Received: 15 December 2015 | Revised: 7 July 2016 | Accepted: 28 July 2016

DOI 10.1002/lrh2.10014

Learning Health Systems

TECHNICAL REPORT

Cross border semantic interoperability for learning health systems: The EHR4CR semantic resources and services

Christel Daniel[1,2] | David Ouagne[1] | Eric Sadou[1,2] | Nicolas Paris[2] | Sajjad Hussain[1] | Marie-Christine Jaulent[1] | Dipak Kalra[3]

[1] Sorbonne Universités, UPMC Univ Paris 06, INSERM UMR_S 1142, LIMICS, F-75006, Paris, France

[2] AP-HP, Paris, France

[3] EUROREC Institute, Sint-Martens-Latem, Belgium

Correspondence
Christel Daniel, DSI-WIND-AP-HP 5, rue Santerre 75012 Paris, France.
Email: christel.daniel@aphp.fr

FUNDING
This work was supported by the Innovative Medicines Initiative Joint Undertaking, under grant agreement no 115189, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013).

This article was published online on 21 October 2016. Subsequently, an author's name was found to be incorrect, and the correction was published on 31 October 2016.

Abstract

With the development of platforms enabling the integration and use of phenome, genome, and exposome data in the context of international research, data management challenges are increasing, and scalable solutions for cross border and cross domain semantic interoperability need to be developed. Reusing routinely collected clinical data, especially, requires computable portable phenotype algorithms running across different electronic health record (EHR) products and healthcare systems. We propose a framework for describing and comparing mediation platforms enabling cross border phenotype identification within federated EHRs. This framework was used to describe the experience gained during the EHR4CR project and the evaluation of the platform developed for accessing semantically equivalent data elements across 11 European participating EHR systems from 5 countries. Developers of semantic interoperability platforms are beginning to address a core set of requirements in order to reach the goal of developing cross border semantic integration of data.

KEYWORDS

biomedical research, data integration and standardization, electronic health records, interoperability, knowledge representation, terminology as topic

1 | INTRODUCTION

Within a learning health system, every patient interaction with the healthcare system provides an opportunity to generate data that can be used to create new evidence and knowledge, which in turn can be used to improve clinical practice at the point of care.[1] The phenotype of an individual results from the interplay between the genome and the external/environmental elements to which it is exposed. Biomedical research interest in environmental exposures, as well as "omics" data as determinants of physiopathological processes, is raising as such data increasingly become available.[2] Therefore, integration of genome, phenome, and exposome data is gaining an important supporting role in different areas such as clinical research, patient safety, comparative effectiveness, and public health monitoring.

In clinical research, specific topics of interest include eligibility determination—finding clinical trials for which a patient is eligible,[3] or identifying a cohort that is eligible for a clinical trial in order to provide clinical trial planners with a better understanding of the eligible cohorts[4] or to support targeted patient recruitment.[5,6] Example systems include ASPIRE,[7] caMatch,[8] EON,[9] AIDS2,[10] PROforma,[11] Asbru,[12] GLIF,[13] SAGE,[14] GUIDE,[15] TRANSFoRm,[16] and EHR4CR.[17] Another topic of interest is "single-source data entry" at the point of clinical care[18,19]. In epidemiology and public health, relevant solutions support public health strategy by increasing the ability to provide actionable insights at the point of care. These solutions include systems enabling early identification of the combined effects of environment, lifestyle, and genetics on public health (descriptive analysis), advanced simulation methods to study causal mechanisms, to improve current risk stratification methodologies or to improve forecasts of spatial and temporal development of ill-health and disease (predictive analysis) and systems turning huge amount of data into actionable information to authorities for planning public health activities (prescriptive analysis).

Data management challenges are increasing with improvements in techniques and high-throughput pipelines in the 3 domains—genome (high-throughput "omics" pipelines), exposome (high-throughput pipelines of patient-centric or non-patient centric exposome data), and phenome (increasing availability of imaging and multi-channel physiologic datasets). In this context, the aim of biomedical researchers and

funding bodies, such as the US National Institutes of Health,[a] the UK National Health Service,[b] or EU commission,[c] is to promote advances in technologies and methods for data management and analytics in order to better acquire, manage, share, model, process, and exploit big data collections and extract knowledge from them. The US National Institutes of Health Big Data to Knowledge (BD2K) Initiative is now fostering several activities to support the full exploitation of the large collection of data available in health care and funding big data excellence centers, which will be in charge of translating ideas into actionable tools and case studies.[20]

Rather than improving existing isolated systems, initiatives now focus on how to merge those large data collections to obtain even larger data sets, holding the promise to improve our understanding of diseases at a considerable higher pace than today.[21]

Large data integration efforts based on the data warehousing approach, which extracts, transforms, and loads data from heterogeneous sources into a single view schema, have started to improve outcome research, such as PCORI in the United States. Moreover, collaborative efforts are designed not only to share data but also to disseminate large-scale analytics to bring out the value of observational health data, such as the Observational Health Data Sciences and Informatics Research Network (http://www.ohdsi.org/).

An interesting area that may successfully exploit big data technologies is electronic phenotyping. Currently, phenotype algorithms are most commonly represented as non-computable descriptive documents and knowledge artifacts that detail the protocols for querying diagnoses, symptoms, procedures, medications, and/or text-driven medical concepts and are primarily meant for human comprehension. Current initiatives aim at developing complex computerized queries to clinical data repositories that allow ascertaining a clinical condition or characteristic (phenotype).[22] The eMERGE network is especially designed to combine -omics biorepositories with electronic medical record data for supporting genetic research (https://emerge.mc.vanderbilt.edu/).[23–25] Finally, large data repositories involving patients, such as "Patients Like Me" (http://www.patientslikeme.com/), can also be used to foster research.

In Europe, two international projects focus on cross-border clinical data integration for supporting clinical research. TRANSFoRm (http://www.transformproject.eu/) is an EU FP7 project that seeks to develop an infrastructure for the Learning Health System in European primary care. The project is based on 3 clinical use cases, a genotype-phenotype study in diabetes, a randomized controlled trial with gastroesophageal reflux disease, and a diagnostic decision support system for chest pain, abdominal pain, and shortness of breath.[16] The EHR4CR (Electronic Health Records for Clinical Research (http://www.ehr4cr.eu/) is an IMI (Innovative Medicines Initiative) project funded jointly by the European Commission and by the European Federation of Pharmaceutical Industries and Associations.[17] The aim of the project is to reduce the cost of conducting clinical trials, through better leveraging routinely collected clinical electronic health record (EHR) data at key points in the trial design and execution life-cycle. EHR4CR implementations have been installed at 11 pilot hospital sites within 5 European countries (France, Germany, Poland, Switzerland, and United Kingdom). These hospital EHRs collectively contain data from over 7 000 000 patients. The EHR4CR platform is a loosely coupled service platform, which orchestrates independent services addressing semantic interoperability, data protection, privacy, security, and end-user platform services to ease and speed the conduct of clinical trials, in particular during the phases of protocol feasibility study (PFS), patient identification and recruitment services (PRS), and clinical trial execution (CTE). Both TRANSFoRm and EHR4CR projects are based on the mediation approach relying on mappings between a mediation model and the model of original sources, and transformation of queries into specialized queries to match the models of the original databases. Both projects adopted the "Local As View" approach specifying mappings from entities in the original sources to the mediation model as opposed to the "Global As View" approach where such mappings are specified from entities in the mediation model to entities in the original sources.

Our hypothesis is that semantic mediation is achievable internationally within federated database systems by implementing (i) a consistent integrative semantic abstraction on top of existing application proprietary models and (ii) a semantically enabled query engine. The mediation model provides a homogeneous view of the clinical data available in disparate systems so that data users can access these data using a library of standard queries that have been written based on the mediation model.

Our goal is twofold, first to define a set of desiderata for developing a mediation platform for computing phenotype algorithms within a federated database system and second to use this evaluation framework to describe the strengths and limitations of the EHR4CR platform.

2 | METHODS

Our approach consisted first of extending the desiderata for computable representations of EHRs-driven phenotype algorithms proposed by Mo et al[22] in order to propose a conceptual framework for comparing mediation approaches and semantic interoperability solutions developed by platforms supporting cross border research. Second, we instantiated the conceptual framework in the context of the EHR4CR project in order to evaluate how far the development of the mediation model and the standardization efforts met the expected requirements of the project.

The conceptual framework of computable representations of phenotype algorithms consists of a set of requirements related to the 3 main components of a mediation platform: (i) query language and model, (ii) patient data model (mediation model), and (iii) standardization pipeline for data providers.

2.1 | The need for a high quality query language and model

There is a need to manage eligibility criteria in order to accelerate the development of new clinical research protocols and related clinical

[a]https://datascience.nih.gov/bd2k

[b]http://www.nhs.uk/NHSEngland/thenhs/records/healthrecords/Pages/caredata.aspx

[c]https://ec.europa.eu/digital-agenda/en/big-data

research documents (eg, case report forms, data collection forms, and training materials). Related efforts include EligWriter[26] and Designa-Trial[27] that supported the re-use of eligibility criteria during clinical trial protocol authoring, as well as ERGO[28] and ASPIRE[7] that supported the annotation of eligibility criteria. The definition of computable phenotype algorithms requires interoperability with patient data. The knowledge representation requirements for eligibility criteria in this context are more stringent, including highly expressive language(s) to achieve executable eligibility rules, a patient information model, and an appropriate clinical terminology to facilitate mapping from eligibility concepts to patient data.

2.2 | The need for a high quality mediation model (patient data model)

Because each data source is not designed with a primary focus of cross-domain integration, initiatives for integrating clinical data have been often limited to non-scalable, system (or vendor)-specific efforts.[6,18]

In an expanding research landscape, cooperation infrastructures are now being built to allow research projects to reuse patient data from federated systems from many different sites in different countries and therefore in a multilingual settings. Non-standard, and often conflicting, vendor approaches to representing data pose challenges to infrastructure developers, who must build solutions to work with clinical data across multiple formats.

Systems developed during the last decade in order to compute eligibility criteria—including GUIDE, GLIF3, SAGE, ERGO, and CRFQ—largely adopted some form of Virtual Medical Records (VMR) based on the HL7 Reference Information Model (RIM),[29] which provides an abstraction layer on top of a real EHR.

A controlled mediation model is required to support federated access to heterogeneous data sources. Mediation models must be based on the adoption and integration of multiple standards, themselves being aligned to be consistent, coherent, and cross-compatible.[25,30] Although there is no consensus in the medical informatics community regarding a standard patient information model, HL7 FHIR specifications are gaining interest and show promise to mitigate the classic site-specific data mapping problem. In addition, as part of the Clinical Information Modeling Initiative (CIMI) launched in 2011, an international consortium of representing national bodies, Standards Development Organizations, healthcare organizations, and vendors are building collaboratively a process and tools for constructing a single curated collection of shared implementable clinical information models that are free for use at no cost.[d]

2.3 | The need for an efficient standardization pipeline within participants data providers

Beyond the creation and continuous extension of the standard-based mediation model, the process of harmonizing heterogeneous data sources, called "data standardization" in this paper, relies also on the capability of different actors in hospital sites to align the local structures and content of their EHR systems or Clinical Data Repositories to the mediation model. Few EHR systems or Clinical Data Repositories in hospitals implement standard reference models such as HL7 RIM, EN ISO 13606, or openEHR. Most of them rely on proprietary models. Furthermore, although the need for controlled vocabularies in EHR systems is widely recognized, system developers have often dealt with this need by creating ad hoc sets of controlled terms for use in their applications so that information in one system cannot be recognized and used by other systems. Although some standardizations of codes are now occurring, data are not consistent vendor to vendor, or even organization to organization for the same vendor.

Therefore, mapping local models and/or controlled vocabularies is a challenging and time-consuming task for terminologists in participant hospitals.

Efficient supportive mapping tools are required to enable terminologists to develop and maintain structural and semantic mapping between the proprietary models and the mediation model.

3 | RESULTS

3.1 | Evaluation framework of mediation platforms enabling phenotype identification with federated databases

3.2 | EHR4CR use case

Table 1 also provides a qualitative evaluation the strengths and limitations of the EHR4CR platform in the implementation of phenotype algorithms and its capacity to support the different actors in accomplishing their tasks during the data standardization process at both setup and execution phases of the EHR4CR use cases.

3.2.1 | Query model and language

In this section, we describe the characteristics of the *EHR4CR Eligibility Criteria Model (EC Model) and ECLECTIC language* regarding the 8 requirements stated in the "A-Query model and language" section of the conceptual framework.

- Req A.1: Implement set operations and relational algebra for modeling phenotype algorithms, represent phenotype criteria with structured rules

The EHR4CR Eligibility Criteria Model (EC Model) is an extensible query model representing eligibility criteria in the UML language to meet the expressivity needs of computationally viable eligibility criteria. An ad hoc language ECLECTIC (Eligibility Criteria Language for European Clinical Trial Investigation and Construction) has been developed in order to ensure that it can express only queries that the object model can represent. The UML class diagram and language grammar are 2 alternative representations of the same logical model. The resultant object model, although hidden away from the user's eyes, lies at the heart of the query engine and is key for model transformation and query serialization in different forms. Eligibility criteria

[d]http://www.opencimi.org/

TABLE 1 List of 23 requirements defined for the 3 main components involved in a mediation platform enabling phenotype identification in federated databases: (i) query language and model, (ii) patient data model (mediation model), and (iii) standardization pipeline for data providers

		Desiderata proposed by Mo et al[22]	EHR4CR use case
	A-The need of high quality query language ("semantic discovery")		
Req A.1	Implement set operations and relational algebra for modeling phenotype algorithms, represent phenotype criteria with structured rules	Mo 2015; Req.4 and 5	++
Req A.2	Support both human readable and computable representations of phenotype algorithms	Mo 2015; Req.3	++
Req A.3	Support defining temporal relations between events	Mo 2015; Req.6	++
Req A.4	Provide representations for text searching and capture the coding output of natural language processing (NLP)	Mo 2015; Req.8	
Req A.5	Query language shall be generic and standard based		
Req A.6	Query builder shall be intuitive		++
Req A.7	Provide interfaces for external software algorithms	Mo 2015; Req.9	
Req A.8	Maintain backward compatibility	Mo 2015; Req.10	
	B-The need of high quality patient data model ("semantic mediation")		
Req B.1	The mediation model shall be based on standard domain knowledge and reference models provided by standard development organizations that are and will be used by EHR vendors, clinicians, and government mandates (eg, Meaningful Use Stage 3 in US).		+++
Req B.2	The mediation model shall use standard terminologies, ontologies, and value sets that are multilingual and internationally used	Mo 2015; Req.7	+++
Req B.3	Support customization for the variability and availability of EHR data among sites. Possible use of internally defined extensions of existing standard terminologies (in order to add any missing concept or any missing description in any specific language)	Mo 2015; Req.2	++
Req B.4	The mediation model shall use mappings between reference terminologies (eg, SNOMED-MedDRA, and SNOMED CT-NCI Thesaurus) in order to allow end users to access semantically equivalent content through different terminologies		+
Req B.5	The mediation model shall be expressive enough to represent (i) multimodal (sign, symptoms, diseases, outcomes, procedures, care plans, etc, as well as images, signals, etc) and multi-scale clinical data including molecular findings such as genomics information; (ii) specimen related information, family related information, etc; (iii) multiple granularities, multiple consistent views, context representation		
Req B.6	The mediation model shall be scoped to the needs of the users of the research network in the context of dedicated use cases but scalable and sustainable (designed to be rapidly and efficiently scoped to cover any new requirement, extensible in terms of structure and content)		++
Req B.7	The mediation model shall be represented using standard formal languages allowing semantic reasoning (eg, semantic web languages) in order to recognize redundancy or inconsistency		
Req B.8	A robust version management process shall be provided for any type of semantic resource of the mediation model		++
Req B.9	A dedicated tool is required for supporting the authors of the mediation model to efficiently create/update the semantic resources of the model. The editor shall support a collaborative editing process. The creation and update process shall be user-friendly and adapted to medical experts (through user interface, but also through import of simple csv files used to capture medical knowledge in a format that is understandable for medical experts). The editor shall allow the authors to create new semantic resources from standard terminologies (eg, SNOMED CT, LOINC, ATC, and ICD-O) or value sets. The standard resources are imported from the official terminology providers and up-to-date.		++
Req B.10	The semantic resources shall be accessible to any application through standardized semantic services based on new web technologies, such as Representational State Transfer (REST)-based APIs/web services, recently been adopted by HL7.		+++
	C-The need of efficient standardization pipeline within data providers ("data standardization")		
Req C.1	Automatic mapping algorithms supporting terminologists in identifying corresponding concepts in the mediation model on one side and local models on the other side. These algorithms shall (i) use the descriptions and synonyms of the concepts; (ii) address multi lingual issues; (iii) use existing mappings between reference terminologies (eg, when local sources are mapped to a standard terminology that is not used in the mediation model (eg, NCI Thesaurus), using the mapping between SNOMED CT and NCI Thesaurus to propose automatic mappings between local concepts and SNOMED CT concepts in the mediation model)	Mo 2015; Req.1	
Req C.2	Natural language processing for semantic annotation of text	Mo 2015; Req.8	
Req C.3	Formal representation of mappings and version management		
Req C.4	Use case driven support for prioritizing the mapping effort. The terminologist needs to know within the list of the data elements of the mediation model that are not yet mapped to local data elements, the ones that need to be mapped in priority according to different criteria (eg, data elements that are the most frequently used in distributed queries and data elements corresponding to a specific phenotype algorithm)		

(*Continues*)

TABLE 1 (Continued)

		Desiderata proposed by Mo et al[22]	EHR4CR use case
Req C.5	Mappings shall be accessible to any application through standardized semantic services based on new web technologies, such as Representational State Transfer (REST)-based APIs/web services, recently been adopted by HL7		+

are expressed as ECLECTIC queries corresponding to a set of rules defining a phenotype. In ECLECTIC queries, clinical events are embedded within predicates evaluated as true or false for each patient. The predicate may detect the existence of some attribute such as a diagnosis or compare a value (numerical or categorical) to a reference range. Each rule must specify whether the first-occurring or last-occurring (most recent) reading is to be used.[31]

- Req A.2: Support both human readable and computable representations of phenotype algorithms

ECLECTIC is also a human-readable serialization of the object hierarchy, which allows us to reason about the model and perform validation prior to implementation.

- Req A.3: Support defining temporal relations between events

ECLECTIC rules may optionally have a temporal constraint requiring that a clinical event must have occurred before or after some temporal anchor.

- Req A.4: Provide representations for text searching and natural language processing

None

- Req A.5: Query language shall be generic and standard based

ECLECTIC is an ad hoc query language.

- Req A.6: Query builder shall be intuitive

Using the EHR4CR query builder (see Figure 1), a study manager can drag and drop data elements stored in the mediation model (marked as "1" in Figure 1) and logical and temporal operators (marked as "2" in Figure 1) in order to populate query-templates designed for representing formally the eligibility criteria of the clinical trial (marked as "3" in Figure 1).

- Req A.7: Provide interfaces for external software algorithms: none
- Req A.8: Maintain backward compatibility: not addressed

3.3 | Mediation model

Our approach is based on the realistic assumption that there will remain a co-existence of several standard semantic artifacts—namely information models (eg, EN ISO 13606 information model and archetypes, openEHR, HL7 RIM, C-CDA and FHIR specifications, and CDISC ODM) and terminologies/ontologies (eg, LOINC, ATC, and SNOMED CT)—as well as proprietary implementations for representing the

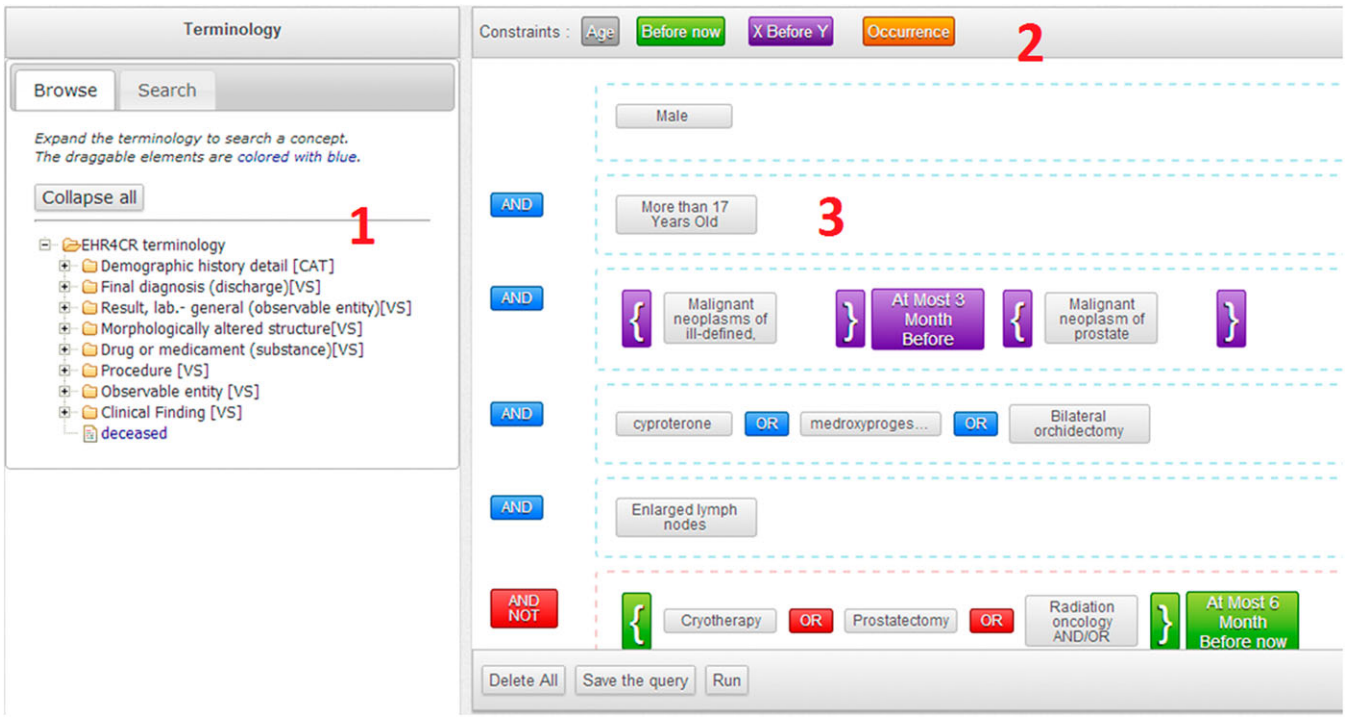

FIGURE 1 EHR4CR query builder demonstrating Protocol Feasibility Study module

content of health information in systems. Therefore, achieving broad-based, scalable, and computable semantic interoperability across multiple domains and systems requires a consistent use of multiple standards, clinical information models, and terminology models.

The *common EHR4CR semantic resources* consist of a shared set of standard-based templates and data elements with their associated value sets and concepts that enable to mediate across heterogeneous representations of patient-centric health information. The common EHR4CR semantic resources are stored and maintained in a metadata registry framework extending the ISO/IEC 11179 and are accessed through standardized interfaces—the *EHR4CR semantic interoperability services (SIS)*.

In this section, we describe the characteristics of the *EHR4CR Common Information Model (CIM)* regarding the 10 desiderata stated in the previous section.

- Req B.1: The mediation model shall be based on standard domain knowledge and reference models.

We considered the efforts in the domain of patient care, focusing on specifying both the syntax and the semantics of clinical information. The HL7 Reference Information Model (RIM) and EN ISO 13606 standards defined the semantics of patient care data and clearly demonstrate the need for "layers of semantic expressiveness" including the following: (i) generic reference information models of concepts and relationships (eg, EN ISO 13606, openEHR Reference Model, or HL7 RIM and additional FHIR specifications) each capable of binding terms from terminology models (eg, SNOMED-CT, and LOINC) and associated with a data type models such as ISO 21090; and (ii) more detailed models (eg, EN ISO 13606 or openEHR Archetypes/Templates, or HL7 Detailed Clinical Models, that instantiate generic reference models (eg, HL7's Clinical Document Architecture (CDA) meta-standard and the derived Continuity of Care Document (CCD) or FHIR resources).

The *EHR4CR Common Information Model (CIM)* consists in a set of multilingual semantic resources based on multiple standards (see Figures 1 and 2). The EHR4CR templates are based on FHIR resources (Patient, Encounter, Condition, Observation, Procedure, and Medication Statement) (see Table 1). FHIR-based resources were organized into categories based on HL7 CCD sections and UMLS semantic types: Demographics, Encounters, Advance directives, Problems, Family History, Social History, Alerts, Medications, Immunizations, Vital Signs, Results (lab, anatomic pathology), Procedures, Plan of Care, Lifestyle Choice, Ethical consideration. FHIR resources were enriched in order to fulfill the requirements of the project and represent the required semantic content. Some specific value sets were defined for some data elements of the FHIR templates.

- Req B.2: Use of standard terminologies, ontologies, and value sets that are multilingual and internationally used

EHR4CR templates are composed of data elements that are bound to a set of international reference terminologies selected by the project: ICD, SNOMED-CT, LOINC, ATC, ICD-O, Pubcan, TNM, PathLex. These terminologies are, when possible, imported into the collaborative editor from the official source of the terminology provider in order to bind the EHR4CR resources to up-to-date terminologies.

The terminology binding is done through the definition of value sets corresponding to the data elements of each template. Figure 2 illustrates the terminology binding done for the Observable entity: "ECOG performance status." The EHR4CR editing tool supports faceted templates. We defined a limited set of generic templates (eg, Observation) with facets, so that it is possible for each code of the template (eg, Observable entity SCT/423740007/ECOG performance status) to define its corresponding value set (eg, SCT/424122007/ECOG performance status finding).

As much as possible, we enriched and/or merged reference terminologies in order to build multilingual terminologies and value sets (in English, French at least and when possible in the 4 languages of the EHR4CR partners: English, French, German, and Polish).

- Req B.3: Possible use of internally defined extensions of existing standard terminologies (in order to add any missing concept or any missing description in any specific language)

An EHR4CR terminology was created in order to create concepts that are in the scope of the project but do not exist in the selected reference terminologies.

- Req B.4: Mappings between reference terminologies (eg, SNOMED-MedDRA, SNOMED CT-NCI Thesaurus)

We integrated the UMLS CUI in order to allow multi-terminology binding.

- Req B.5: Expressiveness

The current limited set of FHIR-based templates allows the representation of the main textual clinical data (signs, symptoms, diseases, outcome, procedures, care plans, etc.). We defined context-dependent value sets for representing multiple views or contextual information (eg, organ specific scores or histologic types).

- Req B.6: Scoped to the needs of the users of the research network in the context of dedicated use cases but scalable and sustainable (designed to be rapidly and efficiently scoped to cover any new requirement, extensible in terms of structure and content)

The EHR4CR mediation model *(EHR4CR CIM)* has been developed and can be extended, through a global consensus-based development process in order to cover the scope of both (i) eligibility criteria and data items identified from a given set of specific clinical trials (bottom up approach resulting in the creation of "useful data elements") and (ii) standard reference clinical information models or data elements (eg, CDISC SHARE) (top down approach). Although scoped to the needs of the users of the EHR4CR platform in the context of the 3 use cases of the project described earlier (PFS, PRS, or CTE) (see Figure 3), its structure ensures its scalability so that it can be extended in terms of structure and content to cover any new need. The *EHR4CR CIM* was developed and evolved through repeated cycles using a "Learning by Doing" approach in order

FIGURE 2 Copy screen of the EHR4CR collaborative editing tool. Left: Organization of FHIR-based resources into categories. The clinical observable entity: "Eastern Cooperative Oncology Group (ECOG) performance status" is defined using the template designed for clinical observations (see Table 1). Right: Terminology binding. The data element: "code" (DataType=ConceptDescriptor (CD)) is associated to a Value set defined as a set of TOP SNOMEDCT or LOINC codes, eg, SCT/423740007/ECOG performance status. The data element: "value" (DataType=ConceptDescriptor (CD)) is associated to a Value set defined as a set of concepts (ordered children of SCT/424122007/ECOG performance status finding: 0/SCT/425389002-ECOG 0; 1/SCT/422512005-ECOG 1; 2/SCT/422894000-ECOG 2; 3/SCT/423053003-ECOG 3; 4/SCT/423237006-ECOG 4; 5/SCT/423409001-ECOG 5)

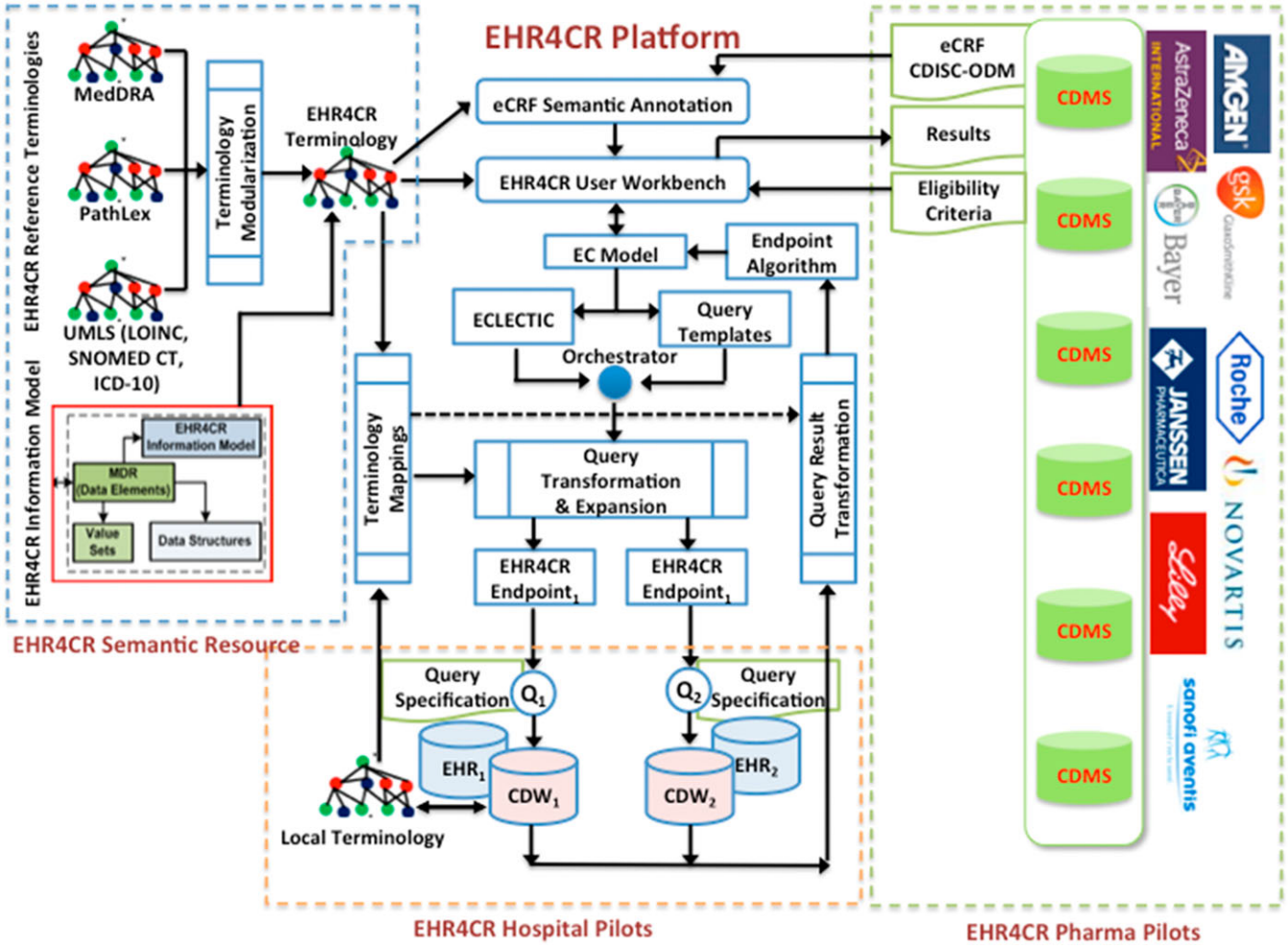

FIGURE 3 EHR4CR Semantic interoperability platform: a set of EHR4CR Semantic Resources and Semantic Interoperability Services (SIS) are used during the setup and execution phases of the EHR4CR use cases : Protocol Feasibility Study (PFS), Patient Identification and Recruitment Services (PRS) and Clinical Trial Execution (CTE).

to cover the scope of 14 first clinical trials selected to demonstrate the PFS use case, then 17 additional clinical trials selected to demonstrate the PRS use case and finally 28 additional clinical trials selected to demonstrate the CTE use case. Each new version of the *EHR4CR CIM* had an extended scope and improved quality.

The current version of the EHR4CR CIM includes 6 FHIR-based templates (and 6 additional specialized templates) and a subset of 15 corresponding data elements. Table 2 describes the content scope of the templates. Four **patient** demographic data elements (gender, birth time, deceased indicator, and deceased time) are part of the patient template. Four data elements (code, discharge disposition code, effective time, and length of stay) are part of the **Encounter** template. We distinguished 2 types of **Conditions**: diseases on one hand and signs and symptoms on the other hand. We defined 25 categories of diagnoses (including discharge diagnosis, primary diagnosis, secondary diagnosis, and admitting diagnosis). Diseases are encoding using codes from a value set combining ICD 10 (n = 12 318 codes) and a subset of SNOMED CT codes.

In the current version we defined 4 specialized **Observation** templates and defined clinical observable entities (n = 26), vital signs (n = 5), laboratory observable entities (n = 2000), and anatomic pathology observable entities (n = 80). Value sets corresponding to categorical observable entities were defined and populated with more than 1000 codes from SNOMED CT, ICD-O (Pubcan), TNM, PathLex, and EHR4CR-T.

We defined as part of the **Procedure** template a small value set SNOMED CT procedures (n = 57). As part of the **MedicationStatement**, we selected ATC (n = 5655 codes) as the value set attached to the data element consumableCode.

The terminology binding of the *EHR4CR CIM* involves more than 21 500 concepts from reference terminologies internationally used. All the concepts are at least bilingual (English and French).

- Req B.7: Standard formal languages allowing semantic reasoning

The semantic resources are stored within a semantic metadata repository. We use the term metadata (literally “data about data”) to distinguish “data collection structures” from patient data that populate those structures, ie, instance-level. Metadata should be described using a well-defined metadata schema so as to represent the semantics of the instance data and should include concepts and relationships as well as bindings to terminologies. A metadata scheme may be expressed in a number of different programming languages, eg, HTML, XML, UML, and RDF. We used the international standard ISO/IEC 11179 to define our metadata. This standard provides the definition of a “data element”

TABLE 2 Description and structure of the 6 core FHIR-based-templates of the EHR4CR mediation model

Template (nb. of data elements)	Template scope	Specialized template scope	Data element	Terminology binding Value set	Nb. of concepts
Patient (*n* = 4)	A Patient is a uniquely identified person. Clinical statements attached to this Patient may be recorded within the source systems.		administrativeGenderCode birthTime deceasedInd deceasedTime	SCT gender types	4
Encounter (*n* = 4)	An Encounter occurrence correspond to a period of time a Patient continuously receives medical services from one or more providers at a care site in a given setting within the health care system.		code dischargeDispositionCode effectiveTime lengthOfStayQuantity	SCT encounter types	6
Condition (*n* = 2)	Conditions state the presence of a clinical disease, sign or symptom, etc.	**nonDiseaseCondition:** correspond to symptoms (observed by the patient) or signs (observed by a care provider).	category code	SCT condition types Subset of SCT findings	4 16
		diseaseCondition: are inferred from medical claims data, textual clinical document, collected via forms (eg, from a problem list), etc.	category code	SCT diagnostic types Diseases (ICD10+subset of SCT diseases)	25 12500
clinicalObservation (*n* = 2)	A (numerical or categorical) Observation is a sign or a symptom or the result of any procedure which is either observed by a Provider or reported by the Patient.	**clinicalObservation**: records of measurements performed by a clinician at bed side (including scores, grades, stages, etc.)	name value	subset of SCT observable entities value sets specific to each categorical observable entity	26 95
		vitalSignObservation: refer to blood pressure, body temperature, pulse rate, and respiratory rate.	name value	subset of SCT vital signs	5
		laboratoryObservation: refer to laboratory tests.	name value	subset of LOINC codes (Top 2000) value sets specific to each categorical observable entity	2000 >500
		anatomicPathologyObservation: records of measurements performed by a pathologist analyzing tissues/cells with a microscope (including scores, grades, stages, etc).	name value	subset of LOINC codes (Top 80) value sets specific to each categorical observable entity (eg, ICD-O, TNM, etc)	80 >500
Procedure (*n* = 1)	A Procedure occurrence correspond to the record of an activity or process ordered by, or carried out by, a healthcare provider on the patient with a diagnostic or therapeutic purpose. Procedures are inferred from medical claims include, computerized orders in EHRs, etc.		code	subset of SCT procedures	57
Medication Statement (*n* = 2)	A medication statement is inferred from clinical events associated with orders, prescriptions written, pharmacy dispensing, procedural administrations, and other patient-reported information. Medication includes medicines, vaccines, and large-molecule biologic therapies.		administrationUnitCode consumableCode	ATC codes	6000

registry, describing disembodied data elements. It is important to note that ISO/IEC 11179 covers just the definition of elements and does not dictate the persistence structures or retrieval strategies. In the healthcare domain, another ISO standard—ISO 21090—plays a key role in the ISO/IEC 11179-based data element definitions because it provides the appropriate formal representation of the data type for Data Element Concept and of any type of the Value Domain data type. ISO 21090 especially provides a formal representation of the coded data types and addresses the binding with terminologies.

- Req B.8: Version management

Version management is provided for any type of semantic resource (terminologies, value sets, data elements, templates)

- Req B.9: Prototype of collaborative authoring tool

A tool was developed for authoring and maintaining the shared semantic resources of the mediation model. The *EHR4CR CIM Editor* allows a user to (Req. 1-4):

- Browse/search the repository of EHR4CR semantic resources (Common Element Templates (eg, observations, procedures, and substance administrations), Common Data Elements, Value Sets and Terminologies)
- Import semantic resources from external providers (eg, UMLS, BioPortal, HL7, and IHTSDO)
- Export any type of EHR4CR semantic resources in standard formats (eg, SKOS)
- Create/modify the model of the EHR4CR semantic resources
- Req B.10: Standard semantic services based on new web technologies

The *semantic interoperability services (SIS)* are developed to enable EHR4CR end-user services to assess and consume the semantic resources of the mediation model (terminologies, value sets, data elements, templates) and the mappings. SIS are used at the workbench by the *EHR4CR query builder* for query specification (representation of free text eligibility criteria using the data elements of the mediation model) and at the *EHR4CR endpoints* for query transformation. This goal was realized via the expansion of the original functionality outlined in HL7's Common Terminology Service—Release 2 (CTS2) Specification. The functional profiles of the SIS include capabilities for searching and query code system content, value set content, and template content. The technical specifications of the EHR4CR SIS rely on Representational State Transfer (REST)-based APIs/web services, recently been adopted by HL7.

3.3.1 | Standardization pipeline for data providers

In this section, we describe the characteristics of the *EHR4CR standardization pipeline* regarding the 5 requirements stated in the "Standardization pipeline" section of the conceptual framework.

- Req C.1Automatic mapping algorithms

Once hospital clinical data repositories are connected to the EHR4CR platform, their data source information models need to be mapped to the *EHR4CR CIM*. In the current state, the concepts used in the definitions of the central data elements were manually mapped to corresponding local terms used in pilot sites. Tools supporting this function are still under development. The current version of the Terminology Mapping Editor has limited functionalities: it allows the Terminology Mapper to upload subset of local value sets and to create their mapping to central value sets defined within the EHR4CR CIM.

- Req C.2: Natural language processing for semantic annotation of text: none
- Req C.3: Formal representation of mappings and version management: Mappings are available in SKOS format. Version management is provided.
- Req C.4: Use case driven support for prioritizing the mapping effort: none
- Req C.5: Standard semantic services: Mappings are available through REST-based APIs/web services

3.3.2 | Evaluation

This paper describes the design of the semantic interoperability architecture within the EHR4CR platform. As indicated above, successive versions of the platform were implemented during the project, and the iterative evolution of its design was informed by the experiences of its use by the 11 hospital pilot sites across 5 European countries in the UK, France, Germany, Poland, and Switzerland. Early evaluations used a specially created test data set of robustly de-identified patient records, in order to verify the reproducibility of the query results within different deployment settings. Later deployments and evaluations of the platform used locally de-identified data from each hospital site and therefore tested the full functionality of the semantic medication architecture described here. In order to undertake these evaluations, the semantic mediation architecture, Terminology Mapping Editor, and query language were used to map local data corresponding to data elements that had been identified the pilot sites as being required to respond to the majority of eligibility criteria within clinical trial protocols.[32] The final project evaluations have been published elsewhere and so are not reported here.[4,33] These evaluations report on the end to end success of the EHR4CR platform to accurately anticipate the eligible patient numbers that correspond to example clinical trial protocols, which therefore included the successful use and functioning of the semantic interoperability services documented here.

4 | DISCUSSION

With the development of platforms enabling the use of routinely collected clinical data in the context of international research, scalable solutions for cross border and cross domain semantic interoperability need to be developed.

An expression language, an underlying model of patient data and the codification of eligibility concepts are essential constructs for a

formal knowledge representation for eligibility criteria. There is currently an intense scientific focus directed toward developing and maintaining shareable, multipurpose, high-quality computable phenotype algorithms in order to mediate between different EHR products and healthcare systems.

4.1 | Contribution

4.1.1 | The evaluation framework

Mo et al[22] proposed 10 desired characteristics for a flexible, computable phenotype representation model.

We extended this list and proposed 23 requirements classified to match the 3 main components of a mediation platform for computing phenotype algorithms within a federated database system: (i) query language and model, (ii) patient data model (mediation model), and (iii) standardization pipeline for data providers. A correspondence with the desiderata for computable representations of EHRs-driven phenotype algorithms proposed by Mo et al is provided in Table 1.

4.1.2 | The EHR4CR case study

The EHR4CR mediation platform fulfills—at least partially—most of the 23 requirements of the proposed conceptual framework.

The **mediation model** is based on multiple standards: standard models (HL7 FHIR templates, ISO 21090, ISO11179), standard value sets, and terminologies. Integrating these different multi-level standards is challenging, and terminology binding is especially a difficult issue while contextual and versioning issues need to be addressed. We developed specific data structures—faceted templates—to get a good balance between complexity (a limited set of generic templates) and expressiveness (major scalability in terms of structure and content thanks to the facets). As much as possible, we enriched and/or merged reference terminologies in order to build multilingual terminologies and define multilingual value sets (at least in the 4 languages of the EHR4CR partners: English, French, German, and Polish). An EHR4CR terminology was created in order to create concepts that are in the scope of the project but do not exist in the selected reference terminologies.

We developed a prototype of a **collaborative editing tool** handling the management of any type of the EHR4CR complex semantic resources (faceted templates, data elements, value sets, concepts from huge and complex terminologies, eg, SNOMED CT) and of their relationships. We addressed the versioning issues for every type of resource, deriving CTS2 approaches for vocabulary updates. A Terminology Mapping Editor, under development, enables participant EHRs to develop and maintain **semantic mappings** between their proprietary models and the mediation model. This tool is still at its infancy and does not yet fulfill the expected requirements (such as use case driven support for prioritizing the mapping effort, contextual terminology mapping, automatic mapping algorithms addressing multilingual issues).

The semantic resources (mediation models and mappings) are accessible to any component of the EHR4CR platform through standardized **semantic services** based on new web technologies, such as Representational State Transfer (REST)-based APIs/web services, recently been adopted by HL7.

Within the EHR4CR project, we identified the need for a **governance body and process** for ensuring the quality of the data standardization pipeline within the network. Because a set of complex and sometimes time-consuming activities is required at the hospital side at the connection phase (initial mapping to a core of semantic resources) and at the setup phase of each new study (update of the mappings in the specific context of the study), it is important that those activities are well organized and properly synchronized with central efforts. Thus, it is not just a matter of content scope of the semantic resources but also a matter of reaching agreements on how they are represented and accessed. The governance body and process will be especially important in the context of any operational use of the EHR4CR platform at a broader scale within an extended network.

4.2 | Limits, related works, and perspectives

The evaluation framework of mediation platforms computing phenotype identification algorithms within federated databases was useful for identifying the limitations of the EHR4CR mediation platform. Some of these limitations correspond to biomedical informatics challenges that are subject to active ongoing research.

4.2.1 | The need of enhanced query engines

Expression languages employed to represent eligibility logic include ad hoc expressions (with or without the use of templates), the Arden Syntax, logic-based languages (ie, PAL, SQL, and DL), object-oriented languages (ie, GELLO), and temporal query languages (eg, Asbru and Chronus II).[34] Ad hoc formalisms are functional in many existing systems and can provide interesting features regarding expressiveness, as in the EHR4CR project. SQL-based queries on a clinical database are expressive but not extensible for knowledge re-use or inference. These mechanisms all suffer from the lack of scalability. Multiple query languages were used for different types of logic within the same model or system. Ontologies are increasingly being used as common terminological resources to automatically reconcile data heterogeneity and implement large-scale, distributed data management systems. Ontology-aware query interfaces that are integrated EHR systems can subsequently leverage the ontology annotations to support extensive query answering functionalities.[35]

4.2.2 | The need of shared high quality mediation models

Over the past decade, medical informatics researchers have been studying issues related to clinical information models associated with terminologies and have begun to articulate some requirements for **"high quality" models.**[22,34,36] There are several efforts trying to address the **interoperability between the clinical research and patient care domains** in building a common data model where the interoperating systems are required to interact through this well-defined mediation model. In this approach, a top-level knowledge model agreement is forced for the underlying data models of the interoperating parties for successful data exchange. Some projects, adopting this top-down strategy, proposed solutions that have been carried forward into practice, and new experience has been gained: OMOP CDM,[14] FDA Mini-Sentinel,[37] I2B2-SHRINE,[38,39] STRIDE,[40]

eMERGE,[24,25,41] SHARPn,[42,43] and other initiatives.[19,44,45] CDISC SHARE is an important initiative in addressing the interoperability between care and research domains through maintaining common data elements built upon BRIDG DAM where they are annotated with CDISC data sets like CDASH and SDTM, and other CDISC terminologies.[46] CDISC SHARE CDEs need to be considered for enriching the EHR4CR mediation model.

Several other efforts like the DebugIT[47] or SALUS projects[45] propose an ontological framework where each local system can continue to use its own local models and terminology systems, while both structural mapping and terminology mapping are handled through rule-based reasoning on formal representations of reference models and terminology systems. The rationale of these projects is that addressing syntactic and semantic interoperability should not be separated from each other, because the binding between models of use and models of meaning also has an impact on Semantic Interoperability.

CDEs and existing classification terms can be incorporated into domain ontologies to be used as a single domain model with multiple levels of abstraction that can be easily integrated with a variety of informatics tools.[35] Representing all the knowledge through formal means, as ontologies, and establishing the necessary links again through ontological constructs give an enhanced capability of semantic mediation and terminology reasoning.

Challenges that are usually not yet addressed in the ongoing projects are the use of terminology mappings between reference terminologies (eg. mappings between SNOMEDCT and MedDRA, NCI-T, ICD-9, ICD-10, ICD-O) in order to fully support multi-terminology binding and enhance the expressiveness of query engines. Representing multi-scale clinical data including molecular findings such as genomics information and representing multiple contextual views using existing formal models like CIMI models or FHIR templates are still challenging.

Developing a smart user interface for searching and/or browsing within complex semantic resources remains problematic. It is also important to enhance the tooling for collaborative edition of models by medical experts (using the user interface and/or CSV files) and to support a robust distribution process of these models (with 3 modes: full, snapshots, and/or deltas).

Another challenging issue is to extend the mediation platform in order to query federated data lakes combining unstructured and varied data and do not necessarily require a (often complex) master relational schema to structure and define all data.

International collaboration as part of initiatives such CIMI, FHIR, or the European Institute for Innovation through Healthcare Data (I--HD),[e] launched in March 2016 is important for improving the interoperability of healthcare systems through shared implementable clinical information models.

5 | CONCLUSION

Cross-border networking coordination and new technologies for data integration facilitates interoperability among research networks. Clinical research is on the threshold of a new era in which EHRs are gaining an important novel supporting role. The EHR4CR project developed an instance of a platform, providing communication, security, and semantic interoperability services to the 11 participating hospitals located in 5 European countries and 10 pharmaceutical companies.[17,48] This paper proposed desiderata for mediation platforms for computing phenotype algorithms within a federated database system and described the strengths and limitations of the EHR4CR mediation platform.

What was already known on the topic?

- Semantic interoperability is one of the main challenges to address to enable the reuse of hospital EHR data to support research.
- Several efforts aim at proposing a common information model used to mediate between heterogeneous EHRs within research networks.

What this study added to our knowledge?

- A common set of requirements for high-quality query languages, mediation models, and standardization pipelines can be defined.
- The EHR4CR mediation platform fulfills most of the requirements and demonstrated the feasibility of computing phenotype identification algorithms within multilingual federated databases
- Some requirements remain problematic
- The scope of the mediation model needs to be continuously adapted to the user's needs. Because the update can hardly be fully automatized (eg, through automatic coding of free text clinical trial protocols), a collaborative editor needs to efficiently support the creation of new semantic resources scoped to any additional use case.
- Despite recent efforts, formal representation of multimodal and multi-level data supporting data interoperability across clinical research and care domains is still challenging
- Terminology mapping in hospital sites is the major bottleneck of the data standardization pipeline. Supportive tools are still at their infancy
- Semantic interoperability within a broad international research network reusing clinical data from EHRs requires a rigorous governance process to ensure the quality of the data standardization process.

ACKNOWLEDGMENTS

We thank the members of EHR4CR WPG2 for their contribution to the design and development of the EHR4CR semantic interoperability platform.

AUTHORSHIP AND CONTRIBUTORSHIP

DO, SH, and CD designed the semantic interoperability platform with substantial input from ES, and NP; DO and ES developed the platform. CD, SH, and MJ developed the literature review method. CD and SH

[e]http://www.i-hd.eu/

drafted the original paper with the support of MJ. DO, NP, and DK revised the paper for submission.

COMPETING INTERESTS

None.

PROVENANCE AND PEER REVIEW

Not commissioned; externally peer reviewed.

REFERENCES

1. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med*. 2010;2(57):57cm29.
2. Martin Sanchez F, Gray K, Bellazzi R, Lopez-Campos G. Exposome informatics: considerations for the design of future biomedical research information systems. *J Am Med Inform Assoc*. 2014;21(3):386–390.
3. Moreno-Conde A, Moner D, da Cruz WD, et al. Clinical Information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. *J Am Med Inform Assoc*. 2015;22(4):925–934.
4. Doods J, Bache R, McGilchrist M, Daniel C, Dugas M, Fritz F. Work package 7. Piloting the EHR4CR feasibility platform across Europe. *Methods Inf Med*. 2014;53(4):264–268.
5. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. *Int J Med Inform*. 2011;80:371–388.
6. Schreiweis B, Trinczek B, Köpcke F, et al. Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. *Int J Med Inform*. 2014;83(11):860–868.
7. Niland J. ASPIRE: agreement on standardized protocol inclusion requirements for eligibility. In: An unpublished web resource; 2007.
8. Cohen Eea. caMATCH: a patient matching tool for clinical trials. In: caBIG Annual Meeting; 2005.
9. Musen MA, Tu SW, Das AK, Shahar Y. EON: a component-based approach to automation of protocol-directed therapy. *J Am Med Inform Assoc*. 1996;3:367–388.
10. Ohno-Machado L, Parra E, Henry S, Tu S, Musen M. AIDS2: a decision-support tool for decreasing physicians' uncertainty regarding patient eligibility for HIV treatment protocols. In: Proceedings of the Annual Symposium on Computer Application in Medical Care; 1993;429–433.
11. Sutton DR, Fox J. The syntax and semantics of the PROforma guideline modeling language. *J Am Med Inform Assoc*. 2003;10:433–443.
12. Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. *Artif Intell Med*. 1998;14:29–51.
13. Boxwala A. GLIF3: a representation format for sharable computer interpretable clinical practice guidelines. *J Biomed Inform*. 2004;37:147–161.
14. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc*. 2015;22(3):553–564.
15. Quaglini S, Stefanelli M, Lanzola G, Caporusso V, Panzarasa S. Flexible guideline-based patient careflow systems. *Artif Intell Med*. 2001;22:65–80.
16. Delaney BC, Curcin V, Andreasson A, et al. Translational medicine and patient safety in Europe: TRANSFoRm-architecture for the learning health system in Europe. *Biomed Res Int*. 2015;2015:961526
17. De Moor G, Sundgren M, Kalra D, et al. Using electronic health records for clinical research: the case of the EHR4CR project. *J Biomed Inform*. 2015;53:162–173.
18. El Fadly A, Rance B, Lucas N, et al. Integrating clinical research with the healthcare enterprise: from the RE-USE project to the EHR4CR platform. *J Biomed Inform*. 2011;44(Suppl 1):S94–102.
19. Jiang G, Evans J, Oniki TA, et al. Harmonization of detailed clinical models with clinical study data standards. *Methods Inf Med*. 2015;54(1):65–74.
20. Margolis R, Derr L, Dunn M, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*. 2014;21(6):957–958.
21. Schneeweiss S. Learning from big health care data. *N Engl J Med*. 2014;370(23):2161–2163.
22. Mo H, Thompson WK, Rasmussen LV, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc*. 2015;22(6):1220–1230.
23. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. *Genet Med*. 2013;15(10):761–771.
24. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc*. 2013;20(e1):e147–e154.
25. Pathak J, Wang J, Kashyap S, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc*. 2011;18:376–386.
26. Gennari J, Sklar D, Silva J. Cross-tool communication: from protocol authoring to eligibility determination. In: Proceedings of the AMIA symposium 2001;199–203.
27. Nammuni K, Pickering C, Modgil S, et al. Design-a-trial: a rule-based decision support system for clinical trial design. *Knowl-Based Syst*. 2004;17:121–129.
28. ERGO: a template-based expression language for encoding eligibility criteria, http://128.218.179.58:8080/homepage/ ERGO_Technical_Documentation.pdf/>; 2009 [accessed 24.11.15].
29. Jenders R, Sujansky W, Broverman C, Chadwick M. Towards improved knowledge sharing: assessment of the HL7 reference information model to support medical logic module queries. In: Proceedings of the AMIA annual fall symposium 1997;308–312.
30. Hammond WE, Jaffe C, Kush RD. Healthcare standards development. The value of nurturing collaboration. *J AHIMA*. 2009;80:44–50. quiz 51–52
31. Bache R, Taweel A, Miles S, Delaney BC. An eligibility criteria query language for heterogeneous data warehouses. *Methods Inf Med*. 2015;54(1):41–44.
32. Doods J, Botteri F, Dugas M, Fritz F. EHR4CR WP7. A European inventory of common electronic health record data elements for clinical trial feasibility. *Trials* 2014;15:18.
33. Soto-Rey I, Trinczek B, Girardeau Y, et al. Efficiency and effectiveness evaluation of an automated multi-country patient count cohort system. *BMC Med Res Methodol*. 2015;15:44.
34. Weng C, Tu SW, Sim I, et al. Formal representation of eligibility criteria: a literature review. *J Biomed Inform*. 2010;43:451–467.
35. Sahoo SS, Lhatoo SD, Gupta DK, et al. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. *J Am Med Inform Assoc*. 2014;21(1):82–89.
36. Ahn S, Huff SM, Kim Y, Kalra D. Quality metrics for detailed clinical models. *Int J Med Inform*. 2013;82(5):408–417.
37. Curtis LH et al. Design considerations, architecture, and use of the MiniSentinel distributed data system. *Pharmacoepidem Drug Saf*. 2012;21:23–31.
38. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc*. 2012;19:181–185.

39. McMurry AJ, Murphy SN, MacFadden D, et al. SHRINE: enabling nationally scalable multi-site disease studies. *PLoS One*. 2013;8(3) e55811:
40. Lowe HJ et al. STRIDE – an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*. 2009; 2009:391–395.
41. Herr TM et al. Practical considerations in genomic decision support: the eMERGE experience. *J Pathol Inform*. 2015 Sep 28;6:50
42. Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *J Am Med Inform Assoc*. 2013;20(e2): e341–e348.
43. Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. *J Biomed Inform*. 2012;45(4):763–771.
44. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21(2):221–230.
45. Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. *J Biomed Inform*. 2013;46(5):784–794.
46. CDISC SHARE. <http://www.cdisc.org/cdisc-share> [accessed 24.9.15]
47. Schober D, Boeker M, Bullenkamp J, et al. The DebugIT core ontology: semantic integration of antibiotics resistance patterns. *Stud Health Technol Inform*. 2010;160(Pt 2):1060–1064.
48. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. *J Intern Med*. 2013;274(6): 547–560.

How to cite this article: Daniel C, Ouagne D, Sadou E, Paris N, Hussain S, Jaulent M-C, Kalra D. Cross border semantic interoperability for learning health systems: the EHR4CR semantic resources and services. *Learn Health Sys*. 2017;1:e10014. doi: 10.1002/lrh2.10014