Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biomedical Informatics 115 (2021) 103697

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models

Miguel Pedrera-Jiménez [a,b,*], Noelia García-Barrio [a], Jaime Cruz-Rojo [a], Ana Isabel Terriza-Torres [a], Elena Ana López-Jiménez [a], Fernando Calvo-Boyero [a], María Jesús Jiménez-Cerezo [a], Alvar Javier Blanco-Martínez [a], Gustavo Roig-Domínguez [a], Juan Luis Cruz-Bermúdez [a], José Luis Bernal-Sobrino [a], Pablo Serrano-Balazote [a], Adolfo Muñoz-Carrero [c]

[a] Hospital Universitario 12 de Octubre, Av. de Córdoba, s/n, 28041 Madrid, Spain
[b] ETSI Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain
[c] Digital Health Research Dept., Instituto de Salud Carlos III, Av. de Monforte de Lemos, 5, 28029 Madrid, Spain

ARTICLE INFO

Keywords:
Detailed clinical models
COVID-19
Electronic health records
Semantics
Standards
Real world data

ABSTRACT

Background: COVID-19 ranks as the single largest health incident worldwide in decades. In such a scenario, electronic health records (EHRs) should provide a timely response to healthcare needs and to data uses that go beyond direct medical care and are known as secondary uses, which include biomedical research. However, it is usual for each data analysis initiative to define its own information model in line with its requirements. These specifications share clinical concepts, but differ in format and recording criteria, something that creates data entry redundancy in multiple electronic data capture systems (EDCs) with the consequent investment of effort and time by the organization.

Objective: This study sought to design and implement a flexible methodology based on detailed clinical models (DCM), which would enable EHRs generated in a tertiary hospital to be effectively reused without loss of meaning and within a short time.

Material and methods: The proposed methodology comprises four stages: (1) specification of an initial set of relevant variables for COVID-19; (2) modeling and formalization of clinical concepts using ISO 13606 standard and SNOMED CT and LOINC terminologies; (3) definition of transformation rules to generate secondary use models from standardized EHRs and development of them using R language; and (4) implementation and validation of the methodology through the generation of the International Severe Acute Respiratory and emerging Infection Consortium (ISARIC-WHO) COVID-19 case report form. This process has been implemented into a 1300-bed tertiary Hospital for a cohort of 4489 patients hospitalized from 25 February 2020 to 10 September 2020.

Results: An initial and expandable set of relevant concepts for COVID-19 was identified, modeled and formalized using ISO-13606 standard and SNOMED CT and LOINC terminologies. Similarly, an algorithm was designed and implemented with R and then applied to process EHRs in accordance with standardized concepts, transforming them into secondary use models. Lastly, these resources were applied to obtain a data extract conforming to the ISARIC-WHO COVID-19 case report form, without requiring manual data collection. The methodology allowed obtaining the observation domain of this model with a coverage of over 85% of patients in the majority of concepts.

Conclusion: This study has furnished a solution to the difficulty of rapidly and efficiently obtaining EHR-derived data for secondary use in COVID-19, capable of adapting to changes in data specifications and applicable to other

* Corresponding author at: Health Informatics Dept., Hospital Universitario 12 de Octubre, Av. de Córdoba, s/n, 28041 Madrid, Spain.
E-mail addresses: miguel.pedrera@salud.madrid.org (M. Pedrera-Jiménez), ngbarrio@salud.madrid.org (N. García-Barrio), jaime.cruz@salud.madrid.org (J. Cruz-Rojo), anaisabel.terriza@salud.madrid.org (A.I. Terriza-Torres), elenaana.lopez@salud.madrid.org (E.A. López-Jiménez), fernando.calvo@salud.madrid.org (F. Calvo-Boyero), mjcerezo@salud.madrid.org (M.J. Jiménez-Cerezo), alvar.blanco@salud.madrid.org (A.J. Blanco-Martínez), gustavo.roig@salud.madrid.org (G. Roig-Domínguez), juanluis.cruz@salud.madrid.org (J.L. Cruz-Bermúdez), joseluis.bernal@salud.madrid.org (J.L. Bernal-Sobrino), pserranob@salud.madrid.org (P. Serrano-Balazote), adolfo.munoz@isciii.es (A. Muñoz-Carrero).

https://doi.org/10.1016/j.jbi.2021.103697
Received 11 October 2020; Received in revised form 18 December 2020; Accepted 1 February 2021
Available online 3 February 2021

organizations and other health conditions. The conclusion to be drawn from this initial validation is that this DCM-based methodology allows the effective reuse of EHRs generated in a tertiary Hospital during COVID-19 pandemic, with no additional effort or time for the organization and with a greater data scope than that yielded by conventional manual data collection process in ad-hoc EDCs.

1. Introduction

1.1. Background and significance

COVID-19 ranks as the single largest health incident worldwide in decades [1,2], registering over 27,486,960 confirmed cases and 894,983 related deaths around the globe up to 9 September 2020 [3]. This study was undertaken at the Hospital Universitario 12 de Octubre [4], a 1300-bed tertiary Hospital situated in Madrid Region (Spain), where 156,026 confirmed cases and 8817 deaths had been recorded as of 10 September 2020 [5]. During the pandemic, average length of stay at this hospital increased by around 15%. Likewise, the burden of managing COVID-19 patients rose to become an overload that saturated healthcare resources. In such a scenario, electronic health records (EHRs) should provide a timely response to healthcare needs (decision-making, whether for clinical or for resource-planning purposes) [6–8], without generating errors [9]. These needs also extend to data uses that go beyond direct medical care and are known as secondary uses, which include biomedical research [10]. It is usual for each data analysis initiative to define its own information model in line with its data requirements [11]. Although they share clinical concepts, these models differ in format and recording criteria, something that creates data entry redundancy in multiple electronic data capture systems (EDCs). Moreover, in a situation like that caused by a new disease such as COVID-19, data is needed in a short time and advances in research result in data specifications constantly changing. In order to overcome these issues, an innovative methodology, which enables semantics to be incorporated into the process of the reuse of routine healthcare data, must be defined and implemented [12]. In this way, EHRs can be reused for multiple purposes in a brief period and adapted to changes in data specifications, while maintaining their original meaning and an acceptable quality.

Nevertheless, current health information systems incorporate data semantics very poorly, which then hinders their combination and reuse. This is due to the fact they are “single-level” systems, in which the concept model is implicit in the data model. Advanced healthcare information systems and clinical data warehouses such as i2b2 and OMOP [13,14], implement a dual paradigm, which separates the data model and the concept model. This is based on Detailed Clinical Model (DCM) paradigm [15]; in which the reference model defines the set of generic components for constructing interoperable EHRs, and the archetype model formalizes concepts of the clinical domain, constructed by the combination of the components and constraints of the reference model [16]. Some standards applying the dual model are the ISO 13606 standard and the OpenEHR specification [17,18], which has published specific resources for COVID-19 [19]. The archetypes make it possible to define terminology binding that associates each component with standard terminologies, such as Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) and Logical Observation Identifiers Names and Codes (LOINC) [20,21]. SNOMED CT has published several new concepts and related descriptions pertaining to COVID-19 [22], while LOINC has published a set of codes for laboratory tests for the diagnosis of this new disease [23]. In this study, a flexible methodology based on this paradigm is proposed to help resolve existing difficulties arising in the rapid and efficient collection of data for COVID-19 research, considered extensible to other organizations and applicable to other conditions.

1.2. Objectives

The aim of this study was to design and implement a flexible methodology based on the DCM paradigm that would enable EHRs to be effectively reused for COVID-19 secondary uses, without loss of meaning and within a short time. This implies a series of particular objectives, such as:

- specifying an initial and expandable set of relevant variables for COVID-19 on which to apply the methodology;
- selecting and applying the appropriate modeling and terminological standards to the clinical concepts identified;
- defining the necessary transformation rules to generate EHR-derived models from the standard information model; and,
- implementing and validating the methodology through the generation of a data extract in accordance with a validated COVID-19 information model.

2. Material and methods

The proposed methodology should allow the representation and reuse of EHRs on any health condition, with no changes in their original meaning. It is supported by previous studies [24–27], and it comprises four stages:

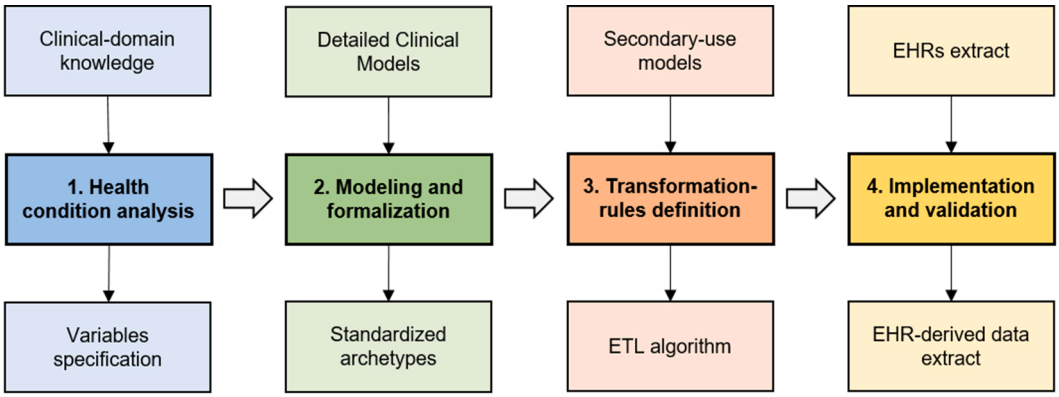

Fig. 1. Stages of the methodology for obtaining EHR-derived data.

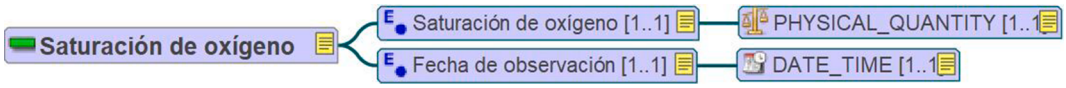

Fig. 2. Mind map of the "Oxygen saturation" ("Saturación de oxígeno" in Spanish) archetype.

1. **health condition analysis and specification of relevant variables**, i.e., analysis and identification of an initial and expandable set of relevant variables for healthcare and secondary use purposes;
2. **modeling and formalization of the concepts of the clinical domain** i.e., making use of resources based on the DCM paradigm to model and formalize the identified concepts;
3. **definition of rules to generate EHR-derived models,** i.e., analysis of secondary use models and design of rules of transformation to them from standardized EHRs; and,
4. **implementation and validation of the methodology** i.e., implementation of EHRs registration, extraction and transformation mechanisms. For validation purposes, a secondary use model is generated, and the data coverage achieved is analyzed.

Fig. 1 depicts the component stages that make up the methodology and the deliverables of each of them.

This study applies the methodology to COVID-19 in order to enhance the efficiency of data collection for the many data initiatives that have arisen around this condition. The methodology is valuable in this

```
definition
    ENTRY[at0000] occurrences matches {1..1} matches {  -- Saturación de oxígeno
        items existence matches {0..1} cardinality matches {0..*; unordered; unique} matches {
            ELEMENT[at0001] occurrences matches {1..1} matches {  -- Saturación de oxígeno
                value existence matches {1..1} matches {
                    PHYSICAL_QUANTITY[at0002] occurrences matches {1..1} matches {  -- PHYSICAL_QUANTITY
                        value existence matches {1..1} matches {|>0..<100|}
                        unit existence matches {1..1} matches {
                            CODED_SIMPLE[at0006] occurrences matches {1..1} matches {  -- CODED_SIMPLE
                                value existence matches {1..1} matches {"%"}
                            }}}}}
            ELEMENT[at0003] occurrences matches {1..1} matches {  -- Fecha de observación
                value existence matches {1..1} matches {
                    DATE_TIME[at0005] occurrences matches {1..1} matches {  -- DATE_TIME
                        value existence matches {1..1} matches {*}
             }}}}}
ontology
    terminologies_available = <"SNOMED-CT", ...>
    term_definitions = <
        ["es"] = <
            items = <
                ["at0000"] = <
                    text = <"Saturación de oxígeno">
                    description = <"">
                >
                ["at0001"] = <
                    text = <"Saturación de oxígeno">
                    description = <"">
                >
                ["at0002"] = <
                    text = <"PHYSICAL_QUANTITY">
                    description = <"">
                >
                ["at0003"] = <
                    text = <"Fecha de observación">
                    description = <"">
                >
                ["at0005"] = <
                    text = <"DATE_TIME">
                    description = <"">
                >
                ["at0006"] = <
                    text = <"CODED_SIMPLE">
                    description = <"">
                >>>>
    term_binding = <
        ["SNOMED-CT"] = <
            items = <
                ["at0001"] = <[SNOMED-CT::103228002]>
                ["at0003"] = <[SNOMED-CT::439771001]>
                ["at0000"] = <[SNOMED-CT::103228002]>
            >>>
```

Fig. 3. Code in ADL of the "Oxygen saturation" ("Saturación de oxígeno" in Spanish) archetype.

pandemic scenario, when data is needed urgently and reference specifications change constantly, providing the timing and flexibility required. This process is innovative compared to that of manual data entry, in which effort and time is proportional to the number of patients to be included, and changes in the secondary use model involve data re-entry.

2.1. Stage 1: condition analysis and specification of relevant variables

To identify the gaps in standardization to which the methodology would be applicable, the different EHR domains in healthcare information systems were analyzed. It could be concluded that evaluations, instructions and actions had adequate modeling and standardization. Observable entities (OE), however, constituted very wide-ranging, heterogeneous sets that render reuse difficult. This is a domain where the DCM paradigm can make a major contribution, since it is essential not only to have codified value-sets, but also to ensure that each clinical domain concept, such as "Oxygen saturation" or "D-dimer", is represented formally without loss of meaning. Although in this case it was not necessary, other EHR domains would proceed the same way, with definitions of more general archetypes such as "Prescription" or "Health problem".

The requirements established for defining the initial set of variables were that it had to cover the necessary span for both patient care and secondary uses, and be parsimonious, since the data were to be recorded in healthcare practice, and it was important not to increase the health professionals' workload [28]. Thus, a work team was created in March 2020, consisting of health professionals attached to the main hospital departments tasked with the care of COVID-19 patients. A total of 58 health OE, 22 clinical and 36 laboratory-related, were identified by this group based on their clinical knowledge and scientific evidence. During this task, the proposed methodology allowed the concept model to be expanded as the medical team identified new relevant variables for COVID-19. In the same way, owing to the fact that COVID-19 is a new disease, these concepts are just an initial set, expandable according to increased understanding of it. DCM provides a real solution to the extension of this initially defined concept model without altering the information systems that implement it.

2.2. Stage 2: Modeling and formalization of concepts

The modeling and formalization of concepts were performed in accordance with the ISO 13606 standard with this being adapted to the technical capacities of the hospital information systems. This standard was used for several reasons: (1) it defines a rigorous and stable information architecture for defining clinical domain concepts and communicating EHRs, (2) it allows adding clinical concepts without altering the databases structure, (3) it has current applications in health organizations through tools based on it [24,29], (4) it is used by the Spanish Ministry of Health and the different Regions as the standard for the definition of exchangeable EHR extracts in the country [30], and (5) it was adopted by the Hospital for the management and governance of the clinical concepts and modeling resources [27].

ISO 13606 standard is based on DCM paradigm and defines a reference model and an archetype model. Its reference model defines the Entry component as "*a result of one clinical action, one observation, one clinical interpretation, or an intention*". This component may, in turn, contain several component Elements, "*The leaf node of the EHR hierarchy, containing a single data value*". Each OE defined in this study was modeled using an entry component such as "Blood pressure", which, at the same time, contains the component elements relating to the specific concepts associated with it: "Systolic blood pressure", "Diastolic blood pressure" and "Mean blood pressure". Lastly, the Entry component contains a component element for representing the date on which the observation was made. ISO 13606 reference model also establishes the types of data permitted accordingly to ISO 21090 [31]. It was necessary to use the following four to cover the requirements of this use case:

- Physical Quantity (PQ): for OE whose result is a numeric value with unit of measurement, e.g., systolic blood pressure measured in mmHg;
- Coded Value (CV): for OE whose result is a set of possible coded values, e.g., the result of the SARS-COV-2 virus detection test, which may be positive, negative or inconclusive;
- Integer: for OE whose result is an integer value, e.g., Glasgow Coma Scale score; and,
- Date Time: for OE whose value is a time point, e.g., date of initiation of smoking habit or date on which an observation was made.

Fig. 2 shows the mind map relating to the archetype "Oxygen saturation" ("Saturación de oxígeno" in Spanish), composed by an Entry and two Elements, "Oxygen saturation" ("Saturación de oxígeno" in Spanish) of Physical Quantity data type and "Observation date" ("Fecha de observación" in Spanish) of Date Time data type.

The archetype model makes use of the above-defined components to formalize the concepts of the clinical domain. On the one hand, the "definition" section specifies the components of the archetype, along with their cardinality, type of data, minimum and maximum values, unit of measurement, codified value-set and other metadata. The full definition of the information model and its constraints ensure the completeness and consistency of EHR extracts [32]. On the other hand, the "ontology" section defines the terminology binding used, incorporating the semantics to the information model. The Archetype Definition Language (ADL) was employed for archetype development using LinkEHR Studio [29]. Fig. 3 shows an ADL code fragment of the "Oxygen saturation" archetype.

Terminology binding was constructed with SNOMED CT and LOINC, since both are internationally adopted, and form part of the semantic specifications issued by the Spanish Ministry of Health [33]. While LOINC was used to represent laboratory OE, e.g., "94315-9 |SARS coronavirus 2 and gene [Presence] in Unspecified specimen by NAA with probe detection", the SNOMED CT 'observable entity' axis was employed to represent concepts of clinical OE, e.g., "103228002 |Hemoglobin saturation with oxygen (observable entity)|". Here, it was necessary to resort to the terminology extension mechanism for five concepts. This allows each SNOMED CT National Reference Center (*Centro Nacional de Referencia/CNR*) to publish its own concepts [34], which are then proposed for inclusion in the international edition of this terminology. Lastly, the SNOMED CT 'finding' and 'qualifier' axes were used for OE responses reporting a set of possible values, e.g., "77176002 | Smoker (finding) |" and "10828004 |Positive (qualifier value)|".

2.3. Stage 3: Definition of secondary use generation rules

Firstly, secondary use models of COVID-19 were studied to quantify the coverage that could be achieved on the basis of the standard concepts defined. If a concept was not covered by the initial specification, the utility of including it in the standard information model was analyzed by the clinical team. Expanding concept model is one of the advantages of a DCM-based methodology.

Following this, the rules to generate EHR-derived models were designed based on the format of these specifications. A total of five data operations were identified, considered applicable to any health condition:

1. **Inference of specific variables from general concepts**, e.g., inferring a yes/no response for an "active smoker" variable from a "smoking habit" concept that assumes "non-smoker", "ex-smoker" and "active smoker" as possible values.
2. **Transformations between coding systems**, e.g., transforming a concept "10828004 |Positive (qualifier value)|" into a local code 'P'.

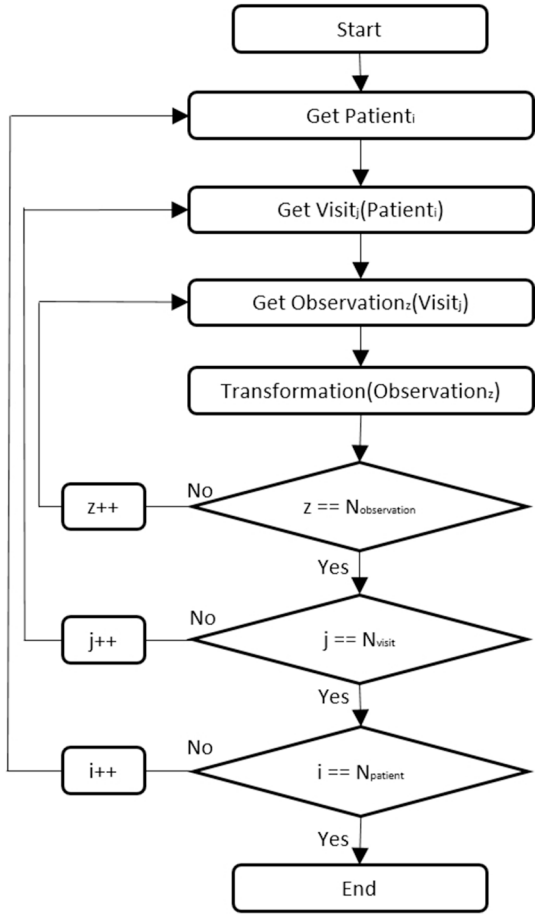

Fig. 4. Iterative algorithm for generation of EHR-derived data extracts.

3. **Transformations between units of measurement**, e.g., transforming a variable "C-Reactive Protein" measured in "mg/dL" into a variable relating to the same concept measured in "mg/L".
4. **Selection according to specific values**, e.g., selecting "Oxygen saturation" with value under 92%.
5. **Selection of data at a given time point**, e.g., selecting "Body temperature" value on admission to hospital.

The transformation rules were documented and shared with the clinical team for review. Once validated, an algorithm was developed in R language, version 3.6.1 [35], which performed the combination of transformations needed to obtain the secondary use model. Fig. 4 shows the flow chart of the algorithm developed. It functions iteratively selecting the data relating to the concepts of interest (index 'i'), from each visit (index 'j'), for each patient (index 'z'), and then applying the abovementioned operations to these.

2.4. Stage 4: Implementation and validation of the methodology

The starting point for the implementation of the methodology was the definition of the clinical archetypes in the multiple hospital information systems affected. For this purpose, the clinical concepts were identified or created in each information system and then mapped (in the system itself) from the local identifier to the standard code defined by the semantic of the archetype. Hence, data are stored following a key-value structure (observable entity-finding), in which each observation is identified in a standard and homogeneous way. This mechanism enables data to be extracted from the different systems for reuse, while maintaining their meaning unaltered and ensuring acceptable data quality: clinical archetypes are used to guarantee completeness and consistency

```
<DATA_RECORD>
   <CODE>103228002</CODE>
   <DESCRIPTION>Saturación de oxígeno</DESCRIPTION>
   <DATE>2020/05/16</DATE>
   <TIME>07:31:39</TIME>
   <VALUE>100</VALUE>
   <UDS>%</UDS>
</DATA_RECORD>
<DATA_RECORD>
   <CODE>103228002</CODE>
   <DESCRIPTION>Saturación de oxígeno</DESCRIPTION>
   <DATE>2020/05/15</DATE>
   <TIME>07:28:18</TIME>
   <VALUE>97</VALUE>
   <UDS>%</UDS>
</DATA_RECORD>
<DATA_RECORD>
   <CODE>103228002</CODE>
   <DESCRIPTION>Saturación de oxígeno</DESCRIPTION>
   <DATE>2020/05/14</DATE>
   <TIME>07:37:34</TIME>
   <VALUE>99</VALUE>
   <UDS>%</UDS>
</DATA_RECORD>
```

Fig. 5. Extract of semantically interoperable EHR.

by fully defining the information model and its constraints. Thus, if a datum is not compliant with the archetype, it is not used in the generation of the secondary use model. Fig. 5 shows an EHR extract related to "Oxygen saturation" ("Saturación de oxígeno" in Spanish) archetype, implemented in Extensible Markup Language (XML).

The transformation rules were applied to these EHR extracts to generate data files in accordance with secondary use models. To this end, different modules were designed and developed for each type of operation identified. The effort is not multiplied for each secondary use model: instead, these operations are adjusted in line with its specific requirements. This allows the generation processes to be reusable and scalable to any secondary use model. Fig. 6 shows an example which selects the "Oxygen saturation" values (identified via SNOMED CT code "103228002") between the starting and finishing dates of the admission episode, and only the maximum and minimum values.

In view of the support shown by the clinical and scientific community [36], the rapid case report form (CRF) proposed by the Severe Acute Respiratory and emerging Infection Consortium (ISARIC-WHO) was chosen as the secondary use model to transform to for technical validation of the methodology [37]. Although Spain has not yet issued a COVID-19 data specification at a national level at the date of writing, this could be generated in the same way with the proposed methodology. The information model designed by ISARIC-WHO for the rapid CRF defines around 200 data elements, 68 of which are OE concerning to 36 concepts. It is structured in three modules: the first for hospital admission data; the second for the first day of admission to the intensive care unit (ICU) and as many times as possible across hospitalization; and the third for the date of patient discharge or death. By virtue of this model's volume of OE concepts and the data-registration criteria it establishes, it is optimal for validating the methodology. Thus, this model was generated from EHRs of 4489 patients hospitalized due to COVID-19 from 25 February 2020 to 10 September 2020. Fig. 7 shows an overview of the methodology implementation process, based on the components described above.

3. Results

The results of this study are the deliverables defined in the different stages of the methodology. Its implementation into the Hospital began on March 15, 2020 and the first EHR-derived extract was generated and validated on April 20, 2020.

3.1. Standard catalog of observable entities in COVID-19

The first result obtained in this study was the specification and standardization of a set of 22 clinical OE and 36 laboratory-related OE of

```
start_date<-VISITS%>%filter(TYPE='ADM')%>%select(STARTDATE)
end_date<-VISITS%>%filter(TYPE='ADM')%>%select(ENDDATE)
sao2_min<-OBSERVATIONS%>%filter(CODE=='103228002')%>%
    filter(MRN==mrn,(DATE>start_date & DATE<end_date))%>%filter(VALUE==min(VALUE))
sao2_max<-OBSERVATIONS%>%filter(CODE=='103228002')%>%
    filter(MRN==mrn,(DATE>start_date & DATE<end_date))%>%filter(VALUE==max(VALUE))
```

Fig. 6. Code in R for generating data related to "Oxygen saturation" concept.

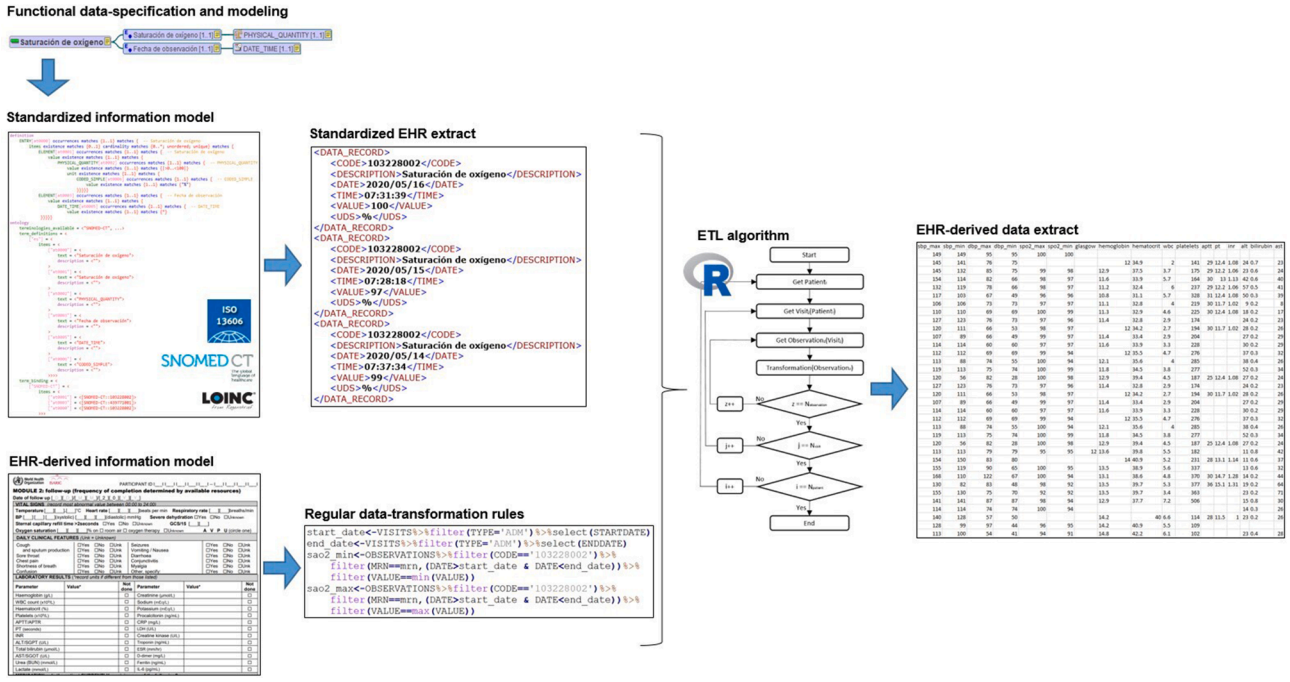

Fig. 7. Overview of the methodology implementation process.

interest in COVID-19 (included in Appendix A). These concepts, in consonance with the ISO 13606 standard and semantically linked to standard terminologies, are implemented in the multiple Hospital healthcare information systems, allowing homogenous data entry via clinical record forms or, transparently, through integration with laboratory equipment. Data are stored in each system's database, following a dual key-value structure: standard concept of the OE and finding reported. This allows the reuse of data, while maintaining their original meaning unaltered.

3.2. Secondary use model generation rules

The second result achieved was the design and development of transformation rules to be applied on EHRs, based on standard archetypes, for obtaining secondary use models. In order to address the generation of the proposed ISARIC-WHO information model, data transformations rules were adapted to the specific criteria, without the need of creating any operations in addition to those identified in Stage 3 of the methodology. An algorithm in R was thus implemented: this selects data for each patient in line with the standard OE concepts, to which it then applies the rules defined for generating the ISARIC-WHO COVID-19 information model.

3.3. ISARIC-WHO COVID-19 data extract

Lastly, the set of OE proposed by ISARIC-WHO for the cohort of 4489 hospitalized patients due to COVID-19 (4286 confirmed by laboratory test and 203 with clinical diagnosis) from 25 February 2020 to 10 September 2020 was obtained. Of a total of 36 OE that define this model, 34 could be generated. As the concepts "Capillary refill time" and "Mid-upper arm circumference" were not identified in Stage 1 of the methodology, it was proposed that they should be included in the information model and, by extension, in the hospital information systems. The proposed methodology allowed expanding the concept model without altering the data model, through the definition of new clinical archetypes. Table 1 shows the volume of data that could be automatically generated from EHRs. Firstly, it shows the total records directly extracted from health information systems, prior to being processed. Secondly, it shows the data after application of the generation algorithm for modules 1 and 2 of the ISARIC-WHO information model (module 3 does not include OE concepts), with the following breakdown: total number of records generated; number of patients to whom these refer; and the percentage with respect to the total cohort covered.

As can be seen, the majority of OE had a patient coverage of over 85%. Some basic vital constants, e.g., blood pressure and oxygen saturation, as well as SARS-COV-2 detection test and common laboratory tests, e.g., hemogram, sodium or potassium, had a high coverage since these measurements are performed daily on most hospitalized COVID-19 patients. Even so, there were concepts, such as the Glasgow Coma Scale score or specific laboratory tests, e.g., IL-6 and lactate, in which the percentage of patients covered was in the region of 10%. This is due to not all patients underwent the complete set of OE included in the model. The fact that these are real world data means that each patient exclusively generated data relating to the observations which professionals found necessary in healthcare activity.

Table 1
ISARIC-WHO OE dataset generated from healthcare data.

	EHR	ISARIC-WHO MODULE 1			ISARIC-WHO MODULE 2		
	Records (N)	**Records (N)**	**Patients (N)**	**Patients (%)**	**Records (N)**	**Patients (N)**	**Patients (%)**
SARS-COV-2	9179	4286	4286	95.48	–	–	–
Height	6781	1060	1060	23.61	–	–	–
Weight	7596	1070	1070	23.84	–	–	–
Temperature	148,184	3926	3926	87.46	42,015	4405	98.13
Heart rate	131,251	3799	3799	84.63	39,849	4342	96.73
Respiratory rate	6456	364	364	8.11	3205	1142	25.44
Systolic blood pressure	107,477	3773	3773	84.05	39,430	4308	95.97
Diastolic blood pressure	107,388	3773	3773	84.05	39,425	4308	95.97
Oxygen saturation	132,486	2873	2873	64.00	36,506	4203	93.63
Glasgow Coma score	1012	478	478	10.65	737	677	15.08
Hemoglobin	37,683	4195	4195	93.45	21,971	4219	93.99
Leukocytes	37,326	4194	4194	93.43	21,965	4218	93.96
Hematocrit	37,318	4194	4194	93.43	21,965	4218	93.96
Platelets	37,322	4195	4195	93.45	21,967	4219	93.99
aPTT	21,978	4044	4044	90.09	13,766	4131	92.02
Prothrombin time	21,992	4044	4044	90.09	13,767	4130	92.00
INR	22,001	4044	4044	90.09	13,769	4130	92.00
ALT/SGPT	35,031	4109	4109	91.53	21,249	4193	93.41
Bilirubin	34,435	3974	3974	88.53	21,061	4192	93.38
AST/SGOT	34,302	3973	3973	88.51	20,964	4163	92.74
Urea	9896	1661	1661	37.00	6564	2363	52.64
Lactate	383	110	110	2.45	259	209	4.66
Creatinine	38,226	4168	4168	92.85	22,415	4208	93.74
Sodium	37,458	4161	4161	92.69	22,338	4207	93.72
Potassium	37,257	4130	4130	92.00	22,229	4204	93.65
Procalcitonin	3621	367	367	8.18	3133	1371	30.54
C reactive protein	29,695	4078	4078	90.84	20,372	4154	92.54
LDH	26,188	3934	3934	87.64	17,542	4104	91.42
Creatine kinase	14,965	1852	1852	41.26	11,538	3573	79.59
Troponin T	5091	751	751	16.73	3804	1714	38.18
ESR	286	14	14	0.31	64	47	1.05
D-dimer	7351	1864	1864	41.52	6238	2861	63.73
Ferritin	7613	615	615	13.70	4381	3160	70.39
IL-6	1046	63	63	1.40	807	626	13.95

4. Discussion

The proposed methodology takes the DCM paradigm as its basis, being initially applied successfully to the creation of an i2b2 data warehouse in the Hospital [27,38]. However, this study broadens its scope given that, for an effective reuse of health data, it is necessary to create a mechanism that offers data to consumers in the format they demand. In comparison with previous studies focus on DCM approach for data extraction from heterogeneous sources [39], the proposed methodology serves not only to extract and standardize the data currently generated, but also to improve the Hospital information systems. Consequently, it is possible to record data with the modeling and standardization requirements needed for transforming them into the information models demanded by the different initiatives dedicated to the collection, integration, and harmonization of COVID-19 data. In this sense, ISARIC has implemented an EDC, based on ISARIC-WHO CRF, for reporting COVID-19 cases to generate monthly clinical data [40]; the 4CE Consortium has designed a common model of aggregated COVID-19 data to perform combined studies [41]; TriNetX has defined an essential set of data elements to build COVID-19 research cohorts from EHRs [42]; the European Health Data Evidence Network (EHDEN) has launched a rapid call to homogenize COVID-19 data in a European network of OMOP repositories [43]; and the National COVID Cohort Collaborative (N3C) has created an open scientific community focused on the analysis of patient-level data from multiple centers [44]. The aim of the methodology proposed in this study is not to replace these initiatives, but to obtain data conforming to the information model designed in each one of them rapidly and efficiently.

In parallel to archetype-based initiatives, such as ISO 13606 standard or OpenEHR specification [19], the Fast Healthcare Interoperability Resources standard (FHIR) of Health Level Seven (HL7) has been applied to model COVID-19 information by different standardization initiatives [45]. This standard offers a rapid mechanism for information exchange between different systems without loss of meaning. To achieve this, FHIR provides a series of common health information resources, which incorporate semantics as an element of the information model itself, defining a generic "Observation" resource for representing and exchanging any observable entity. Nonetheless, for formalizing the concept model of multiple information systems, it is necessary to implement an archetype for each clinical concept, defining its specific components and constraints. This just means that both standards, ISO 13606 and FHIR, can be used in conjunction, applying each of them for its design purpose. In relation to this, the group of experts from the Technical Committee ISO/TC 215 Health informatics is working on the "Guidelines for implementation of HL7/FHIR based on ISO 13940 and ISO 13606" [46].

Therefore, use was made of the ISO 13606 standard, parts 1 (reference model) and 2 (archetype model), because of its stability and adaptability, as well as its adoption as a reference standard by Spain and our Hospital. On the one hand, the reference model was used to model concepts pertaining to the OE to be implemented in healthcare information systems. On the other hand, the archetype model made it possible to formalize the information models and link their components to standard terminologies, which represent their clinical meaning. Adopting the ISO 13606 standard enabled the methodology to be a systematic process, homogenizing the data extracts to be transformed and ensuring the completeness and consistency of data through the full definition of the information model and its constraints. In addition, implementation of clinical archetypes allows these to be published and shared for subsequent use. Thus, reuse of clinical archetypes and the designed ETL process allow the methodology to be extended to other health organizations and applicable to other conditions with minimum

effort. If an organization decides to apply it, the only manual work is required to implement the clinical archetypes in the information systems of the organization (creation and mapping of standard concepts) at Stage 4. In the case of applying the methodology to a different condition, it may be necessary to include new clinical concepts at Stage 1 and 2, as well as to adapt the transformation rules to the specified EHR-derived model at Stage 3. This reproducibility is essential in a country like Spain, which has 17 Regions with transferred health authority, so it could be applicable to each of them to standardize the clinical concept models of their multiple information systems [47].

The fact that this methodology was developed in a scenario of a new disease means that the specification of relevant variables should be expanded in the future: it is preferable to collect useful data at this time rather than wait for a perfect model. DCM allows extending the initial concept model defined without altering the information systems that implement them. Thus, ISO 13606-compliant archetypes were used as a basis for implementing the clinical domain concepts that render the multiple Hospital healthcare information systems conceptually homogeneous. Some applications of the ISO 13606 standard in the methodology will be expanded in next studies: due to Hospital information systems are not prepared for automatically incorporating archetypes, the definition of the clinical concepts was performed manually in each of these systems on the basis of the defined archetypes (terminology binding and metadata). Similarly, a structure implemented in XML and Delimiter-Separated Values (DSV) was chosen for the EHR extracts on which to apply the transformation operations, since it allows data to be processed without loss of meaning. In order to make these extracts completely interoperable, use must be made of a common structure towards which to converge among different organizations. Accordingly, a constraint to be resolved in future studies is to employ the ISO 13606 archetype model for automatic definition of concepts in any healthcare information systems and generation of EHR extracts in line with these, as proposed by previous papers on the topic [48].

Terminology binding of the OE was effected using only two terminologies: SNOMED CT and LOINC. SNOMED CT has been used for clinical OE, and of a total of 22 concepts, only five could not be found in the International Edition, with resort being had to the concept-extension mechanism defined by this terminology. LOINC was used for laboratory OE, and a total of 36 concepts were found in the terminology. The use of only two terminological standards to cover the complete spectrum of OE registered in healthcare information systems differs completely from conventional methodologies based on implementation of specific data collection forms with their own coding, where the same data is recorded in multiple systems in multiple ways [11]. This amounts to a real and initial implementation of something that international studies propose as a line to be pursued in health research based on real world data from multiple sources [49].

In accordance with the archetypes implemented, transformation rules for generating the ISARIC-WHO information model were defined and then validated by the clinical team. These rules were designed with a multipurpose approach, so they can be adapted to generate any EHR-derived model that might require these OE. In this case, it was only necessary to adjust parameters regarding temporality and values of interest in accordance with the specific requirements of the model. These rules process EHR extracts implemented in XML and DSV and then generate EHR-derived data extracts conforming to ISARIC-WHO, which can be directly used by consumers or stored in shared repositories [40]. By way of complementing the above, this study is to be followed by systematic application of these transformation rules to EHR extracts in accordance with ISO 13606 as in previous studies on transformation between information models [50,51].

Lastly, the automatic generation of the ISARIC-WHO COVID-19 CRF had a patient coverage of over 85%. The fact of reusing data from EHRs means that each patient exclusively generates data relating to the observations which professionals found necessary to obtain in healthcare activity. In this line, the EHR2EDC project has developed a seamless and acceptable method for reusing hospital EHR data within clinical trials. Its first objective was to transfer at least 15% of the specified data, and it was possible to achieve up to 37% [52]. Comparing with manual data collection methodologies, reusing health data has made greater data scope achievable, without the need for any additional effort on the organization side. At the same time of this project, a relevant COVID-19 study, based on manual data entry using ISARIC-WHO CRF, was conducted in 208 acute care hospitals of England, Wales and Scotland [36,37]. It recollected adequately information of 20,133 hospitalized patients of domains identified by the proposed methodology as less problematic, such as demographic data, visits, comorbidities, symptoms or treatment. Nevertheless, the results of this study only included one clinical OE, smoking habit, and none laboratory-related OE. This underscores the need to standardize this highly extensive and heterogeneous data domain. Moreover, the cohort of this study is composed of patients admitted with COVID-19 between February 6, 2020 and April 19, 2020. In manual data collection processes, the number of patients included determines the effort and time required by the organization. Our methodology was applied to a cohort of 4489 patients hospitalized from 25 February 2020 to 10 September 2020. This process has no such limitation as once the process of generating the secondary use model from EHRs has been implemented, the number of cases to be included does not imply additional effort or time. That said, EHR data have certain characteristics that differ from those collected manually for a specific purpose [53]. Although the archetypes allow setting a basic control of the data quality, this study will be followed by another into the quality, validity and utility of EHR-derived data in research and other secondary uses.

5. Conclusions

This study has furnished a real and novel solution to the difficulty of rapidly and efficiently obtaining EHR-derived data for secondary use in COVID-19, capable of adapting to changes in data specifications and ensuring acceptable data quality. Thus, a flexible methodology based on DCM paradigm was designed and implemented in a tertiary Hospital of Madrid Region, Spain. This country has 17 Health Services with health-authority transferred, so the methodology could be applicable to each Region, and even to other countries, to homogenize the data-reuse process for COVID-19 and other health conditions. The exposed methodology was divided in four stages. First, a total of 58 OE were identified as an initial set of relevant concepts for COVID-19. These were then modeled and formalized via parts 1 and 2 of the ISO 13606 standard, and semantically linked to standards such as SNOMED CT and LOINC. Selection and transformation rules for generating EHR-derived models were, therefore, designed and implemented. Lastly, the transformation process was validated by generating the information model proposed by ISARIC-WHO for the 4489 COVID-19 cases identified at the hospital up to 10 September 2020. Of the 36 OE included in the ISARIC-WHO model, it was possible to obtain 34 with a coverage, in most instances, of over 85% of patients in the cohort. The conclusion to be drawn from this initial validation is that this methodology allows the effective reuse of EHRs in a real and complex scenario with a greater scope than that yielded by classic manual-record process in ad-hoc EDC and without requiring additional effort or time on the part of the healthcare professionals.

Table A.1
Standardized set of clinical observable entities.

Concept	Data type	Values/ Unit	SNOMED CT
Height	PQ	cm	50373000 |Body height measure (observable entity)|
Weight	PQ	kg	27113001 |Body weight (observable entity)|
Temperature	PQ	°C	386725007 |Body temperature (observable entity)|
Heart rate	PQ	lat/min	364075005 |Heart rate (observable entity)|
Respiratory rate	PQ	resp/min	86290005 |Respiratory rate (observable entity)|
Systolic blood pressure	PQ	mmHg	271649006 |Systolic blood pressure (observable entity)|
Diastolic blood pressure	PQ	mmHg	271650006 |Diastolic blood pressure (observable entity)|
Oxygen saturation	PQ	%	103228002 |Hemoglobin saturation with oxygen (observable entity)|
Oxygen concentration	PQ	%	425608004 |Delivered oxygen concentration (observable entity)|
Oxygen flow rate	PQ	L/min	427081008 |Delivered oxygen flow rate (observable entity)|
Mean blood pressure	PQ	mmHg	6797001 |Mean blood pressure (observable entity)|
Defecation	INTEGER		162098000 |Frequency of defecation (observable entity)|
Urination	INTEGER		364198000 |Frequency of urination (observable entity)|
Vomit	INTEGER		63361000122100 |Frequency of vomits (observable entity)|
Smoking habit	CV	Non-smoker; Ex-smoker; Smoker	266918002 |Tobacco smoking consumption (observable entity)|
Tobacco exposure	INTEGER		782516008 |Number of calculated pack years for cumulative lifetime tobacco exposure (observable entity)|
Date started smoking	DATE		63371000122105 |Date started smoking (observable entity)
Date ceased smoking	DATE		160625004 |Date ceased smoking (observable entity)|
Glasgow Coma score	INTEGER		248241002 |Glasgow coma score (observable entity)|
qSOFA score	INTEGER		63451000122107 |qSOFA score (observable entity)|
SOFA score	INTEGER		63441000122105 |SOFA score (observable entity)|
NEWS score	INTEGER		63441000122102 |NEWS score (observable entity)|

CRediT authorship contribution statement

Miguel Pedrera Jiménez: Conceptualization, Methodology, Project administration, Writing - original draft. **Noelia García Barrio:** Methodology, Software, Writing - original draft. **Jaime Cruz Rojo:** Data curation, Validation, Writing - review & editing. **Ana Isabel Terriza Torres:** Data curation, Validation, Writing - review & editing. **Elena Ana López Jiménez:** Data curation, Validation, Writing - review & editing. **Fernando Calvo Boyero:** Data curation, Validation, Writing - review & editing. **María Jesús Jiménez Cerezo:** Data curation, Validation, Writing - review & editing. **Alvar Javier Blanco Martínez:** Resources, Writing - review & editing. **Gustavo Roig Domínguez:** Resources, Writing - review & editing. **Juan Luis Cruz Bermúdez:** Supervision, Writing - review & editing. **José Luis Bernal Sobrino:** Supervision, Writing - review & editing. **Pablo Serrano Balazote:** Conceptualization, Supervision, Writing - review & editing. **Adolfo Muñoz Carrero:** Supervision, Writing - review & editing.

Table A.2
Standardized set of laboratory-related observable entities.

Concept	Data type	Values/ Unit	LOINC
SARS-COV-2	CV	Positive; Negative; Equivocal	94315-9 |SARS coronavirus 2 E gene [Presence] in Unspecified specimen by NAA with probe detection
Hemoglobin	PQ	g/dL	718-7 Hemoglobin [Mass/volume] in Blood
Leukocytes	PQ	x1000/µL	6690-2 Leukocytes [#/volume] in Blood by Automated count
Lymphocytes	PQ	x1000/µL	731-0 Lymphocytes [#/volume] in Blood by Automated count
Platelets	PQ	x1000/µL	777-3 Platelets [#/volume] in Blood by Automated count
Neutrophils	PQ	x1000/µL	751-8 Neutrophils [#/volume] in Blood by Automated count
Eosinophils	PQ	x1000/µL	711-2 Eosinophils [#/volume] in Blood by Automated count
Basophils	PQ	x1000/µL	704-7 Basophils [#/volume] in Blood by Automated count
Hematocrit	PQ	%	4544-3 Hematocrit [Volume Fraction] of Blood by Automated count
aPTT	PQ	Sec	3173-2 aPTT in Blood by Coagulation assay
Prothrombin time	PQ	Sec	5902-2 Prothrombin time (PT)
INR	PQ	{INR}	6301-6 INR in Platelet poor plasma by Coagulation assay
Albumin	PQ	g/dL	1751-7 Albumin [Mass/volume] in Serum or Plasma
ALT/SGPT	PQ	U/L	1742-6 Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma
Bilirubin	PQ	mg/dL	1975-2 Bilirubin.total [Mass/volume] in Serum or Plasma
AST/SGOT	PQ	U/L	1920-8 Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma
Urea	PQ	mg/dL	3091-6 Urea [Mass/volume] in Serum or Plasma
Lactate	PQ	mmol/L	2524-7 Lactate [Moles/volume] in Serum or Plasma
Creatinine	PQ	mg/dL	2160-0 Creatinine [Mass/volume] in Serum or Plasma
Sodium	PQ	mEq/L	2951-2 Sodium [Moles/volume] in Serum or Plasma
Potassium	PQ	mEq/L	2823-3 Potassium [Moles/volume] in Serum or Plasma
Procalcitonin	PQ	ng/mL	33959-8 |Procalcitonin [Mass/volume] in Serum or Plasma
C reactive protein	PQ	mg/dL	1988-5 C reactive protein [Mass/volume] in Serum or Plasma
LDH	PQ	U/L	2532-0 Lactate dehydrogenase [Enzymatic activity/volume] in Serum or Plasma
Creatine kinase	PQ	U/L	2157-6 Creatine kinase [Enzymatic activity/volume] in Serum or Plasma
Troponin T	PQ	ng/L	67151-1 Troponin T.cardiac [Mass/volume] in Serum or Plasma by High sensitivity method
ESR	PQ	mm/h	30341-2 Erythrocyte sedimentation rate
Fibrinogen	PQ	mg/dL	3255-7 Fibrinogen [Mass/volume] in Platelet poor plasma by Coagulation assay
D-dimer	PQ	ng/mL	48067-3 Fibrin D-dimer FEU [Mass/volume] in Platelet poor plasma by Immunoassay
Triglyceride	PQ	mg/dL	2571-8 Triglyceride [Mass/volume] in Serum or Plasma
Ferritin	PQ	ng/mL	2276-4 Ferritin [Mass/volume] in Serum or Plasma
IL-6	PQ	pg/mL	26881-3 Interleukin 6 [Mass/volume] in Serum or Plasma
pO2	PQ	mmHg	2703-7 Oxygen [Partial pressure] in Arterial blood
pCO2	PQ	mmHg	

(*continued on next page*)

Table A.2 (*continued*)

Concept	Data type	Values/ Unit	LOINC
			2019-8 Carbon dioxide [Partial pressure] in Arterial blood
FiO2	PQ	%	3150-0 Inhaled oxygen concentration
SaO2	PQ	%	2708-6 Oxygen saturation in Arterial blood

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Hospital 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, "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 Department, Instituto de Salud Carlos III (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 Carlos III Health Institute 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 Mercedes Alfaro, Arturo Romero, Jorge Rangil, María Jesús López de Cuellar, Luis Lapuente, Ana Delgado, Rosalía Fernández and the SNOMED CT National Reference Center for Spain for the support in the standardization and creation of new concepts. We would also like to thank María Elena Hernando (Bioengineering and Telemedicine Centre GBT-UPM) for the support in the revision of the manuscript.

Appendix A. Standardized set of observable entities relating to COVID-19

See Tables A.1 and A.2.

References

[1] N. Zhu, D. Zhang, W. Wang, et al., A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733, https://doi.org/10.1056/NEJMoa2001017.
[2] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, Lancet 395 (2020) 689–697, https://doi.org/10.1016/S0140-6736(20)30260-9.
[3] Situation Report of WHO (COVID-19). https://www.who.int/docs/default-source/coronaviruse/weekly-updates/wou-9-september-2020-cleared-14092020.pdf?sfvrsn=68120013_2. Accessed December 14, 2020.
[4] Hospital Universitario 12 de Octubre. https://www.comunidad.madrid/hospital/12octubre/. Accessed December 14, 2020.
[5] Situation Report of Health Ministry of Spain (COVID-19). https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Actualizacion_204_COVID-19.pdf. Accessed December 14, 2020.
[6] K. Häyrinen, K. Saranto, P. Nykänen, Definition, structure, content, use and impacts of electronic health records: a review of the research literature, Int. J. Med. Inform. 77 (2008) 291–304, https://doi.org/10.1016/j.ijmedinf.2007.09.001.
[7] J.J. Reeves, H.M. Hollandsworth, F.J. Torriani, et al., Rapid response to COVID-19: health informatics support for outbreak management in an academic health system, J. Am. Med. Inform. Assoc. 27 (6) (2020) 853–859, https://doi.org/10.1093/jamia/ocaa037.
[8] A. Atreja, S.M. Gordon, D.A. Pollock, et al., Opportunities and challenges in utilizing electronic health records for infection surveillance, prevention, and control, Am. J. Infect. Control 36 (2008) 37–46, https://doi.org/10.1016/j.ajic.2008.01.002.
[9] M.O. Kim, E. Coiera, F. Magrabi, Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review, J. Am. Med. Inform. Assoc. 24 (2017) 246–250, https://doi.org/10.1093/jamia/ocw154.
[10] C. Safran, M. Bloomrosen, E. Hammond, et al., Toward a national framework for the secondary use of health, J. Am. Med. Inform. Assoc. 14 (2007) 1–9, https://doi.org/10.1197/jamia.M2273.Introduction.
[11] R.L. Richesson, J. Krischer, Data standards in clinical research: gaps, overlaps, challenges and future directions [published correction appears in J Am Med Inform Assoc. 2008 Mar-Apr;15(2):265], J. Am. Med. Inform. Assoc. 14 (6) (2007) 687–696, https://doi.org/10.1197/jamia.M2470.
[12] H. Sun, K. Depraetere, J. De Roo, et al., Semantic processing of EHR data for clinical research, J. Biomed. Inform. 58 (2015) 247–259, https://doi.org/10.1016/j.jbi.2015.10.009.
[13] S.N. Murphy, G. Weber, M. Mendis, et al., Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. 17 (2) (2010) 124–130, https://doi.org/10.1136/jamia.2009.000893.
[14] G. Hripcsak, J.D. Duke, N.H. Shah, et al., Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, Stud Health Technol Inform. 216 (2015) 574–578.
[15] Beale T. Archetypes. Constraint-based domain models for future-proof information systems, 2001. http://www.openehr.org/publications/archetypes/archetypes_beale_web_2000.pdf. Accessed December 14, 2020.
[16] W. Goossen, A. Goossen-Baremans, M. van der Zel, Detailed clinical models: a review, Healthc. Inform. Res. 16 (2010) 201, https://doi.org/10.4258/hir.2010.16.4.201.
[17] M.R. Santos, M.P. Bax, D. Kalra, Building a logical EHR architecture based on ISO 13606 standard and semantic web technologies, Stud. Health Technol. Inform. 160 (2010) 161–165, https://doi.org/10.3233/978-1-60750-588-4-161.
[18] B. Christensen, G. Ellingsen, Evaluating model-driven development for large-scale EHRs through the openEHR approach, Int. J. Med. Inform. 89 (2016) 43–54, https://doi.org/10.1016/j.ijmedinf.2016.02.004.
[19] M. Li, H. Leslie, B. Qi, et al., Development of an openEHR template for COVID-19 based on clinical guidelines, J Med Internet Res 22 (2020) e20239, https://doi.org/10.2196/20239.
[20] K. Donnelly, SNOMED-CT: The advanced terminology and coding system for eHealth, Stud. Health Technol. Inform. 121 (2006) 279–290.
[21] C.J. McDonald, S.M. Huff, J.G. Suico, et al., LOINC, a universal standard for identifying laboratory observations: A 5-year update, Clin. Chem. 49 (2003) 624–633, https://doi.org/10.1373/49.4.624.
[22] SNOMED CT resources for COVID-19. https://www.snomed.org/news-and-events/articles/march-2020-interim-snomedct-release-COVID-19. Accessed December 14, 2020.
[23] LOINC resources for COVID-19. https://loinc.org/sars-coronavirus-2/. Accessed December 14, 2020.
[24] R. Lozano-Rubí, A. Muñoz Carrero, P. Serrano Balazote, et al., OntoCR: a CEN/ISO-13606 clinical repository based on ontologies, J. Biomed. Inform. 60 (2016) 224–233, https://doi.org/10.1016/j.jbi.2016.02.007.
[25] A. Muñoz, R. Somolinos, M. Pascual, et al., Proof-of-concept design and development of an EN13606-based electronic health care record service, J. Am. Med. Inform. Assoc. 14 (1) (2007) 118–129, https://doi.org/10.1197/jamia.M2058.
[26] R. Sánchez-de-Madariaga, A. Muñoz, R. Lozano-Rubí, et al., Examining database persistence of ISO/EN 13606 standardized electronic health record extracts: relational vs. NoSQL approaches. BMC Med Inform Decis. Mak. 17(1) (2017) 123. Published 2017 Aug 18. doi:10.1186/s12911-017-0515-4.
[27] M. Pedrera, P. Serrano, A. Terriza, et al., Defining a standardized information model for multi-source representation of breast cancer data, Stud. Health Technol. Inform. 270 (2020) 1243–1244, https://doi.org/10.3233/SHTI200383.
[28] R.L. Gardner, E. Cooper, J. Haskell, et al., Physician stress and burnout: the impact of health information technology, J. Am. Med. Inform. Assoc. 26 (2) (2019) 106–114, https://doi.org/10.1093/jamia/ocy145.
[29] J.A. Maldonado, D. Moner, D. Boscá, J.T. Fernández-Breis, C. Angulo, M. Robles, LinkEHR-Ed: a multi-reference model archetype editor based on formal semantics, Int. J. Med. Inform. 78 (8) (2009) 559–570, https://doi.org/10.1016/j.ijmedinf.2009.03.006.
[30] Health Ministry of Spain: Clinical modeling resources, reference ISO 13606 archetypes. https://www.mscbs.gob.es/profesionales/hcdsns/areaRecursosSem/Rec_mod_clinico_arquetipos.htm. Accessed December 14, 2020.
[31] S. Sun, T. Austin, D. Kalra, A data types profile suitable for use with ISO EN 13606, J. Med. Syst. 36 (6) (2012 Dec) 3621–3635, https://doi.org/10.1007/s10916-012-9837-z. Epub 2012 Mar 8 PMID: 22399066.
[32] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (1) (2013) 144–151, https://doi.org/10.1136/amiajnl-2011-000681.
[33] Health Ministry of Spain: Minimum data set for clinical reports. https://www.boe.es/eli/es/rd/2010/09/03/1093. Accessed December 14, 2020.
[34] Health Ministry of Spain: SNOMED CT resources for COVID-19. https://www.mscbs.gob.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/SNOMED_CT_COVID-19.htm. Accessed December 14, 2020.
[35] R Foundation for Statistical Computing. https://www.r-project.org/about.html. Accessed December 14, 2020.
[36] A.B. Docherty, E.M. Harrison, C.A. Green, et al., Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study, BMJ (2020;369:m1985.), https://doi.org/10.1136/bmj.m1985. Published 2020 May 22.

[37] ISARIC-WHO CRF for COVID-19. https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf/. Accessed December 14, 2020.
[38] L. González, D. Pérez-Rey, E. Alonso, et al., Building an I2B2-based population repository for clinical research, Stud. Health Technol. Inform. 270 (2020) 78–82, https://doi.org/10.3233/SHTI200126.
[39] S. Lim Choi Keung, L. Zhao, J. Rossiter, et al., Detailed clinical modelling approach to data extraction from heterogeneous data sources for clinical research, AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci. 2014 (2014) 55–59, https://doi.org/10.1016/j.ic.2014.12.007.
[40] ISARIC-WHO COVID-19 Data Management & Hosting. https://isaric.org/research/covid-19-clinical-research-resources/covid-19-data-management-hosting/. Accessed December 14, 2020.
[41] G.A. Brat, G.M. Weber, N. Gehlenborg, et al., International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium, NPJ Digit Med. 3 (2020) 109. Published 2020 Aug 19. doi:10.1038/s41746-020-00308-0.
[42] TriNetX: COVID-19 Clinical Data. https://trinetx.com/coronavirus/. Accessed December 14, 2020.
[43] EHDEN: COVID19 Rapid Collaboration Call. https://www.ehden.eu/open-calls/04-2020-covid19-data-partner-call/. Accessed December 14, 2020.
[44] H. Melissa, C. Christopher, G. Kenneth, The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment [published online ahead of print, 2020 Aug 17], J. Am. Med. Inform. Assoc. 2020;ocaa196. doi:10.1093/jamia/ocaa196.
[45] Logica COVID-19 FHIR Profile Library. https://covid-19-ig.logicahealth.org/. Accessed December 14, 2020.
[46] ISO/AWI TR 24305 “Guidelines for implementation of HL7 / FHIR based on ISO 13940 and ISO 13606”. https://www.iso.org/standard/78390.html. Accessed December 14, 2020.
[47] Health Ministry of Spain: Electronic Health Records of the National Health System (HCDSNS). https://www.mscbs.gob.es/profesionales/hcdsns/home.htm. Accessed December 14, 2020.
[48] G. Duftschmid, T. Wrba, C. Rinner, Extraction of standardized archetyped data from Electronic Health Record systems based on the Entity-Attribute-Value Model, Int. J. Med. Inform. 79 (8) (2010) 585–597, https://doi.org/10.1016/j.ijmedinf.2010.04.007.
[49] B. Claerhout, D. Kalra, C. Mueller, et al., Federated electronic health records research technology to support clinical trial protocol optimization: evidence from EHR4CR and the InSite platform, J. Biomed. Inform. 90 (2019) 103090, https://doi.org/10.1016/j.jbi.2018.12.004.
[50] S. Mate, F. Köpcke, D. Toddenroth, et al., Ontology-based data integration between clinical and research systems [published correction appears in PLoS One. 10(3) 2015 e0122172]. PLoS One. 2015;10(1):e0116656. Published 2015 Jan 14. doi:10.1371/journal.pone.0116656.
[51] J.A. Maldonado, M. Marcos, J.T. Fernández-Breis, V.M. Giménez-Solano, M.D.C. Legaz-García, B. Martínez-Salvador, CLIN-IK-LINKS: a platform for the design and execution of clinical data transformation and reasoning workflows [published online ahead of print, 2020 Jun 25], Comput. Methods Programs Biomed. 197 (2020) 105616. doi:10.1016/j.cmpb.2020.105616.
[52] EHR2EDC project. https://www.i-hd.eu/index.cfm/r-d-and-collaborative-projects/research-projects/ehr2edc/. Accessed December 14, 2020.
[53] N.G. Weiskopf, G. Hripcsak, S. Swaminathan, C. Weng, Defining and measuring completeness of electronic health records for secondary use, J. Biomed. Inform. 46 (5) (2013) 830–836, https://doi.org/10.1016/j.jbi.2013.06.010.