Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210831

Making EHRs Reusable: A Common Framework of Data Operations

Miguel PEDRERA[a,b,1], Noelia GARCIA[a], Paula RUBIO[a], Juan Luis CRUZ[a], José Luis BERNAL[a] and Pablo SERRANO[a]
[a]*Hospital Universitario 12 de Octubre, Madrid, Spain*
[b]*ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain*

Abstract. Reuse of EHRs requires data extraction and transformation processes are based on homogeneous and formalized operations in order to make them understandable, reproducible and auditable. This work aims to define a common framework of data operations for obtaining EHR-derived datasets for secondary use. Thus, 21 operations were identified from different data-driven projects of a 1,300-beds tertiary Hospital. Then, ISO 13606 standard was used to formalize them. This work is the starting point to homogenize ETL processes for the reuse of EHRs, applicable to any condition and organization. In future studies, defined data operations will be implemented and validated in projects of different purposes.

Keywords. Electronic Health Records, FAIR, Data reusability, Real World Data, Semantics, Standards, ISO 13606, i2b2, OMOP, ISARIC, COVID-19.

1. Introduction

Electronic Health Record (EHR) is defined as the repository of health data that is generated throughout a patient's lifetime. Its primary use is to enable continuous, efficient and quality healthcare [1]. Additionally, there are other uses of EHR, known as secondary uses, including activities such as clinical research or public health [2]. These further uses are only possible if we produce reusable EHR data, which is one of the principles established by FAIR [3].

Reusability is determined by how we manage the semantics of concepts and (meta)data in information systems. A first step is provided by Detailed Clinical Models (DCM), which allow implementing mechanisms for obtaining EHR-derived datasets for secondary use [4, 5]. However, it is essential that the extraction, transformation and loading processes (ETLs) are based on homogeneous and formalized operations, in order to make them understandable, reproducible and auditable [6].

Thus, this work aims to define a common framework of data operations on EHRs, necessary for them to be adequately reusable for secondary purposes.

[1] Corresponding Author, Miguel Pedrera Jiménez, Hospital Universitario 12 de Octubre, Av. de Córdoba s/n, 28041 Madrid Spain; E-mail: miguel.pedrera@salud.madrid.org.

2. Methods

This work was carried out at Hospital Universitario 12 de Octubre (H12O) in Madrid (Spain), as part of its research line on the effective reuse of EHRs [4, 7, 8].

2.1. Detailed Clinical Models

In this study, DCM were used as the basis for the design and formalization of data operations. This paradigm proposes a dual model composed of a reference model and an archetype model [9]. Thus, ISO 13606 standard [10], previously adopted by H12O and Spanish Ministry of Health, was selected for this purpose.

The reference model of this standard defines the components for building an interoperable EHR: Folder, Composition, Section, Entry, Cluster and Element. It also establishes the types of data permitted, which allows limiting the valid data types for each operation. In the present work, it was necessary to use the following subset:

- **Coded Value (CV)**: for concepts whose result is a set of possible coded values, e.g., SARS-COV-2 test, which may be positive, negative or inconclusive;
- **Physical Quantity (PQ)**: for concepts whose outcome is a numerical value with unit of measurement, e.g., oxygen flow rate measured in liters per minute;
- **Integer**: for concepts whose result is an integer value, e.g., Glasgow Comma Scale score; and,
- **Date Time**: for concepts whose value is a time point, e.g., date of symptom onset.

Likewise, the archetype model allows the information models to be formalized and linked with terminologies at two levels: the semantic binding, for specifying the meaning of its components, and the value binding, to define the set of values of a CV element.

A reusable EHR must be supported by appropriate modeling and standardization practices. Therefore, it is necessary to use common reference models and standard terminologies, such as SNOMED CT [11] and LOINC [12], that do not contain miscellaneous, grouped, calculated or inferred concepts. Thus, the execution of formalized data operations on standardized EHR extracts (structure and content) allows ETL process to be applicable regardless of the condition and organization.

2.2. Secondary use data models

Secondary use models allow data to be represented and persisted for other uses in addition to healthcare. Consequently, they are less demanding than primary use models in terms of metadata about the registration process or access permissions. We can distinguish two types of secondary use models:

- **Clinical data repositories.** These models centralize data from different sources within a common structure and content. They have not been modeled for a single purpose, but as a data warehouse for multiple secondary uses, e.g., i2b2 (tranSMART Foundation) [13], used in TriNetX Platform (federated network for clinical trials) [14], and OMOP CDM (OHDSI) [15], used in EHDEN Consortium (federated network for observational research) [16].
- **Electronic Data Capture systems (EDC).** These models collect data as it is expected to be analyzed. They are designed according to specific use cases, e.g.,

ISARIC Case Report Form (CRF) for COVID-19 [17] and STOP-CORONAVIRUS EDC [18].

In order to define the set of common data operations, the different models that have been used for different data-driven projects in H12O were analyzed, considering both typologies. Table 1 shows the list of specifications reviewed.

Table 1. 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 analytics
2	EHDEN Consortium	OMOP repository	Observational studies
3	ISARIC Consortium	Specific EDC	Case reports and analytics
4	STOP-CORONAVIRUS	Specific EDC	Observational studies

2.3. Identification and formalization of data operations

Once the data models of the different projects have been analyzed, data operations were identified and then classified according to several categories. These high-level operations, parents of the fully defined operations (FDO), were as follows:

- **Selection (S).** Operations to select and extract the required data under the restrictions of the secondary use model. Two subtypes were defined:
  - **Selection with reference (S.1),** e.g., selection of "Oxygen saturations" less than 96%.
  - **Selection without reference (S.2),** e.g., selection of the "Oxygen saturation" with the lowest value.
- **Transformation (T).** Operations to transform the data to the format of the secondary use model. Two subtypes were defined:
  - **Transformation maintaining meaning (T.1),** e.g., changing the measurement unit of a concept "C-Reactive Protein" from mg/dL to mg/L.
  - **Transformation altering meaning (T.2),** e.g., calculating a "BMI" concept from "Weight" and "Height".

Operations were formalized by specifying the valid data types and the cardinality of the argument, input and output of them. For this purpose, the data types specified in the ISO 13606 reference model were employed.

3. Results

3.1. Identification of data operations

The first result obtained was the set of data operations, classified according to the categories defined in the methodology section. Table 2 shows this specification, indicating, for each FDO, an example and the projects that required them.

Table 2. Identified data operations for EHRs reuse.

ID	Operation	Example	Project
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*	Pre-hospitalization medication	All
S.1.3	Selection of data after *date*	Medication during hospitalization	All
S.1.4	Selection of data higher than *value*	Temperatures higher than 37 ºC	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 maintaining meaning	-	-
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 altering meaning	-	-
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

3.2. Formalization of data operations

The second result was the formalized set of FDO. To this end, data types for argument, input and output of operations were specified according to ISO 13606, as well as the cardinality (arguments have unique cardinality). Table 3 shows this specification.

Table 3. Formalized data operations for EHRs reuse.

Operation ID	Argument Datatype	Input Data type	Output Data type	Input Card.	Output Card.
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

4. Conclusions

In this study, a common framework of data operations was theoretically defined for obtaining secondary use models from EHRs. For this purpose, four data-driven projects in which H12O participates were studied (Table 1).

Thus, 21 operations were identified, 15 of which were FDO (Table 2). Data models related to standardized repositories did not involve complex operations. However, specific data models for COVID-19 research required selections with complex criteria and meaning-altering transformations. The set of FDO was formalized (data types and cardinality) using ISO 13606 standard reference model (Table 3). This allows

implementing homogeneous ETL processes based on common criteria and identifying processes with inconsistent operations (e.g., a 'unit change' operation on a CV variable). Moreover, these operations can be adapted in accordance to data sources and secondary use models, being applicable to other organizations and health conditions.

In future studies, data operations will be implemented with programming languages such as R, and validated in COVID-19 projects and studies of other clinical conditions.

Acknowledgment

This work has been supported by Research Projects PI18/00981 and PI18/01047 funded by Instituto de Salud Carlos III, co-funded by ERDF/ESF.

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;77:291–304. doi:10.1016/j.ijmedinf.2007.09.001.
[2] Safran C, Bloomrosen M, Hammond E, et al. Toward a National Framework for the Secondary Use of Health. Jounal Am Med Informatics Assoc 2007;14:1–9. doi:10.1197/jamia.M2273.Introduction.
[3] FAIR Principle R1.3. https://www.go-fair.org/fair-principles/r1-3-metadata-meet-domain-relevant-community-standards/. Accessed July 30, 2021.
[4] 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. doi:10.1016/j.jbi.2021.103697.
[5] Lim Choi Keung S, 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 proceedings AMIA Jt Summits Transl Sci 2014;2014:55–9. doi:10.1016/j.ic.2014.12.007.
[6] Kohane IS, Aronow BJ, Avillach P, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021;23(3):e22219. Published 2021 Mar 2. doi:10.2196/22219.
[7] 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. doi:10.3233/SHTI210114.
[8] González L, Pérez-Rey D, Alonso E, et al. Building an I2B2-Based Population Repository for Clinical Research. Stud Health Technol Inform. 2020;270:78-82. doi:10.3233/SHTI200126.
[9] Beale T. Archetypes: Constraint-based Domain Models for Future- proof Information Systems. OOPSLA 2002 Work Behav Semant 2001;:1–69. doi:10.1.1.147.8835.
[10] ISO 13606 standard. https://www.iso.org/standard/67868.html. Accessed July 30, 2021.
[11] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279-290.
[12] McDonald CJ, Huff SM, Suico JG, *et al.* LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clin Chem* 2003;49:624–33. doi:10.1373/49.4.624.
[13] 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(2):124-130. doi:10.1136/jamia.2009.000893.
[14] TriNetX Platform. https://trinetx.com/. Accessed July 30, 2021.
[15] 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.
[16] EHDEN Consortium. https://www.ehden.eu/. Accessed July 30, 2021.
[17] ISARIC-WHO CRF for COVID-19. https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf/. Accessed July 30, 2021.
[18] STOP-CORONAVIRUS. https://imas12.es/blog/stop-coronavirus-nuevo-proyecto-clinico-llevado-a-cabo-en-el-instituto-i12-para-ofrecer-respuestas-integrales-a-la-covid-19/. Accessed July 30, 2021.