Article published online: 2022-10-11

TransformEHRs: a flexible methodology for building transparent ETL processes for EHR reuse

Miguel Pedrera-Jiménez[1,2] Noelia García-Barrio[1] Paula Rubio-Mayo[1] Alberto Tato-Gómez[1] Juan Luis Cruz-Bermúdez[1] José Luis Bernal-Sobrino[1] Adolfo Muñoz-Carrero[3] Pablo Serrano-Balazote[1]

[1] Data Science Unit, Instituto de Investigación Sanitaria Hospital Universitario 12 de Octubre, Madrid, Spain
[2] ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
[3] Digital Health Research Unit, Instituto de Salud Carlos III, Madrid, Spain

Address for correspondence Miguel Pedrera-Jiménez, Eng, MSc, Health Informatics Department, Hospital Universitario 12 de Octubre, Av. de Córdoba, s/n, 28041 Madrid, Spain (e-mail: miguel.pedrera@salud.madrid.org).

Methods Inf Med 2022;61:e89–e102.

Abstract

Background During the COVID-19 pandemic, several methodologies were designed for obtaining electronic health record (EHR)-derived datasets for research. These processes are often based on black boxes, on which clinical researchers are unaware of how the data were recorded, extracted, and transformed. In order to solve this, it is essential that extract, transform, and load (ETL) processes are based on transparent, homogeneous, and formal methodologies, making them understandable, reproducible, and auditable.

Objectives This study aims to design and implement a methodology, according with FAIR Principles, for building ETL processes (focused on data extraction, selection, and transformation) for EHR reuse in a transparent and flexible manner, applicable to any clinical condition and health care organization.

Methods The proposed methodology comprises four stages: (1) analysis of secondary use models and identification of data operations, based on internationally used clinical repositories, case report forms, and aggregated datasets; (2) modeling and formalization of data operations, through the paradigm of the Detailed Clinical Models; (3) agnostic development of data operations, selecting SQL and R as programming languages; and (4) automation of the ETL instantiation, building a formal configuration file with XML.

Results First, four international projects were analyzed to identify 17 operations, necessary to obtain datasets according to the specifications of these projects from the EHR. With this, each of the data operations was formalized, using the ISO 13606 reference model, specifying the valid data types as arguments, inputs and outputs, and their cardinality. Then, an agnostic catalog of data was developed through data-oriented programming languages previously selected. Finally, an automated ETL instantiation process was built from an ETL configuration file formally defined.

Conclusions This study has provided a transparent and flexible solution to the difficulty of making the processes for obtaining EHR-derived data for secondary use

Keywords
- electronic health record
- FAIR Principles
- data reusability
- real-world data
- standards

received
March 25, 2022
accepted
July 5, 2022
article published online
October 7, 2022

DOI https://doi.org/10.1055/s-0042-1757763.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

understandable, auditable, and reproducible. Moreover, the abstraction carried out in this study means that any previous EHR reuse methodology can incorporate these results into them.

Introduction

Electronic health record (EHR) is defined as the repository of health data generated throughout the patient's life to achieve continuous, efficient, and high-quality health care.[1] Likewise, there are other uses of EHR, known as secondary uses, which include activities such as clinical research, public health, or evaluation of health outcomes.[2] In these EHR-derived uses, it is common for each analysis initiative to define its own data and content model based on its specific needs.[3] Although these models share most clinical concepts, they differ in format and recording criteria, leading to redundant data entry in multiple information systems designed for specific purposes and parallel to the health care information systems. To solve this problem, it is essential that the design of EHR systems follows the FAIR Principles,[4] thus achieving findable, accessible, interoperable, and reusable data. Following these principles, Hospital Universitario 12 de Octubre from Madrid, Spain (H12O) designed, implemented, and applied an innovative methodology for obtaining EHR-derived datasets for research, which allows incorporating the semantics of health data into the EHR reuse process,[5] being considered this process an international reference.[6] In this way, the data generated during health care were reused in multiple additional purposes in an agile manner, and adapted to changes in data specifications, while maintaining their original meaning and acceptable quality.[7] In the context of coronavirus disease 2019 (COVID-19), this line of work has allowed H12O to participate in an efficient and sustainable way, during the most critical moments of the pandemic, in several real world data (RWD) initiatives,[8] such as the 4CE Consortium,[9] the EHDEN Consortium,[10] and TriNetX.[11] Similarly, it has made it possible to optimize data collection in frequently manual processes, such as the automated data loading into the case report form (CRF) of the ISARIC-WHO Consortium.[12]

However, these EHR reuse processes are often based on black boxes on which the final data customer is unaware of how the data uploaded to their research database were recorded, extracted, and transformed. This became evident when, in the context of RWD studies in COVID-19, two publications in top-tier journals, one in The Lancet[13] and other in the New England Journal of Medicine,[14] were retracted less than 2 months after publication due to data quality, among other issues. These retractions highlight that, although medical informatics experts can identify the strengths and shortcomings of EHR as a source of data for health research, editorial teams and clinical readers lack the necessary framework to evaluate these studies in a fully critical manner, making it necessary to develop methodologies to evaluate the EHR data used in research studies. In order to make this possible, it is essential that these extract, transform, and load (ETL) processes are based on homogeneous and formal data operations, making them understandable, reproducible, and auditable.[15] Thus, this study proposes a methodology, according with FAIR Principles, to solve the existing difficulties in the implementation of transparent EHR reutilization processes for research and other purposes, being extendable to other organizations and applicable to any health condition.

Objectives

The main goal of this study was to design and implement a methodology, according with FAIR Principles, for building ETL processes for EHR reuse in a transparent and flexible manner, applicable to any clinical condition and health care organization. This involves several specific objectives, such as:

- Identify and formally define an initial and extendable set of data operations required to convert the EHR into data models for secondary use.
- Select and apply modeling and terminology standards for the formalization of the set of data operations.
- Select and apply appropriate programming languages for the development of the set of data operations.
- Design and implement a mechanism to automate the application of the data operations for specific use cases.

Methods

This work is part of the methodology previously designed by H12O for obtaining data for research from the EHR.[5] Specifically, it is focused on the third phase of that process, which covers the definition and implementation of the extraction and transformation operations on EHR data for its reuse in a multipurpose and transparent way, and without changing its original meaning.[16] This extended process comprises four stages:

- *Analysis of secondary use models and identification of data operations*, i.e., analyzing relevant secondary use models and identifying the data operations necessary to obtain those models from the EHR.
- *Modeling and formalization of data operations*, i.e., making use of standardization resources to model and formalize the identified data operations.
- *Agnostic development of data operations*, i.e., selecting the most appropriate programming languages and employing them in the development of data operations applicable to any specific use case.
- *Automation of the instantiation of data operations*, i.e., building an automated mechanism to instantiate data operations to specific use cases.

►Fig. 1 summarizes the methodology designed to achieve the objectives proposed in this study.

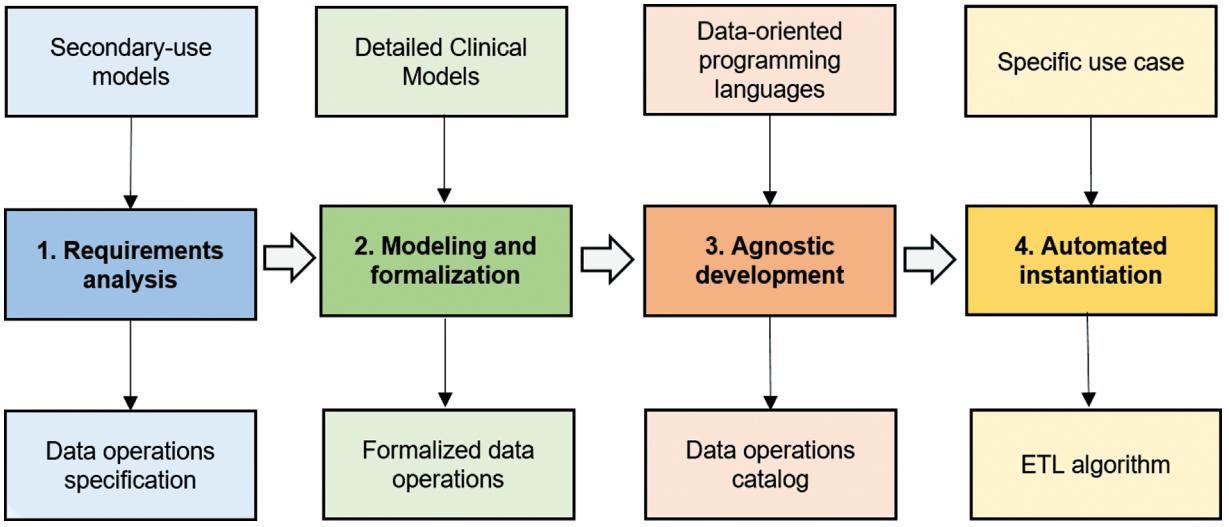

Fig. 1 Stages of the methodology for building transparent ETL processes for EHR reuse. EHR, electronic health record; ETL, extract, transform, and load.

COVID-19 use case has been selected to apply the methodology proposed, in order to improve the transparency of EHR reuse processes for the many data initiatives that have emerged around this new and unknown health condition at H12O.[9–12] The methodology was valuable and innovative in this pandemic scenario, where data were urgently needed to study this disease and neither the time nor the resources were available to conduct manual data recording or to thoroughly evaluate data collection processes. Despite this, the methodology is, by design, applicable to other health conditions and flexible to be adopted by any organization, through parameterization mechanisms according to the restrictions of data sources and data outputs.

Applying the FAIR Principles to the EHR

In the framework of Open Science, the FAIR Principles,[17] formally published in 2016 by the Force11 community,[18] provide guidelines to improve the data generated, being transversal to all scientific disciplines, including health. This specification has been included in the European Commission's data management guidelines, and its conclusions on the costs of not having FAIR research data are particularly relevant.[19] Thus, according to the FAIR Principles, data must be:

- *Findable*, i.e., data should be described with rich and interrelated metadata. The (meta)data should have a unique identifier and should be registered or indexed in a searchable resource.
- *Accessible*, i.e. (meta)data must be retrievable by its identifier, through an open, free, and universally applicable communication protocol, allowing authentication and authorization if necessary. Metadata must be accessible even when the data are no longer available.
- *Interoperable*, i.e., (meta)data should use a specific language for knowledge representation, and should use FAIR-compliant vocabularies, including references to other (meta)data.
- *Reusable*, i.e. (meta)data should be described with attributes, including license and origin information, among others.

The application of these principles to EHR implies a paradigm shift in the conception of EHR, moving from a design focused on the production of clinical documents to one based on the value of health data in patient health care and additional uses, such as research.[20] ►Table 1 describes and compares, briefly, the traditional paradigm based on clinical documents versus the paradigm based on the FAIR Principles.

While achieving this change does not require modifying information systems acquired by the health care organization, which is frequently complicated, but it is necessary to have governance over them in order to build an appropriate framework for recording, managing, and reusing health data.[21] This necessary EHR governance applies to the ability to: (1) centrally manage the clinical domain concept model of the different health care information systems; (2) select and incorporate health information standards into the health care information systems; (3) design and implement data recording and persistence mechanisms according to standardized information models; and (4) extract and process, in a centralized way, these data recorded and stored in the different health care information systems.

Modeling and Formalization of the EHR

The process of reusing health care data must start with an EHR properly modeled, formalized, and persisted in accordance with internationally used health information standards, thus allowing the data to maintain their original meaning intact regardless of the information system that contains them. This prior work on data will determine, to a large extent, the quality, usefulness, and acceptability of the outputs derived from the EHR for specific research purposes. Thus, the suitability of the EHR, on which to apply the proposed methodology, has been based on the paradigm of the Detailed Clinical Models (DCMs),[5] which defines a dual

Table 1 Comparison of EHR paradigms: document-based versus FAIR-based

Document-based paradigm	FAIR Principles-based paradigm
D1. Data entry is modeled around the clinical document, which is the output obtained from them.	F1. Data entry is modeled to respond to the needs of health professionals, allowing its use for health care and secondary purposes.
D2. Free-text or semi-structured data entry predominates.	F2. Structured and coded data entry predominates.
D3. Data do not follow common convergence models, so they are only understandable in the environment where they have been generated.	F3. Data are recorded and persisted based on health information standards, allowing them to be exchanged and transformed into other formats, without loss of meaning.
D4. The clinical domain concepts are implicit in the data tables, so that, the incorporation of new concepts implies an alteration of the data model.	F4. The clinical domain concepts are stored in tables independent of the data tables, making them flexible to incorporate new concepts without altering the data model.

Abbreviation: EHR, electronic health record.

model composed of the reference model and the archetype model. The reference model defines the set of generic components to build interoperable EHR, while the archetype model formalizes any concept of the clinical domain, such as "oxygen saturation" or "discharge report", built by the combination of the components and constraints of the reference model.[22] In this way, a set of concepts was modeled and formalized in the EHR for extraction and transformation with full meaning, according to parts 1 and 2, reference model and archetype model respectively, of the ISO 13606 standard.[23,24] The choice of this standard is due to the fact that: (1) it defines a formal, rigorous, and stable information architecture for the definition of clinical domain concepts and the communication of the EHR[25]; (2) it allows the extension of the clinical concept model without altering the structure of the databases, which is fundamental for providing flexibility to the reuse processes in the face of changes in the output specifications[26]; (3) it has been successfully applied in different scenarios to achieve semantic interoperability of the information[27,28]; (4) it is recommended by the Spanish Ministry of Health as a standard for the definition of interchangeable EHR extracts between the different autonomous regions that compose the country,[29] and (5) it has been adopted by H12O as a base standard for the management and governance of clinical concepts, interoperability, and reuse of the EHR.[30]

Similarly, the archetype model allows information models to be formalized and linked to controlled standard vocabularies to represent the meaning of their components, as well as to establish the value sets for coded data elements.[31] This fully meaningful modeling and formalization of clinical domain concepts makes them semantically unambiguous, which is essential for building effective and homogeneous EHR reuse processes. In this regard, it is essential that the EHR natively incorporates standard terminologies, such as SNOMED CT and LOINC,[32,33] and avoids the use of clinical classifications for data entry in health information systems, since they contain grouped, calculated, or inferred concepts. Starting from these highly granular concepts, e.g., specific histological or clinical diagnosis of breast cancer, enables them to be further translated into more general or aggregated concepts, e.g., variable indicating with a yes/no answer on whether the patient has an oncological disease, which are commonly used in research data models.

Thus, the execution of this methodology on EHR extracts in accordance with DCMs allows the ETL process to be flexible and applicable regardless of condition, information systems, and even health care organization.[34] ►Table 2 describes the set of clinical archetypes, based on the entry component, created in H12O for EHR reuse, indicating the terminologies used in their semantic bindings.

Building Transparent ETL Processes for EHR Reuse

Data operations for EHR reuse constitute the core of the methodology, providing a transparent process for extraction and transformation of health care data into specific models for the development of research studies.[16] They are health condition-agnostic by design, and flexible to adapt to the data sources and constraints of the secondary use data model, as well as to the requirements set by the regulatory authorities. Furthermore, these operations can be implemented within the health care organization's infrastructure according to the information security policies in place, for example, by pseudonymizing the data at the source on which the extraction operations are applied. Hence, first, an analysis of significant models of secondary use employed in data projects was carried out for the identification of the set of necessary operations. Thereby, the identified set was formalized through the DCM paradigm. Subsequently, a catalog of data operations was developed in a use-case agnostic manner. Finally, a process was built to instantiate them, in an automated way, to specific use cases.

Identification of Data Operations

Secondary use models allow data to be represented and persisted for uses in addition to individual patient health care.[2] Consequently, they are less demanding than primary use models in terms of metadata about the registration process or access permissions. We can distinguish three types of secondary use models:

- *Clinical repository*: models that allow the centralization of data from multiple sources, under common data and content models, and are applicable to multiple purposes.

Table 2 Set of clinical archetypes for EHR reuse

Archetype	Description	Terminology binding
Patient	Demographics data, e.g., birthdate, sex, and vital status.	SNOMED CT
Encounter	Data related to inpatient, emergency, and outpatient visits.	SNOMED CT
Location	Patient locations during hospitalization, e.g., ICU admission.	SNOMED CT
Observation	Clinical, laboratory, and patient-reported observations.	SNOMED CT, LOINC
Diagnosis	Health issues and clinical diagnoses.	SNOMED CT
Medication	Pharmacological treatment prescribed.	SNOMED CT
Procedure	Procedures performed, e.g., surgeries and nursing interventions.	SNOMED CT

Abbreviations: EHR, electronic health record; ICU, intensive care unit.

There are specifications that have become standards for this type of resources, such as the i2b2 model,[35] implemented in tools such as TriNetX,[11] and OMOP CDM,[36] adopted in international consortia such as EHDEN.[10]

- *CRF*: data collection models at patient level, which are designed according to a specific purpose. There are specifications agreed upon by experts that, due to their clinical and scientific relevance, are considered standards in a specific domain. Thus, the COVID-19 CRF model proposed by ISARIC and WHO, which was implemented with REDCap platform,[37] and it is therefore compatible with CDISC (Clinical Data Interchange Standards Consortium) standard reference model,[38] constitutes an international reference in this field.[12,39]
- *Aggregate dataset*: models for the representation of aggregated data, commonly used for public health purposes, and designed according to specific use cases. In the COVID-19 scenario, multiple proposals have emerged based on this typology since it does not involve sending data at patient level. An example of this is the aggregated dataset defined by the 4CE Consortium of the i2b2 TranSMART Foundation and Harvard Medical School.[9]

Thus, to define the set of operations, different data models used in several data-driven projects in H12O were analyzed,[9–12] considering the typologies described above. ►**Table 3** shows the list of specifications reviewed, the type of data model they incorporate, and their purpose.

Once the data models of the different projects were analyzed, it was possible to identify the operations required to generate them from the EHR. These operations were classified according to a series of meta-operations in accordance with their typology. These high-level operations, progenitors of the final operations, were as follows:

- *Extraction (E)*: operations to extract the necessary data from the different models formalized in the EHR. Two subtypes were defined:
  - *Extraction with criteria (E.1)*, i.e., extraction of data from the observation domain of the COVID-19 cohort of patients.
  - *Extraction without criteria (E.2)*, i.e., extraction from the observation domain without cohort restriction.
- *Selection (S)*: operations to select the necessary data under the constraints of the secondary use model. Two subtypes were defined:
  - *Selection with reference (S.1)*, e.g., selection of "Oxygen saturations" below 96%.
  - *Selection without reference (S.2)*, e.g., selection of the lowest value of "Oxygen saturations."
- *Transformation (T)*: operations to transform the data to the format of the secondary use model. Two subtypes have been defined:
  - *Transformation without semantic implication (T.1)*, e.g., changing the unit of measurement of concept "C-Reactive Protein" from mg/dL to mg/L.
  - *Transformation with semantic implication (T.2)*, e.g., calculating a concept "Body Mass Index (BMI)" from "Weight" and "Height".

This work has focused on data extraction, selection, and transformation operations since they are the ones with the greatest variability and complexity. The operations to achieve a homogeneous and standard data load are contemplated as a future work, being implemented at this point through ad-hoc data loads in the receiving systems.

Table 3 Data-driven projects analyzed for identification of data operations

ID	Data-driven project	Data model typology	Purpose
1	TriNetX Platform	i2b2 repository	Clinical trials and cohort's analytics
2	EHDEN Consortium	OMOP repository	Observational studies
3	ISARIC Consortium	COVID-19 CRF	Case reports and cohort's analytics
4	4CE Consortium	COVID-19 aggregate dataset	Cohort's analytics

Abbreviation: CRF, case report form.

Formalization of Data Operations

The formalization of data operations is based on the DCM paradigm, selecting for this purpose the reference model proposed by the ISO 13606 standard,[23] due to the reasons described above. This model defines the components needed to build an interoperable EHR, i.e., Folder, Composition, Section, Input, Cluster, and Data Elements, as well as the set of data types for the final data elements. This allowed limiting the valid data types for the inputs, outputs, and arguments of each data operation. In the present work, it was necessary to use the following subset of data types proposed by this standard:

- *Instance identifier (II)*, i.e., unique identifiers of patients within a given jurisdiction, e.g., Medical History Number.
- *Coded value (CV)*, i.e., concepts whose result is a set of possible CVs, e.g., SARS-COV-2 test, which may be positive, negative, or inconclusive.
- *Physical quantity (PQ)*, i.e., concepts whose result is a numerical value with unit of measure, e.g., oxygen flow rate measured in liters per minute.
- *Integer*, i.e., concepts whose result is an integer value, e.g., Glasgow Comma Scale score.
- *Date time*, i.e., concepts whose value is a point in time, e.g., date of symptom onset.

Likewise, the cardinality of the inputs and outputs of the set of data operations was established, i.e., how many data elements they can receive and how many they produce. Thus, while a "select data after a certain event" operation, e.g., clinical observations after an adverse event, can receive and produce multiple data elements of any of the specified data types, a "change unit of measure" operation, e.g., calculate the height in centimeters originally measured in meters, can only receive and produce a single data element of PQ data type.

Development of Data Operations

The development of the data operations was carried out by previously analyzing the data-oriented programming languages most suitable for the different types of operations. The premise for this analysis was that they should constitute technologies widely used in the health domain, and therefore, assumable by the technical and data science teams of a health organization at the level of H12O. Thus, the following technologies were selected:

- *Structured Query Language (SQL)*, i.e., query language designed to manage and retrieve information from relational database management systems, being the standard of the American National Standards Institute in 1986 and the International Organization for Standardization (ISO).[40] In the proposed methodology, it is the technology used for *Extraction* operations.
- *R programming language*, i.e., data-oriented programming language widely used in scientific research, being also very popular in the fields of machine learning, data mining, biomedical research, and bioinformatics.[41] In the proposed methodology, it is the technology used for *Selection* and *Transformation* operations.

In this process, other technologies were analyzed, such as the Archetype Query Language (AQL)[42] for the extraction and selection of data based on the defined archetypes.[43] This technology was discarded at this point due to the complexity of its implementation in the current health care environment, being considered for deployment in future steps of this line of work. With this, the operations were developed according to a design agnostic to specific use cases, so that, they could be instantiated automatically, as a final part of the application of the methodology. ►Table 4 shows the set of agnostic operations developed.

Automated Instantiation in Specific Use Cases

Finally, an automated mechanism for instantiating the different data operations to specific use cases was designed and built. This avoids the ad-hoc development of the ETL process for each secondary use model to be obtained, allowing to build a more efficient and scalable process. For this purpose, a configuration file was designed, formalized, and implemented, based on the data operations catalog and the requirements of the analyzed data models.[9–12] This ETL configuration file, implemented using XML and XSD languages,[44] consists of the following components:

- *Connections*, i.e., component for the configuration of the connection parameters to the different databases of the information systems required in the use case. It is composed by the following elements: *alias*, *driver*, *url*, *url_driver*, *user*, and *password*.
- *Sources*, i.e., component for the configuration of the interaction parameters with the data tables required in the use case. It is composed of the following elements: *connection*, *domain*, *table*, and *attributes*. The latter, in turn, is composed of *id*, *patient*, *provider*, *visit*, *concept*, *value*, *onset_date*, and *end_date*.
- *Operations*, i.e., component for the configuration of the instantiation parameters of the different data operations

Table 4 Developed data operations for EHR reuse, indicating in italics the fields to be instantiated in the application on specific use cases

Operation	Technology	Agnostic development
E	SQL	select *attributes* from *table* where *patient* in (*cohort*)
S	R	*output* < -filter(*input*, *input*[[*concept*]] < *argument*)}
T	R	*output* < - *input*[*input*[[*concept*]] = =*argument*] < -*argument*

Abbreviation: EHR, electronic health record.
Note: Operation: (E) extraction, (S) selection, and (T) transformation.

required in the use case. It is composed of the following elements: *operation*, *argument*, *input*, and *output*.

►Fig. 2 shows, as an example, the XML file necessary to connect to the information system, extract the data from the observations domain of COVID-19 cohort (operation E.1.1), select the data related to the "Oxygen saturation" concept (operation S.1.1), and, from these, the values below 96% (operation S.1.5), from which the "Respiratory risk" concept is inferred (operation T.2.2).

This file is processed by a function implemented in R, to extract the parameters it contains and to build the ETL process code according to the operations previously identified, formalized, and developed. ►Fig. 3 shows the extracted parameters and the R code of the data operations defined in the XML of ►Fig. 2. Thus, this file can be shared among the participating health organizations in a data consortium to facilitate the implementation of the ETL process. The parameterization of the connection and extraction components within the organization that owns the data allows for scalability between sites, while ensuring the security of the data and the confidentiality of the information systems.

```
<?xml version="1.0"?>
- <ETLconfiguration>
  + <connections>
  - <sources>
    - <source>
        <connection>HCIS_3106</connection>
        <domain>OBSERVATION</domain>
        <table>HDOC_INFOBANCO_OBX</table>
      - <attributes>
          <id>CODIGO</id>
          <patient>CODPACI</patient>
          <provider>AUTOR</provider>
          <visit>EPISODIO</visit>
          <concept>OBX</concept>
          <value>VALOR</value>
          <start_date>FECHA</start_date>
          <end_date/>
        </attributes>
      </source>
    </sources>
  - <operations>
    - <operation>
        <type>Extraction</type>
        <id>E.1.1</id>
        <input>HDOC_INFOBANCO_OBX</input>
        <output>OBSERVATION</output>
        <argument>COVID19_COHORTE_COMPLETA</argument>
        <concept>CODPACI</concept>
      </operation>
    - <operation>
        <type>Selection</type>
        <id>S.1.1</id>
        <input>OBSERVATION</input>
        <output>sao2</output>
        <argument>103228002</argument>
        <concept>SNOMEDCT</concept>
      </operation>
    - <operation>
        <type>Selection</type>
        <id>S.1.5</id>
        <input>sao2</input>
        <output>sao2_risk</output>
        <argument>96</argument>
        <concept>VALOR</concept>
      </operation>
    - <operation>
        <type>Transformation</type>
        <id>T.2.2</id>
        <input>sao2_risk</input>
        <output>respiratory_risk</output>
        <argument>96</argument>
        <concept>VALOR</concept>
      </operation>
    </operations>
</ETLconfiguration>
```

Fig. 2 ETL configuration file implemented in XML. ETL, extract, transform, and load.

Results

Methodology for Building ETL for EHR Reuse

The main result obtained in this work has been the design of the methodology for building, in a homogenous and transparent way, ETL processes for EHR reuse, which has been described in detail in the Methods section. ►Figure 4 **shows an overview of the implementation process of this methodology**. Likewise, its application on several COVID-19 data-driven projects at H12O has originated two main implementation resources: a purpose-agnostic catalog of data operations and an automated process for ETL instantiation based on this first resource. They have been designed for being transferable to other clinical conditions, and to any health care organizations. These resources, described below, are accessible on the software repository of the H12O Data Science Unit.[45]

Agnostic Catalog of Data Operations

The first implementation resource obtained was the identification, development, and formalization of the necessary data operations to be applied on the EHR for obtaining research datasets, being use-case agnostic, i.e., data sources, output data model, clinical condition, and health care organization. This set is composed of 17 operations in total, corresponding to two data extractions, ten data selections, and five data transformations. ►Appendix A describes all of them, showing for each operation a real example and the projects that implemented them. This was followed by the formalization, through the DCM approach, of these data operations, indicating the valid data types and cardinality of their inputs and outputs, as shown in ►Appendix B. Finally, the development of each data operation was carried out, through SQL and R languages, whose source code is included in **Supplementary Files**.

Automated Process of ETL Instantiation

The second implantation resource obtained was an automated process for the instantiation of data operations according to specific use cases. This process is based on an XML file, composed of connections, sources, and operations components, which allows the interaction with the databases of the different information systems, and even from different organizations. This file makes the ETL flexible to the data sources, the constraints and clinical condition of the research model to be obtained, and the regulations established in each specific use case. Its formal definition, through the XSD language, is included in **Supplementary Files**, as well as the algorithm developed in R to process the XML file and instantiate the data operations.

Discussion

In this study, a flexible methodology has been designed and implemented for building transparent ETL processes for extracting and transforming EHR into specific formats for

	alias <chr>	driver <chr>	url <chr>	url_driver <chr>	user <chr>	password <chr>
connection	HCIS_3106	oracle.jdbc.OracleDriver	d:/	jdbc:oracle:thin:	user	password

1 row

	connection <chr>	domain <chr>	table <chr>	attributes <chr>
source	HCIS_3106	OBSERVATION	HDOC_INFOBANCO_OBX	CODIGOPACIENTEAUTORVISITAOBXVALORFECHA

1 row

	type <chr>	id <chr>	input <chr>	output <chr>	argument <chr>	concept <chr>
operation	Extraction	E.1.1	HDOC_INFOBANCO_OBX	OBSERVATION	HDOC_COHORTE_COVID19	character(0)
operation.1	Selection	S.1.1	OBSERVATION	sao2	SNOMEDCT:103228002	OBX
operation.2	Selection	S.1.5	sao2	sao2_risk	96	VALOR
operation.3	Transformation	T.2.2	sao2_risk	respiratory_risk	96	VALOR

4 rows

```
S.1.1={ #Selection by CV
     argument<-unlist(str_split(argument,","))
     output<-input[(input[[concept]]%in%argument),]
     },
S.1.5={ #Selection of data less than value
     if(str_detect(argument,",")==TRUE){
       argument<-unlist(str_split(argument,","))
     }
     argument<-as.numeric(argument)
     input[[concept]]<-as.numeric(input[[concept]])
     output<-filter(input, input[[concept]]<argument)
   },
T.2.2={ #Semantic inference
     print(paste("Inferencia por debajo de ", argument,":", sep = ""))
     low_value <- readline()
     print(paste("Inferencia mayor o igual que ", argument,":", sep = ""))
     high_value <- readline()
     output <- input %>% mutate(INFERENCE = case_when(VALOR < argument ~ low_value,
                                                      VALOR >= argument ~ high_value))
   }
```

Fig. 3 Example of extracted parameters and R code of ETL process. ETL, extract, transform, and load.

research and other secondary uses. These processes for obtaining data for purposes additional to health care must follow the FAIR Principles of reuse,[4,17] which establishes that data must be richly described through precise and consistent attributes, as well as the process of recording and preparing them. For this purpose, the DCM-based methodology previously proposed by Pedrera-Jiménez et al, for obtaining EHR-derived datasets for research[5] was used, which is composed of four phases: (1) health condition analysis and specification of important variables, (2) modeling and formalization of the concepts of the clinical domain, (3) definition of rules to generate EHR-derived models, and (4) implementation and validation of the methodology. This study allowed H12O to participate, during the COVID-19 pandemic, in different international health data initiatives such as the ISARIC Consortium, the EHDEN Consortium, the 4CE Consortium, and the TriNetX Platform,[9–12] without requiring manual recording effort parallel to COVID-19 patient health care.[20] However, it is necessary to continue its development to improve each of its component phases. Thus, in this work the third phase of this methodology is expanded, with the aim of making the ETL processes for EHR reuse understandable, auditable, and reproducible,[16] a fundamental requirement for relying on studies that use nonmanually collected data in a controlled and audited manner into study-specific information systems.[15] It is important to highlight that, although guidelines are given for the suitability of EHR for this methodology, data quality aspects such as completeness, correctness, or timeliness[7] are outside the scope of this study for obtaining automated transparent ETL processes. These issues regarding to the EHR quality, and therefore the validity, usefulness, and acceptability of the complete methodology for EHR reuse, are being addressed in next studies of this research line.[46]

The implementation of these EHR extraction and transformation processes has been addressed in several projects through different technologies. Thus, the TRANSFoRm project makes use of the ISO 13606 standard to build a standardized platform for querying and extracting the data, as well as its transformation into health research standards such as CDISC.[38,47] Similarly, the CLIN-IK-LINKS platform allows the development of reusable processes for the transformation of standardized EHR data through the implementation of mappings with XQuery technology.[48] On the other hand, the studies, developed by Sun et al[49] and by Pacaci et al,[50] propose methodologies based on semantic web, through the data representation with Resource Description Framework (RDF), and semantic conversions expressed through Notation 3 (N3) rules. Finally, the Dynamic-ETL project builds a process composed of[51]: (1) complete ETL specifications, described in narrative text and diagrams; (2) D-ETL rules implemented in CSV, which ensures that the rules are human-readable and thus easily examined, maintained, shared, and reused; (3) an ETL rules engine that generates complete SQL statements from the CSV file; and (4) access to the self-generated ETL code to execute, test, and debug the rules. Compared to these previous methodologies, based on technical implementations, this work provides the abstract set of data operations needed to reuse the EHR, its formalization through the DCM paradigm,[22] its use-case agnostic development, and its automated instantiation by means of a formally defined file. This means that this study does not intend to replace these previous methodologies, but rather to propose a common framework that can be used to

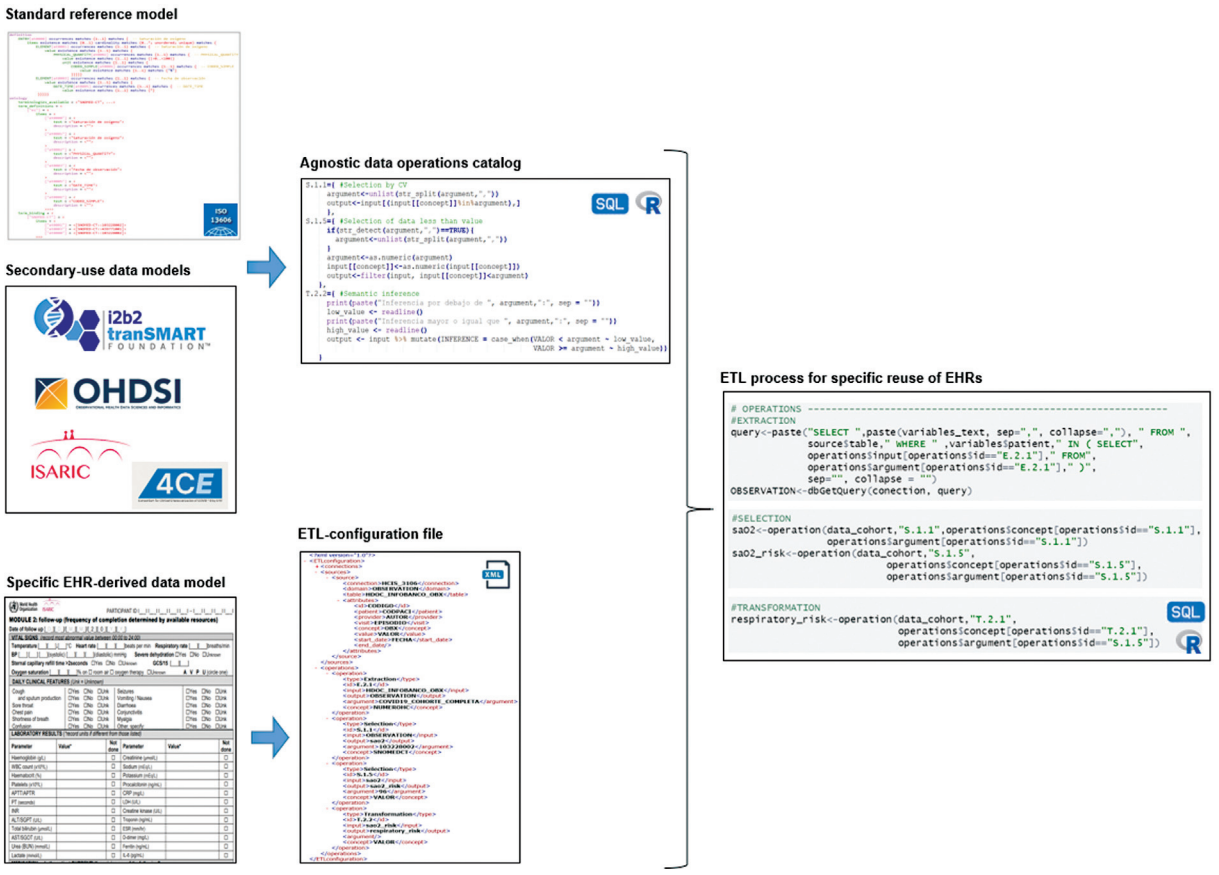

Fig. 4 Overview of the methodology implementation process.

homogenize each of their stages, making them understandable, auditable, and reproducible by technical and functional reviewers.

The identification of the different data operations has been performed by analyzing several data projects of international relevance in which H12O has participated reusing its EHR.[9–12] These projects respond to data models for secondary use of different designs and purposes, such as clinical repository models, CRF models, and aggregated datasets. For the purpose of grouping data operations by typology, a set of progenitor operations were defined, being "Extraction," consisting of "Extraction with criteria" and "Extraction without criteria"; "Selection," consisting of "Selection with reference" and "Selection without reference"; and "Transformation," consisting of "Transformation without semantic implication" and "Transformation with semantic implication." Not all operations were identified and applied to the different typologies of data models. Thus, clinical repository models required data extraction with and without criteria, selections without reference, and transformations without semantic implication. In contrast, CRF and aggregate datasets needed to incorporate, in addition to those mentioned above, selection with reference and transformation with semantic implication. This explains why previous EHR reutilization methodologies have focused on transformation from EHR to standardized repository models such as OMOP CDM,[49–51] instead of complex constraint models such as CRFs and aggregate datasets. Likewise, the identified set of data operations is only an initial specification that may be extended after application of the methodology to new use cases. Therefore, as a continuation of the work, lower level data operations will be identified, e.g., common changes in units of measurement, or published mappings on a terminology server, as well as new typologies of operations through the analysis of data models for different purposes (for COVID-19 and other health conditions), e.g., the evaluation of health outcomes through the standard sets proposed by ICHOM (International Consortium for Health Outcomes Measurement).[52]

Additionally, the formalization of the EHR information models and data operations has been carried out through the paradigm of the DCM and, specifically, making use of the reference model of the ISO 13606 standard,[23,24] due to its stability and adaptability, as well as its adoption as a reference standard by Health Ministry of Spain and H12O. Firstly, the set of clinical archetypes was designed to define the common data and content model to be extracted from the health care information systems for subsequent reutilization. Similarly, the formalization of the data operations was carried out by specifying the data types valid for the arguments, inputs, and outputs of the operations, as well as their cardinality. The use of this common framework makes the operations, and their constraints, understandable by any organization wishing to incorporate them into its EHR reuse process, at whatever point in the process it deems necessary.[47–51] However, a future step of this work will be the

formalization, not only of the input of the operations, but also of the outputs after their application as well. At this point in the development of the methodology, the loading of data into research databases is contemplated as a manual process once the data have been obtained in accordance with the requirements of the output model. To improve this, new *Loading* type operations will be designed based on raw data formats, e.g., CSV, as well as research standards such as CDISC,[38] due to its compatibility with widely used CRF platforms like REDCap,[37] and its adoption by the European Medicines Agency[53] and the American Food and Drug Administration.[54]

Finally, the development of the operations was carried out using data-oriented query and programming languages widely used in the health care environment by technical and health care professionals such as SQL and R.[40,41] Thus, a multipurpose catalog of operations, agnostic to specific use cases, was built. As an extension of this work, more advanced languages, such as AQL,[42] will be used to extract and process the data based on the EHR clinical archetypes. With this, a configuration file of the ETL process was designed and implemented through XML. It is composed of the *connection, source*, and *operation* components, which are processed to extract the parameters for the instantiation of the operations according to the restrictions of the data sources and the data models to be obtained. This file allows clinical researchers to configure and understand the data extraction, selection, and transformation process. Moreover, the process can be adapted to any organization and health condition, just by configuring the parameters of the components of *connection* and *sources*. The configuration of these components and the execution of the process within the organization itself guarantee compliance with the security and privacy measures imposed by the regulations in force in the scenario where it is applied. An improvement to this process will be the implementation of a graphical interface tool that abstracts the user to fill in the XML file directly in a text editor, making it more accessible to nontechnical personnel as previous works have done.[47,48,51]

Conclusions

This study has provided a novel solution to the difficulty of making the ETL processes for obtaining EHR-derived data for secondary use understandable, auditable, and reproducible. Thus, a transparent and flexible methodology was designed based on open standards and technologies, applicable to any clinical condition and health care organization, and even to EHR reuse processes already in place. The proposed methodology was divided into four stages. First, four health data projects of international relevance were analyzed to identify the operations necessary to obtain them from standardized EHR. This made it possible to identify a total of 17 final data operations, classified into categories according to their typology, being two extractions, ten selections, and five transformations. With this, each of the data operations previously identified was formalized. For this purpose, the ISO 13606 reference model was used, specifying the valid data types as arguments, inputs and outputs, and their cardinality. Then, the agnostic catalog of data operations previously identified and formalized was developed through SQL and R languages. Finally, an automated instantiation process of the data operations was built from a formal configuration file implemented in XML. The conclusion drawn from this approach is that the methodology makes the EHR reuse processes transparent and flexible even towards data models with complex constraints, and independent of the health condition, source information systems, health care organization, and regulatory agencies. Moreover, the use-case agnostic abstraction carried on the data operations means that any deployed EHR reuse initiative can incorporate the implemented resources, at any stage of its process, as a common convergence framework.

Funding

Ministerio de Economía y Competitividad Instituto de Salud Carlos III PI18/00981 PI18/01047 PI18CIII/00019

Conflict of Interest

None declared.

Acknowledgment

This work is part of the IMPaCT Data project of the Carlos III Health Institute (ISCIII), as a reference implementation of the process of effective reuse of electronic health records for research and other secondary purposes.

Data Science Unit, Instituto de Investigación Sanitaria Hospital Universitario 12 de Octubre is supported by "Arquitectura normalizada de datos clínicos para la generación de infobancos y su uso secundario en investigación: caso de uso cáncer de mama, cérvix y útero, y evaluación" PI18/00981, and "Infobanco para uso secundario de datos de salud basado en estándares de tecnología y conocimiento: evaluación de la calidad, validez y utilidad de la HCE como origen de datos para el estudio de la infección por VIH" PI18/01047; and Digital Health Research Unit, ISCIII is supported by PI18CIII/00019 "Arquitectura normalizada de datos clínicos para la generación de infobancos y su uso secundario en investigación: solución tecnológica"; funded by the ISCIII from the Spanish National plan for Scientific and Technical Research and Innovation 2017–2020 and the European Regional Development Funds (FEDER).

We would like to thank the rest of the H12O Data Science team for their support and efforts in the data projects described in this article: Blanca Baselga Penalva, Tomás González González, Cristina Díaz Martín, and Bruno Díez Buitrago.

References

1 Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 2008;77(05):291–304

2 Safran C, Bloomrosen M, Hammond WE, et al; Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(01):1–9

3 Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc 2007;14(06):687–696
4 Parra-Calderón CL, Sanz F, McIntosh LD. The challenge of the effective implementation of FAIR principles in biomedical research. Methods Inf Med 2020;59(4-05):117–118
5 Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, et al. Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. J Biomed Inform 2021;115:103697
6 Michaels M, Syed S, Lober WB. Blueprint for aligned data exchange for research and public health. J Am Med Inform Assoc 2021;28 (12):2702–2706
7 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20(01):144–151
8 Makady A, de Boer A, Hillege H, Klungel O, Goettsch W, ; (on behalf of GetReal Work Package 1) What is real-world data? A review of definitions based on literature and stakeholder interviews. Value Health 2017;20(07):858–865
9 Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med 2020;3:109
10 EHDEN Consortium Accessed March 14, 2022, at: https://www.ehden.eu/
11 Pedrera-Jimenez M, Garcia-Barrio N, Hernandez-Ibarburu G, et al. Building an i2b2-based population repository for COVID-19 research. Stud Health Technol Inform 2022;294:287–291
12 ISARIC Clinical Characterisation Group. The value of open-source clinical science in pandemic response: lessons from ISARIC. Lancet Infect Dis 2021;21(12):1623–1624 [published correction appears in Lancet Infect Dis 2021 Dec;21(12):e363]
13 Mehra MR, Desai SS, Ruschitzka F, Patel AN. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. Lancet 2020:S0140-6736(20)31180-6
14 Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Cardiovascular disease, drug therapy, and mortality in Covid-19. N Engl J Med 2020;382(25):e102 [retracted in: N Engl J Med 2020 Jun 4]
15 Kohane IS, Aronow BJ, Avillach P, et al; Consortium For Clinical Characterization Of COVID-19 By EHR (4CE) What every reader should know about studies using electronic health record data but may be afraid to ask. J Med Internet Res 2021;23(03):e22219
16 Pedrera M, Garcia N, Rubio P, Cruz JL, Bernal JL, Serrano P. Making EHRs reusable: a common framework of data operations. Stud Health Technol Inform 2021;287:129–133
17 FAIR Principles Accessed March 14, 2022, at: https://www.go-fair.org/fair-principles/
18 Force11 Accessed March 14, 2022, at: https://force11.org/
19 European Commission. Cost of Not Having FAIR Research Data - Cost-Benefit Analysis for FAIR Research Data. Brussels: European Commission; 2018
20 Pedrera M, Garcia N, Blanco A, et al. Use of EHRs in a tertiary hospital during COVID-19 pandemic: a multi-purpose approach based on standards. Stud Health Technol Inform 2021; 281:28–32
21 Blobel B. Advanced and secure architectural EHR approaches. Int J Med Inform 2006;75(3–4):185–190
22 Beale T. Archetypes: Constraint-based domain models for future-proof information systems. Eleventh OOPSLA Workshop on Behavioral Semantics: Serving the Customer (Seattle, Washington, USA, November 4, 2002). Edited by Kenneth Baclawski and Haim Kilov. Northeastern University, Boston, 2002, pp. 16-32
23 ISO 13606 Standard, Part 1: Reference model. Accessed March 14, 2022, at: https://www.iso.org/standard/67868.html
24 ISO 13606 Standard, Part 2: Archetype model. Accessed March 14, 2022, at: https://www.iso.org/standard/62305.html
25 Muñoz A, Somolinos R, Pascual M, et al. Proof-of-concept design and development of an EN13606-based electronic health care record service. J Am Med Inform Assoc 2007;14(01):118–129
26 Goossen W. Representing knowledge, data and concepts for EHRS using DCM. Stud Health Technol Inform 2011;169:774–778
27 Maldonado JA, Moner D, Boscá D, Fernández-Breis JT, Angulo C, Robles M. LinkEHR-Ed: a multi-reference model archetype editor based on formal semantics. Int J Med Inform 2009;78(08): 559–570
28 Lozano-Rubí R, Muñoz Carrero A, Serrano Balazote P, Pastor X. OntoCR: a CEN/ISO-13606 clinical repository based on ontologies. J Biomed Inform 2016;60:224–233
29 Health Ministry of Spain, Clinical Modeling Resources Accessed March 14, 2022, at: https://www.sanidad.gob.es/profesionales/hcdsns/areaRecursosSem/Rec_mod_clinico_arquetipos.htm
30 Pedrera M, Serrano P, Terriza A, et al. Defining a standardized information model for multi-source representation of breast cancer data. Stud Health Technol Inform 2020;270:1243–1244
31 Coyle JF, Mori AR, Huff SM. Standards for detailed clinical models as the basis for medical data exchange and decision support. Int J Med Inform 2003;69(2–3):157–174
32 Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006; 121:279–290
33 McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003;49(04):624–633
34 Electronic Health Records Archetypes of Hospital Universitario 12 de Octubre. Accessed March 14, 2022, at: https://www.safecreative.org/work/2102196969593-h12o-covid-19-observations-archetypes
35 Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(02):124–130
36 Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574–578
37 Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(02): 377–381
38 CDISC. .Foundational. Accessed March 14, 2022, at: https://www.cdisc.org/standards/foundational
39 ISARIC-WHO CRF for COVID-19 Accessed March 14, 2022, at: https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf/
40 Structured Query Language (SQL) Accessed March 14, 2021, at: https://www.w3schools.com/sql/
41 Project R Accessed March 14, 2022, at: https://www.r-project.org/
42 Archetype Query Language (AQL) Accessed March 14, 2022, at: https://specifications.openehr.org/releases/QUERY/latest/AQL.html
43 Ramos M, Sánchez-de-Madariaga R, Barros J, et al. An archetype query language interpreter into MongoDB: managing NoSQL standardized electronic health record extracts systems. J Biomed Inform 2020;101:103339
44 eXtensible Markup Language (XML). Accessed March 14, 2022, at: https://www.w3schools.com/xml/
45 H12O Data Science software repository. Accessed March 14, 2022, at: https://github.com/DataDoce/EHR-Data-Operations
46 Pedrera-Jimenez M, Garcia-Barrio N, Rubio-Mayo P, et al. Making EHRs trustable: a quality analysis of EHR-derived datasets for COVID-19 research. Stud Health Technol Inform 2022; 294:164–168
47 Lim Choi Keung SN, Zhao L, Rossiter J, et al. Detailed clinical modelling approach to data extraction from heterogeneous data

sources for clinical research. AMIA Jt Summits Transl Sci Proc 2014;2014:55–59
48 Maldonado JA, Marcos M, Fernández-Breis JT, Giménez-Solano VM, Legaz-García MDC, Martínez-Salvador B. CLIN-IK-LINKS: a platform for the design and execution of clinical data transformation and reasoning workflows. Comput Methods Programs Biomed 2020;197:105616
49 Sun H, Depraetere K, De Roo J, et al. Semantic processing of EHR data for clinical research. J Biomed Inform 2015;58:247–259
50 Pacaci A, Gonul S, Sinaci AA, Yuksel M, Laleci Erturkmen GB. A semantic transformation methodology for the secondary use of observational healthcare data in postmarketing safety studies. Front Pharmacol 2018;9:435
51 Ong TC, Kahn MG, Kwan BM, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. BMC Med Inform Decis Mak 2017;17(01):134
52 ICHOM. Standard sets. Accessed March 14, 2022, at: https://www.ichom.org/healthcare-standardization/
53 European Medicines Agency. European Medicines Regulatory Network data standardisation strategy. Accessed March 14, 2022, at: https://www.ema.europa.eu/en/documents/other/european-medicines-regulatory-network-data-standardisation-strategy_en.pdf
54 CDISC. Global regulatory requirements. Accessed March 14, 2022, at: https://www.cdisc.org/resources/global-regulatory-requirements

Appendix A: Identified Data Operations for EHR Reuse

Table A.1 Identified data operations for EHR reuse; project: (1) TriNetX Platform, (2) EHDEN Consortium, (3) ISARIC Consortium, and (4) 4CE Consortium

ID	Operation	Example	Project
E	Extraction	–	–
E.1	Extraction with criteria	–	–
E.1.1	Extraction of data for *patient cohort*	Observations of COVID-19 patients	All
E.2	Extraction without criteria	–	–
E.2.1	Extraction of data for all patients	Data from Observation domain	1, 2
S	Selection	–	–
S.1	Selection with reference	–	–
S.1.1	Selection of data related to *concept*	Data related to COVID-19 test results	All
S.1.2	Selection of data previous to *date time*	Prehospitalization medication	All
S.1.3	Selection of data after *date time*	Medication during hospitalization	All
S.1.4	Selection of data higher than *value*	Temperatures higher than 37°C	3, 4
S.1.5	Selection of data less than *value*	Oxygen saturations less than 96%	3, 4
S.1.6	Selection of data equal to *value*	COVID-19 test results equal to “Positive”	3, 4
S.2	Selection without reference	–	–
S.2.1	Selection of most recent datum	Last COVID-19 test result	3, 4
S.2.2	Selection of oldest datum	First Oxygen saturation on admission	3, 4
S.2.3	Selection of datum with higher value	Higher Temperature	3, 4
S.2.4	Selection of datum with lower value	Lower Oxygen saturation	3, 4
T	Transformation	–	–
T.1	Transformation without semantic implication	–	–
T.1.1	Change of unit of measure	C-Reactive Protein from mg/dL to mg/L	All
T.1.2	Change of coding system	Cough from local code to SNOMED CT	All
T.2	Transformation with semantic implication	–	–
T.2.1	Mathematical operation	BMI from Weight and Height	3, 4
T.2.2	Semantic inference	Fever from Temperature	3, 4
T.2.3	Event count	Number of previous hospitalizations	3, 4

Appendix B: Formalized Data Operations for EHR Reuse

Table B.1 Formalized data operations for EHR reuse

Operation ID	Argument data type	Input data type	Output data type	Input card.	Output card.
E.1.1	II	All	Same than Input	1..N	1..N
E.1.2	II	All	Same than Input	1..N	1..N
S.1.1	CV	All	Same than Input	1..N	1..N
S.1.2	DATETIME	All	Same than Input	1..N	1..N
S.1.3	DATETIME	All	Same than Input	1..N	1..N
S.1.4	PQ, INTEGER	PQ, INTEGER	Same than Input	1..N	1..N
S.1.5	PQ, INTEGER	PQ, INTEGER	Same than Input	1..N	1..N
S.1.6	CV, PQ, INTEGER	CV, PQ, INTEGER	Same than Input	1..N	1..N
S.2.1	–	All	Same than Input	1..N	1..1
S.2.2	–	All	Same than Input	1..N	1..1
S.2.3	–	PQ, INTEGER	Same than Input	1..N	1..1
S.2.4	–	PQ, INTEGER	Same than Input	1..N	1..1
T.1.1	–	PQ	PQ	1..1	1..1
T.1.2	–	CV	CV	1..1	1..1
T.2.1	–	PQ, INTEGER	Same than Input	1..N	1..1
T.2.2	–	All	All	1..N	1..1
T.2.3	–	All	INTEGER	1..N	1..1