and data. This aspect is tracked by ETL metrics assessing ontology resources and categorical values.

I3. Data and metadata include qualified references to other data and metadata.

Each Record instance comprises a reference to a patient, a hospital, and a Feature (qualified references to the database instances), and the dataset from which the value comes (qualified reference to the data).

3 Results

3.1 Implementation and scenario

I-ETL is implemented as a Python software built upon v3.12 and the well-known, scalable, and flexible MongoDB [31] database management system (v7.0.12). The source code is available on GitHub at <https://github.com/DEIB-GECO/i-etl>. It can be easily deployed within medical centers, possibly having various software and hardware setups, employing its Docker image [32]. The software includes a template CSV file for the metadata definition, the ETL, and the computation of interoperability metrics. It produces the target MongoDB database. Our software needs to be set up once for each data provider (e.g., each hospital) and can be re-run to add/update data or metadata.

The selection of relevant datasets and features is left to clinicians. To support them in the definition of the features of interest, the metadata model is provided as a pre-defined CSV file whose header corresponds to the 8 attributes of our one-entity metadata model. Then, clinicians gather information for each feature of interest and fill the metadata file accordingly. The interoperability score of the target database highly depends on the quality and quantity of the provided metadata.

To showcase and evaluate our approach and its implementation, we consider the study of patients with end-stage kidney disease (ESKD) affected by COVID-19 [33]. ESKD is the last stage (stage 5) of chronic kidney disease; this causes kidneys to be at 1/10 of their normal capacity, incapable of effectively removing waste or excess fluid from blood. Due to their renal impairment, ESKD patients are at high risk of severe COVID-19, thus necessitating extra care. From this study, we consider a scenario where two hospital centers collaborate and share their different kinds of data to run a precise federated analysis. Here, clinicians are interested in the following two questions:

- “Which set of ESKD-related genes favor the development of severe forms of COVID-19?”
- “Which COVID symptoms are amplified due to the renal impairment of ESKD patients?”.

We selected the open and real-life datasets provided by [33] at [34], containing 111 patients for which phenotypic, diagnosis, imaging, clinical, and genomic data is provided. We allocate datasets to our two hospitals as follows:

ontology	code	name	dataType
loinc	57723-9	Sample_ID	string
hgnc	ENSG00000250433	ENSG00000250433.1	integer
hgnc	ENSG00000258591	ENSG00000258591.2	integer
hgnc	ENSG00000115902	ENSG00000115902.11	integer
hgnc	ENSG00000130234	ENSG00000130234.13	integer
...

Table 3 The metadata obtained from RNA sequence counts, leading to genomic features. For all features, *visibility*=public, *unit*= \emptyset .

- The first hospital H_1 processes only genomic data. Genomic data corresponds to RNA sequence counts for a panel of 60,649 genes.
- The second one, H_2 , processes phenotypic, clinical, imaging, and diagnosis data. Clinical data comes from flow cytometry for 45 general cells, while imaging data is about radiology evidence of COVID-19 in patient lungs. Diagnosis data provides information about the possible causes of ESKD as well as its severity.

3.2 Metadata creation

Starting from the above-mentioned datasets, we defined the set of relevant features F and described them according to our metadata model. The metadata has been collected and mapped to existing ontologies manually.

3.2.1 Hospital 1 metadata

In Table 3, we report an excerpt of metadata regarding genomic data. Original data contains RNA sequence counts computed for a panel of 60,649 genes for all the patients. In this panel, not all genes are relevant whilst they are costly to process. Therefore, we filtered the panel to keep the subsets of genes that are the 1,000 most expressed genes for each patient. The reason is that a high sequence count typically indicates that many reads are associated with a gene, suggesting a higher level of gene expression. The union of the top 1,000 genes-sets of each patient corresponds to a set of 2,382 genes, each leading to a feature in the metadata. Next, the code associated with each of them corresponds to the gene name without its version (the number after the dot); this has been computed automatically for the 2k genes. All gene codes can be found in the HGNC [35] ontology. In the dataset, values correspond to RNA sequence counts and thus are integers.

3.2.2 Hospital 2 metadata

Metadata for phenotypic data, presented in Table 4, has been collected by gathering personal information for each patient. It contains nine variables, each mapped to a SNOMED-CT code¹. 77% of them are categorical, each of them being associated with up to five categories (column *categories*; each value here has been mapped to

¹Tables 4, 5, 6, and 7 exhibit *code* values composed of several individual codes, joined with operators such as : or =. The process of joining several codes is referred to as *post-coordination* and is helpful when a concept/variable cannot be described with a single code (mainly because it does not exist in any ontology). We further discuss this point in Section 4.1.

code	name	visibility	dataType	unit	categories
422549004	individual_id	public	string	-	-
397731000	ethnicity	public	category	-	asian, white, black, other
734000001	sex	public	category	-	M, F
397669002	calc_age	private	integer	years	-
307294006: 246454002=“IHD”	ihd	public	category	-	no, yes.stent, unknown, yes.no.intervention, yes.cabg
307294006: 246454002=111293003	previous_vte	public	category	-	yes.dvt, yes.pe, yes.other, no
13645005	copd	public	category	-	yes, no, copd, bi
73211009	diabetes	public	category	-	yes.T1, yes.T2, no
365981007	smoking	public	category	-	never, not.current, ex unknown, current

Table 4 The metadata obtained from the patient personal information, leading to phenotypic features. For all features, *ontology*=SNOMED-CT.

code	name	dataType	categories
422549004	individual_id	string	-
46177005	cause_eskd	category	DN, Unknown, other, GN, HTN, Genetic
405162009: 47429007=840539006	WHO_severity	category	moderate, mild, severe, critical
419620001	fatal_disease	boolean	-

Table 5 The diagnosis metadata. For all features, *ontology*=SNOMED-CT, *visibility*=public and *unit*= \emptyset .

code	name	dataType	categories
422549004	individual_id	string	-
840539006: 363589002=	radiology_evidence_covid	category	cxr, no, not.done, yes, yes:CVCX1 yes:extensiveBilateralAirspaceConsolidation, yes:extensiveConsolidation, yes:leftBasalInfiltrate, yes:patchyBilateralConsolidation, yes:patchyOpacificationBothLungFields
363680008			

Table 6 The metadata obtained from the imaging datasets. For all features, *ontology*=SNOMED-CT, *visibility*=public and *unit*= \emptyset .

a SNOMED-CT code). The feature `calc_age` is kept private because patient privacy could be violated by looking at young outliers (note that few patients are much younger than the mean age of 68 years old).

For metadata about diagnoses, four public features are retained (Table 5) and mapped to a SNOMED-CT code. The two categorical features have a reasonable number of associated categories, each mapped to a SNOMED-CT code.

Metadata for imaging data (Table 6) leads to two features: the patient id and the observed anomaly in the radiology scans (`radiology_evidence_covid`). The latter feature has been associated with ten categories of evidence, each extracted from original imaging scans using dedicated image processing techniques and mapped to a SNOMED-CT code.

ontology	code	name	dataType
snomed ct	422549004	individual.id	string
loinc	57723-9	sample.id	string
snomed ct	117400003:260864003=732272000	CD66b+ CD45+	numeric
snomed ct	115412003:260864003=732272000	CD4+ T CD45+	numeric
\emptyset	\emptyset	Siglec-1+ NKG2D-HLA-DR+	numeric
...

Table 7 The metadata obtained from flow cytometry analyses, leading to clinical features. For all features, $visibility = \text{public}$, $unit = \emptyset$.

	H_1		H_2	
	Feature	Record	Feature	Record
Phenotypic	-	-	8	888
Clinical	-	-	45	748
Diagnosis	-	-	3	251
Imaging	-	-	1	70
Genomic	2,382	250,103	-	-

Table 8 Statistics of the databases obtained in experiments.

Finally, metadata of clinical data comes from the flow cytometry analyses of patients. It leads to 47 features; a subset of them is shown in Table 7. Eight features out of 47 could not be mapped to any SNOMED-CT code because one or several of the specified acronyms are not included in the ontology (this can be captured by interoperability metrics, see Section 3.4). This is, for instance, the case of **Siglec-1** and **NKG2D**. All the 45 flow cytometry measures lead to numeric values, do not present units, and are all accessible without restriction ($visibility = \text{public}$).

3.3 ETL execution

When run on the initial datasets from the two hospitals considered in our scenario, two databases H_1 and H_2 are consolidated – see Table 8 for numbers of corresponding features and records grouped by kind of data. Both H_1 and H_2 contain 111 Patients and one Hospital instance.

3.4 Interoperability assessment

Finally, I-ETL reports on the overall interoperability of the target database by computing and displaying our set of metrics (described in Section 2.5). This final step is highly important to check that obtained databases (as described in Table 8) are complete and sound, especially when running federated analyses over several databases. Table 9 lists the scores achieved for our metrics on the databases of H_1 and H_2 . It also provides the total number of objects accounted for the score. The score s ranges from 0 to 1 (included), with higher values indicating better performance.

In general, we appreciate that I-ETL created two highly interoperable databases from the ESKD-COVID patient data. In detail, full interoperability ($s = 1.00$) is achieved for 8 metrics in H_1 and 10 metrics in H_2 . High interoperability ($0.8 \leq s < 1.00$) is achieved for 1 metric in H_1 and 2 in H_2 . Low interoperability ($s < 0.8$) is

achieved for 2 metrics in H_1 and 2 in H_2 . Finally, 3 metrics lead to null scores in H_1 (N/A values in Table 9). This is because they were not applicable to the hospital data. For instance, there is no categorical feature for H_1 , thus (M4) could not be computed. By analyzing in more detail the achieved interoperability scores, we observe the following:

- The ratio of selected features (A1) is low for H_1 because only 2,382 genes were selected among the large panel of 60,649 genes. This drastic gene selection was necessary to run I-ETL in a reasonable time, while not relaxing important ones for federated analyses. In H_2 , a high score is achieved because almost all phenotypic, clinical, diagnosis, and imaging features have been retained. Examples of excluded features are `WHO_temp_severity` (a duplicate of `WHO_severity`), `time_from_first_symptoms` and `time_from_first_positive_swab` (the former containing the hour at which COVID symptoms appeared, the latter being the hour at which the nasal test has been done – both are not useful for understanding correlation between ESKD and COVID).
- Assessment of (M1) in H_2 leads to a score of 0.87 because 8 features out of 62 could not be mapped to existing ontology concepts. This is, for instance, the case of the last feature shown in Table 7 because `Siglec-1`, `NKG2D`, `HLA` and `DR` are not associated with any SNOMED-CT code.
- (M5) leads to very low scores for both H_1 and H_2 , respectively 0 and 0.02. This arises because most of the described features have no associated *unit* in the metadata, e.g., only `calc_age` had a unit in H_2 . Empty units cover two cases (without distinction): there is no unit for the feature (as for a ratio), and there is one but it has not been specified. To distinguish them and improve interoperability, experts should explicitly specify in the metadata when a feature has no unit. Unfortunately, this did not happen for our experimental datasets, leading to mostly empty units for numeric features, thus low scores.
- (E1) scores are very high for both H_1 and H_2 , meaning that almost all ontology resources (associated with features or categorical values) carry a label. This ensures interoperability and shows that ontologies can provide information about their resources.
- (E3) leads to a null score for H_1 and a 0-score for H_2 for the same reason mentioned above for (M5). In H_1 , no numeric value has a unit, thus leading to the N/A score. In H_2 , only the feature `calc_age` has a unit specified in the metadata, but no unit was provided in the data (as in the value “3 years”): thus, the score of 0.

4 Discussion

4.1 Challenges and limitations

The primary challenge we faced was to design a conceptual data model that could fit the various kinds of data brought by hospitals and clinical centers. Reusing existing CDMs was deemed not possible, because many of them are tied to entities of particular use cases, e.g., OMOP [8] allows to represent observational data and is hardly extensible to model genomic information. The more general ones, e.g., FHIR [10], lift the

Metric	H_1 total	H_1 score	H_2 total	H_2 score
A1	60,650	0.04 (L)	65	0.92 (H)
A2	5	1.00 (F)	5	1.00 (F)
M1	2,382	1.00 (F)	62	0.87 (H)
M2	2,382	1.00 (F)	62	1.00 (F)
M3	2,382	1.00 (F)	62	1.00 (F)
M4	N/A	N/A	10	1.00 (F)
M5	2,382	0.00 (L)	46	0.02 (L)
E1	2,382	0.99 (H)	46	1.00 (F)
E2	250,103	1.00 (F)	1,957	1.00 (F)
E3	N/A	N/A	1	0.00 (L)
E4	N/A	N/A	1,028	1.00 (F)
E5	250,103	1.00 (F)	1,957	1.00 (F)
E6	250,103	1.00 (F)	1,957	1.00 (F)
E7	250,103	1.00 (F)	1,957	1.00 (F)

Table 9 Interoperability assessment for the databases located in H_1 and H_2 . Interoperability levels are: full (F: $s = 1$), high (H: $0.8 \leq s < 1$), low (L: $s < 0.8$).

above limitation by exhibiting entities of various kinds. This makes them well suited to initiatives where new data needs to be collected, processed, and stored. However, they are not yet general enough to design automatic integration workflows for existing data (as opposed to hand-made ETL pipelines), notably due to their reliance on specific attributes. Following those observations, we propose a novel common data model based on the notions of features and records – abstract concepts of how any dataset can be represented. It currently represents six kinds of healthcare-related data, but is easily usable with other kinds, e.g., administrative or surgery-related data. This makes our conceptual model general enough to be used as a CDM in a wide variety of healthcare projects. In a broader scope, our framework could be utilized in many other contexts, e.g., journalistic sources, spatial databases or social human sciences sources, while only requiring to design a new CDM (such as the one presented in Figure 2) reflecting entities of the domain and leveraging the notions of feature and record.

For what concerns metadata creation – a crucial step to achieve high interoperability – the main challenge lies in the contribution of clinicians, who often do not have the time and/or knowledge to create it. So far, experts need to manually define all the features they are interested in, specify their related information and map each of them to an ontology code. Creating metadata may represent considerable manual work, especially for federated analyses where several datasets are joined. Nonetheless, this is the only part where experts are required to do a technical work, supported by our easy-to-fill metadata model. Even though each ontology is tailored to a particular type of healthcare data, e.g., HGNC is for genes and LOINC is for clinical measurements, finding appropriate ontologies and then searching them for suitable concepts is very time-consuming. Also, some concepts are very specific, thus are not represented in any well-known -commonly adopted- ontology. They can be created through *post-coordination*, a process to join several existing codes. For instance, the feature `previous_vte` (whether the patient already had a venous thromboembolism) does not exist in SNOMED-CT but can be represented with the following association of codes: “307294006: 246454002=111293003” (meaning that there is an occurrence of venous

thrombosis in the patient’s personal history). Creating post-coordinated codes is even more time-consuming.

Manual mapping also suffers from being error-prone, especially when the number of features is large. To limit experts’ manual efforts and errors, we envision semi-automatic support that (*i*) proposes a set of ontologies in which the concept is likely to appear (e.g., with BioPortal Recommender [36, 37]); (*ii*) lets experts select the most appropriate one; (*iii*) automatically proposes a set of codes that fit the concept in the selected ontology (e.g., by integrating BioPortal Search [38, 39]); and (*iv*) lets the expert select the most appropriate term code. Such methods should be used in a human-in-the-loop process. Indeed, they do rely on various metrics, including semantic similarity measures, but have very vague or no context about the scenario, thus may return inappropriate codes. Moreover, it is crucial that medical experts are provided user-friendly support to share their knowledge on the context, so that accurate domain-specific information can be ensured.

Concerning FAIR principles, the sub-principle I3, stating that qualified references to the data and metadata are necessary, is only partially implemented so far. Indeed, qualified references to the data are already included because each entity in the conceptual data model has an *identifier*. However, qualified references to the metadata are not yet included, but will be in subsequent work by providing a catalog to browse and search datasets based on their metadata.

4.2 Outlook

Our I-ETL framework has been developed in the context of a large European project called BETTER [40], whose overarching objective is to develop a decentralized and federated analysis of healthcare data. In this project, seven clinical centers are involved and they all work on the general domain of genetic rare diseases. Yet, they derived three use cases of interest, namely, pediatric intellectual disability, retinal dystrophies, and self-harm behaviors for autistic patients. While all of them rely on genomic data (at least), they also use different kinds and forms of data, thus highlighting the need for a general and easy-to-use framework to integrate and process them. In practice, each center provides datasets from a plethora of different kinds for the use case they are interested in. Starting from this, we discussed with them the healthcare research questions raised by their use cases, their available data (clinical measurements, genomic variants, MRI scans, etc.) and their ideas in terms of Federated Learning analyses. Next, we designed I-ETL and our two conceptual models (for metadata and data). In parallel, clinical experts discussed the metadata to be considered and filled out the metadata for each of their datasets by leveraging our metadata model. At this stage, hospitals have agreed on common and specific features to include in the metadata. By doing so, they ensure that their databases can be joined for further analyses (otherwise, each hospital would end up with a unique feature set). This is where most of the work happens for medical experts (formulate questions, find datasets, specify metadata); the rest of the pipeline is automatic and leads to a ready-to-use database.

We are currently deploying our framework inside each partner hospital and collecting feedback on this deployment as well as the usage of our tool. The I-ETL pipeline has been well-received by all the different stockholders involved in the BETTER project.

Even if the overhead in the data integration pipeline is costly and demanding, all the actors found that the overall process of creating an interoperable database on their server is worth the effort as long as it allows them to later create AI federated algorithms for medical decision making. Our next task is to discuss with them to finalize their FL scenarios and implement corresponding algorithms.

As future developments, BETTER aims at providing (*i*) a catalog for browsing metadata and aggregated data of target databases, as well as (*ii*) a platform for running decentralized and federated analyses of the data.

The catalog will be a website listing all the accessible databases and providing aggregated views of the data for each of them – we already initiated this work in [41]. For instance, the clinicians of the BETTER project may browse the metadata of different hospitals to check which other institutions they can join forces with. They may also take a deeper look at the aggregated data (while original data and the target database are never accessible outside of centers). For instance, they can investigate the patient age distribution as well as the set of diseases of patients of another institution to understand whether a federated analysis combining their data would make sense.

After deciding which datasets and which institutions can be joined, federated analysis will be run on a platform based on the Personal Health Train (PHT [42]) paradigm. This platform will include statistical and AI-based models for analyzing various data stored in the underlying I-ETL-based databases. In the end, clinicians will be able to explore the results of the federated computations and gain insights toward solving their research healthcare questions.

5 Conclusions

In this paper, we presented I-ETL, a framework for integrating heterogeneous healthcare datasets with interoperability as a first-class citizen. Our contributions are the following. First, we proposed a general data model for a large set of health datasets, including clinical, phenotypic, genomic, diagnosis, imaging, and medication data. This conceptual model serves as a common data model for various healthcare settings. Its main strengths are to take into consideration experts' knowledge (metadata) and to be easily extensible/tunable for other scenarios. Next, we proposed and implemented an ETL pipeline for transforming the input data into a database designed on our conceptual model. Incidentally, I-ETL also allows for resource savings (personnel and servers) because it is easy to put in place and does not require a large-scale centralized server. Finally, I-ETL provides a set of across-pipeline metrics for assessing the interoperability level throughout the whole process of integrating the input data into a target database. Ensuring and assessing interoperability also goes into the direction of data quality; well-conceptualized and homogenized datasets will be easily used for FL analyses. Experiments on a small open-data-based scenario with two hospitals have shown that I-ETL can achieve high interoperability scores, thereby enabling effective collaboration between different medical centers, notably via federated analysis of the target databases.

Several research directions arise from the present work – some of which are already ongoing. First, we are now working on the querying of the interoperable databases

(available at each center) through the catalog. This task is complex because, for privacy reasons, the catalog relies on aggregated data only and the real data in the hospital servers cannot be accessed. Therefore, the challenge here is to find the right balance between super-aggregated data (very safe but not very useful due to the high information loss) and low-aggregated data (more useful but with privacy concerns). A subsequent direction is the design of a human-in-the-loop recommendation module for metadata. This would automatically recommend ontologies and codes for a given set of features (recall Section 4.1), allowing experts to save time and reduce errors while keeping control of the obtained metadata. Another interesting addition would be to add more context to the records, e.g., to know whether a value has been observed before or after surgery. This would contribute to a richer common data model while remaining as general as possible. In parallel with these three directions, the BETTER partners work on the implementation of the federated analysis platform to enable the design and secure execution of Federated Learning tasks.

With this project, we learned that there is no “one-size-fits-all” solution, especially when working in large consortium and projects. Despite these challenges, bringing computer science methods and developments to the healthcare sector opens the road to better health systems, improving citizens’ global health.

List of abbreviations.

CDM: Common Data Model
CSV: Comma-Separated Value
DICOM: Digital Imaging and Communications in Medicine
ETL: Extract-Transform-Load
FAIR: Findable, Accessible, Interoperable, Reusable
HGNC: HUGO Gene Nomenclature Committee
LOINC: Logical Observation Identifiers Names and Codes
OMIM: Online Mendelian Inheritance in Man
PHT: Personal Health Train
SNOMED-CT: SNOMED Clinical Terms
VCF: Variant Call Format

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The datasets used in this manuscript are available in the Zenodo repository at <https://zenodo.org/records/7410194> in open source.

Competing interests. The authors declare that there are no conflicts of interest.

Funding. This work is supported by the Horizon Europe project BETTER, Grant agreement n. 101136262.

Authors' contributions. NB, AB, BB, and PP conceived the research; NB and BB jointly conceptualized the framework; NB designed/implemented the ETL pipeline and software and performed the experiments; BB curated the ontology mappings; AB

and PP acquired funding; PP supervised the project; NB drafted the manuscript; AB, BB, and PP revised/edited the manuscript.

Acknowledgments. We thank all the partners involved in the BETTER project for their valuable contributions and feedback.

References

- [1] Dash, S., Shakyawar, S.K., Sharma, M., Kaushik, S.: Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* **6**(1), 1–25 (2019)
- [2] Bietz, M.J., Bloss, C.S., Calvert, S., Godino, J.G., Gregory, J., Claffey, M.P., Sheehan, J., Patrick, K.: Opportunities and challenges in the use of personal health data for health research. *Journal of the American Medical Informatics Association* (2016). Oxford
- [3] The WHO Global Health Observatory. <https://www.who.int/data/gho>. Accessed April 10th, 2025.
- [4] World health statistics 2024: monitoring health for the SDGs, sustainable development goals. <https://www.who.int/publications/i/item/9789240094703>. Accessed April 10th, 2025.
- [5] European involvement for rare diseases. https://health.ec.europa.eu/rare-diseases-and-european-reference-networks/rare-diseases_en. Accessed April 10th, 2025.
- [6] Puttmann, D., Groot, R., Keizer, N., Cornet, R., *et al.*: Assessing the FAIRness of databases on the EHDEN portal: A case study on two Dutch ICU databases. *International Journal of Medical Informatics* **176**, 105104 (2023)
- [7] Blacketer, C., Voss, E.A., DeFalco, F., Hughes, N., Schuemie, M.J., Moinat, M., Rijnbeek, P.R.: Using the data quality dashboard to improve the EHDEN network. *Applied Sciences* **11**(24), 11920 (2021)
- [8] Stang, P.E., Ryan, P.B., Racoosin, J.A., Overhage, J.M., Hartzema, A.G., Reich, C., Welebob, E., Scarneccchia, T., Woodcock, J.: Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of internal medicine* **153**(9), 600–606 (2010)
- [9] Reščić, N., Alberts, J., Altenburg, T.M., Chinapaw, M.J.M., De Nigro, A., Fenoglio, D., Gjoreski, M., Gradišek, A., Jurak, G., Kiourtis, A., Kyriazis, D., Langheinrich, M., Mancuso, E., Mavrogiorgou, A., Pechenizkiy, M., Pratola, R., Ribeiro, J., Sorić, M., Taj, F., Tammelin, T.H., Vastenburg, M., Vilanova, A., Vrijkotte, T.G.M., Luštrek, M.: SmartCHANGE: AI-based long-term health risk

- evaluation for driving behaviour change strategies in children and youth. In: ICAMCS (2023)
- [10] The FHIR framework. <https://hl7.org/fhir/summary.html>. Accessed April 10th, 2025.
 - [11] i2b2: Informatics for Integrating Biology and the Bedside. <https://www.i2b2.org/>. Accessed April 10th, 2025.
 - [12] Bernasconi, A., Ceri, S., Campi, A., Masseroli, M.: Conceptual modeling for genomics: building an integrated repository of open data. In: Conceptual Modeling: International Conference, pp. 325–339 (2017). Springer
 - [13] Bernasconi, A., García S, A., Ceri, S., Pastor, O.: A comprehensive approach for the conceptual modeling of genomic data. In: International Conference on Conceptual Modeling, pp. 194–208 (2022). Springer
 - [14] Ceri, S., Bernasconi, A., Canakoglu, A., Gulino, A., Kaitoua, A., Masseroli, M., Nanni, L., Pinoli, P.: Overview of GeCo: a project for exploring and integrating signals from the genome. In: International Conference on Data Analytics and Management in Data Intensive Domains, pp. 46–57 (2017). Springer
 - [15] Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., *et al.*: The human cell atlas. *elife* **6**, 27041 (2017)
 - [16] OpenEHR. <https://openehr.org/>. Accessed April 10th, 2025.
 - [17] SMART Health IT. <https://smarthealthit.org/>. Accessed April 10th, 2025.
 - [18] Ong, T.C., Kahn, M.G., Kwan, B.M., Yamashita, T., Brandt, E., Hosokawa, P., Uhrich, C., Schilling, L.M.: Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. *BMC medical informatics and decision making* **17**, 1–12 (2017)
 - [19] Kouremenou, E., Kiourtis, A., Kyriazis, D.: A data modeling process for achieving interoperability. In: International Conference on e-Health and Bioengineering (2023). Springer
 - [20] Stearns, M.Q., Price, C., Spackman, K.A., Wang, A.Y.: SNOMED Clinical Terms: overview of the development process and project status. In: Proceedings of the AMIA Symposium, p. 662 (2001). American Medical Informatics Association
 - [21] Huff, S.M., Rocha, R.A., McDonald, C.J., De Moor, G.J., Fiers, T., Bidgood Jr, W.D., Forrey, A.W., Francis, W.G., Tracy, W.R., Leavelle, D., *et al.*: Development of the logical observation identifier names and codes (LOINC) vocabulary. *Journal of the American Medical Informatics Association* **5**(3), 276–292 (1998)

- [22] McDonald, C.J., Huff, S.M., Suico, J.G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., *et al.*: LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry* **49**(4), 624–633 (2003)
- [23] Weinreich, S.S., Mangon, R., Sikkens, J., Teeuw, M.E., Cornel, M.: Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde* **152**(9), 518–519 (2008)
- [24] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L.B., Bourne, P.E., *et al.*: The FAIR guiding principles for scientific data management and stewardship. *Scientific data* **3**(1), 1–9 (2016)
- [25] Parciak, M., Suhr, M., Schmidt, C., Bönisch, C., Löhnhardt, B., Kesztyüs, D., Kesztyüs, T.: FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital. *BMC Medical Informatics and Decision Making* **23**(1), 94 (2023)
- [26] Pereira, A., Lopes, R.P., Oliveira, J.L.: SCALEUS-FD: a FAIR data tool for biomedical applications. *BioMed research international* **2020**(1), 3041498 (2020)
- [27] Candela, L., Mangione, D., Pavone, G.: The FAIR assessment conundrum: Reflections on tools and metrics. *Data Science Journal* **23**(1) (2024)
- [28] BioPortal Ontologies. <https://bioportal.bioontology.org/ontologies>. Accessed April 10th, 2025.
- [29] Chen, P.P.-S.: The entity-relationship model—toward a unified view of data. *ACM transactions on database systems (TODS)* **1**(1), 9–36 (1976)
- [30] Batini, C., Ceri, S., Navathe, S.: Conceptual design—an entity-relationship approach. Redwood: Benjamin Cummings (1992)
- [31] MongoDB. <https://www.mongodb.com/>. Accessed April 10th, 2025.
- [32] Docker. <https://www.docker.com/>. Accessed April 10th, 2025.
- [33] Gisby, J.S., Buang, N.B., Papadaki, A., Clarke, C.L., Malik, T.H., Medjeral-Thomas, N., Pinheiro, D., Mortimer, P.M., Lewis, S., Sandhu, E., McAdoo, S.P., Prendekie, M.F., Willicombe, M., Pickering, M.C., Botto, M., Thomas, D.C., Peters, J.E.: Multi-omics identify falling LRRC15 as a COVID-19 severity marker and persistent pro-thrombotic signals in convalescence. *Nature Communications* **13**(1) (2022)
- [34] Dataset for Multi-omics identify LRRC15 as a COVID-19 severity predictor and

- persistent pro-thrombotic signals in convalescence. <https://zenodo.org/records/7410194>.
- [35] Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., Wain, H.: The HUGO Gene Nomenclature Committee (HGNC). *Human genetics* **109**(6) (2001)
 - [36] BioPortal Recommender. <https://bioportal.bioontology.org/recommender>. Accessed April 10th, 2025.
 - [37] Martínez-Romero, M., Jonquet, C., O'connor, M.J., Graybeal, J., Pazos, A., Musen, M.A.: NCBO ontology recommender 2.0: an enhanced approach for biomedical ontology recommendation. *Journal of biomedical semantics* **8**, 1–22 (2017)
 - [38] BioPortal Search. <https://bioportal.bioontology.org/search>. Accessed April 10th, 2025.
 - [39] Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research* **39**(suppl_2), 541–545 (2011)
 - [40] Better real-world health-data distributed analytics research platform. <https://www.better-health-project.eu/>. Accessed April 10th, 2025.
 - [41] Barret, N., Bernasconi, A., Cappiello, C., Palu, G., Pinoli, P.: Leveraging profiling to bridge healthcare silos for federated analyses. In: International Conference on Advanced Information Systems Engineering, pp. 20–28 (2025). Springer
 - [42] Beyan, O., Choudhury, A., Van Soest, J., Kohlbacher, O., Zimmermann, L., Stenzhorn, H., *et al.*: Distributed analytics on sensitive medical data: the personal health train. *Data Intelligence* **2**(1-2), 96–107 (2020)