Cross border semantic interoperability for clinical research: the EHR4CR semantic resources and services

Christel Daniel, MD, PhD[1,2]; David Ouagne, PhD[1]; Eric Sadou[1,2]; Kerstin Forsberg[3]; Mark Mc Gilchrist[4]; Eric Zapletal[2], PhD; Nicolas Paris[2], Sajjad Hussain, PhD[1]; Marie-Christine Jaulent[1]; Dipka Kalra MD, PhD[5]

[1]INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; [2]AP-HP, Paris, France, [3]AstraZeneca, Sweden; [4]Dundee University, UK, [5]University College London, UK.

Abstract

With the development of platforms enabling the use of routinely collected clinical data in the context of international clinical research, scalable solutions for cross border semantic interoperability need to be developed. Within the context of the IMI EHR4CR project, we first defined the requirements and evaluation criteria of the EHR4CR semantic interoperability platform and then developed the semantic resources and supportive services and tooling to assist hospital sites in standardizing their data for allowing the execution of the project use cases. The experience gained from the evaluation of the EHR4CR platform accessing to semantically equivalent data elements across 11 European participating EHR systems from 5 countries demonstrated how far the mediation model and mapping efforts met the expected requirements of the project. 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: Electronic Health Records, Biomedical Research, Terminology as Topic, Interoperability, Data Integration and Standardization, Knowledge representation

1 Introduction

Electronic Health Records (EHRs) contain a large variety of patient-centric data and are gaining an important supporting role in different area such as clinical research, patient safety and comparative effectiveness [10,27]. Specific topics of interest include providing clinical trial planners with a better understanding of the eligible cohorts [6,11], supporting targeted patient recruitment [4,31] and “single-source data entry” at the point of clinical care [7,16].

However, because EHRs are not designed with a primary focus of cross-domain integration, initiatives for integrating EHRs and clinical research have been often limited to non-scalable, system (or vendor)-specific efforts [7, 31]. In an expanding research landscape, cooperation infrastructures are now being built to allow research projects to reuse patient data from federated EHR systems from many different sites in different countries and therefore in a multilingual settings. Non-standard, and often conflicting, vendor approaches to representing EHR data pose challenges to infrastructure developers, who must build solutions to work with clinical data across multiple formats.

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 (EFPIA)[20]. The aim of the project is to reduce the cost of conducting clinical trials, through better leveraging routinely collected clinical EHR data at key points in trial design and execution life-cycle. EHR4CR implementations have been installed at 11 pilot hospital sites within five 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).

Unfortunately, standards in clinical care and clinical research have often been developed through parallel – and therefore somewhat inconsistent – efforts. Furthermore, integration profiles collaboratively developed by CDISC and Integrating the Healthcare Enterprise (IHE) to enable the use of data collected in clinical research and pharmacovigilance areas are limited to resolving lexical/syntactic data integration issues [12]. They do not yet fully address the needs of initiatives like EHR4CR that address the semantic barriers. To fulfill this need, the challenge is to provide semantic alignment of data collected in disparate contexts under different EHR systems connected from 11 disparate hospital information systems in the framework of EHR4CR.

Our hypothesis is that cross-systems semantic interoperability is achievable internationally by implementing a consistent integrative semantic abstraction on top of existing application proprietary models. This additional “layer” of multilingual metadata then acts as a mediation model between systems’ metadata. The mediation model provides a homogeneous view of the clinical data available in disparate EHRs so that data users can access these data using a library of standard queries that have been written based on the mediation model. Mediation models must be based on the adoption and integration of multiple standards themselves being aligned to be consistent, coherent, and cross-compatible [9,19].

Our goal is to develop and evaluate a semantic interoperability platform consisting of 1) a standard-based expressive and scalable mediation model, 2) a set of mappings between each hospital’s system-specific metadata and the mediation model and 3) a set of semantic services operating during set up and execution phases of the EHR4CR use cases (PFS, PRS and CTE) that correspond to the three first steps of conducting a clinical trial.

2 Methods

Our approach consisted first of defining a set of technical requirements related to the main components of the semantic interoperability platform: 1) a mediation model, 2) an authoring tool for maintaining it, 3) supportive tooling for mapping local models to the mediation model within the hospital sites and 4) semantic services for accessing and using semantic resources and mappings. Second, as part of the evaluation of the use of the EHR4CR platform across the participating 11 EHR systems, we evaluated how far the development of the mediation model and the standardization efforts met the expected requirements of the project.

2.1 The need of high quality mediation model

The execution of the EHR4CR use cases requires identification of patient cohorts based on pre-defined eligibility criteria (PFS, PRS) and extraction of patient-specific data for pre-populating individual forms of a research protocol (CTE). In any case, a controlled mediation model is required to support federated access to heterogeneous data sources. We identified a (non exhaustive) set of 12 desiderata for the development of a standard, reusable, multipurpose controlled mediation model.

- Req.1: Used as part of a mediation model these semantic resources need to 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 (e.g. Meaningful Use Stage 3 in US).
- Req.2: Bound to widespread, internationally and multilingual used terminologies
- Req.3: Possibly bound to internally defined extensions of existing internationally used terminologies (in order to add any missing concept or any missing description in any specific language)
- Req.4: Possibly bound to different reference terminologies (in order to allow end users to access semantically equivalent content through different terminologies (e.g. SNOMED CT or MedDRA, SNOMED CT or NCI Thesaurus)
- Req.5: Expressive enough to represent 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.
- Req.6: Expressive enough to specimen related information, family related information, etc.
- Req.7: Expressive enough to represent multiple granularities, multiple consistent views, context representation
- Req.8: Scoped to the needs of the users of the EHR4CR platform in the context of the three use cases of the project (PFS, PRS or CTE)
- Req.9: Scalable and sustainable (designed to be rapidly and efficiently scoped to cover any new requirement, extensible in terms of structure and content)
- Req.10: Represented using standard formal languages allowing semantic reasoning (e.g. semantic web languages) in order to recognize redundancy or inconsistency
- Req. 11: 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 need to 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 need to address the versioning issues for any type of semantic resource.
- Req. 12: The semantic resources need to be 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.

2.2 The need of efficient tools to support data standardization within participants hospitals

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. Differences between the controlled vocabularies of two systems exist even when both systems were created by the same developers. 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 semantic mapping between the proprietary models and the mediation model. Mapping tools need to provide:

- Req. 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 need to use the descriptions and synonyms of the concepts.

- Req. 2: Automatic mapping algorithms addressing multi lingual issues supporting the mapping between terminologies in different languages
- Req. 3: Automatic mapping algorithms using existing mappings between reference terminologies (e.g. when local sources are mapped to a standard terminology which is not used in the mediation model (e.g. 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)
- Req. 4: Formal representation of mappings
- Req. 5: Version management of mappings
- Req. 6: 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 (e.g. data elements that are the most frequently used in distributed queries, data elements corresponding to a specific clinical trial running in the hospital, etc.)
- Req. 7: Standardized web-based access to mappings

3 Results

A first version of the EHR4CR semantic interoperability platform has been designed and implemented 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.

3.1 Mediation model: the EHR4CR Common Information Model

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

- Based on standards (Req.1)

We considered the efforts done 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: i) generic reference information models of concepts and relationships (e.g. EN ISO 13606, openEHR Reference Model, or HL7 RIM and additional FHIR specifications) each capable of binding terms from terminology models (e.g. SNOMED-CT, LOINC, etc.) and associated with a data type models such as ISO 21090; and ii) more detailed models (e.g. EN ISO 13606 or openEHR Archetypes/Templates, or HL7 Detailed Clinical Models (DCMs), that instantiate generic reference models (e.g. 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 figure 1 & 2). The EHR4CR templates are based on FHIR resources (Patient, Encounter, Condition, Observation, Procedure and MedicationStatement) (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 fulfil 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.

- Terminology binding (Req. 2-4)

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 (e.g. Observation) with facets, so that it is possible for each code of the template (e.g. Observable entity SCT/423740007/ECOG performance status) to define its

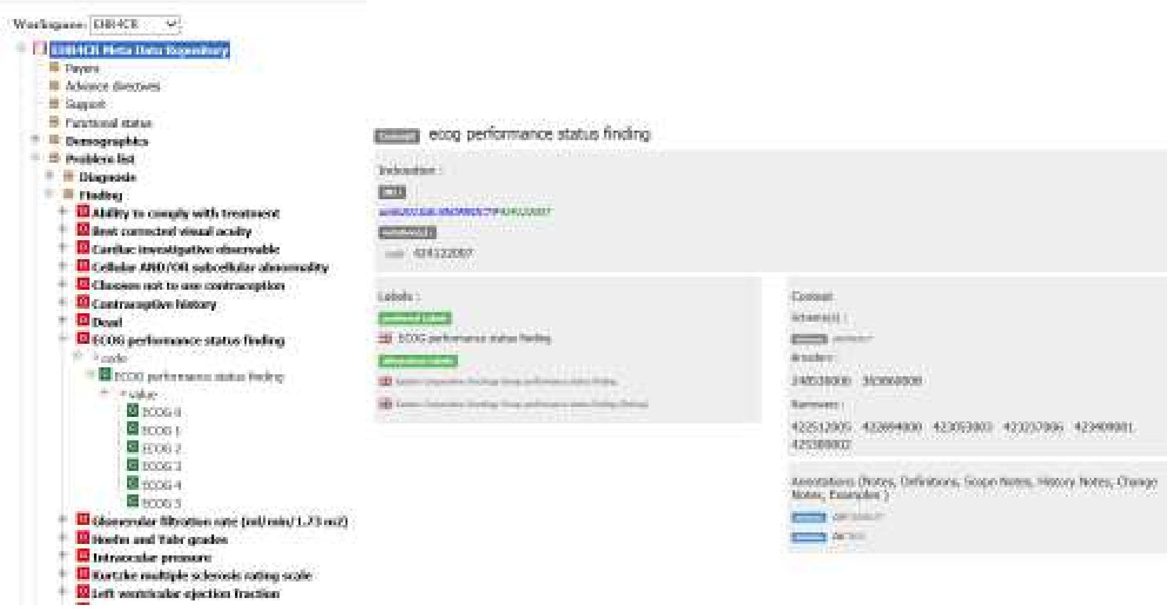
corresponding value set (e.g. SCT/424122007/ECOG performance status finding).
As much as possible, we enriched and/or merged reference terminologies in order to build multilingual terminologies and value sets (in English, French at least and when possible in the four languages of the EHR4CR partners: 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 also integrated the UMLS CUI in order to allow multi-terminology binding.

- Expressiveness (Req. 5-7)

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 (e.g. organ specific scores or histologic types, etc.).

- Scope and scalability (Req. 8-9)

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) standards reference clinical information models or data elements (e.g. CDISC SHARE) (top down approach). Although scoped to the needs of the users of the EHR4CR platform in the context of the three use cases of the project (PFS, PRS or CTE), its structure ensures its scalability so that it can be extended in terms of both 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 to cover the scope of 14 first clinical trials selected to demonstrate the PFS use case, then of 17 additional clinical trials (PRS use case) and finally of 28 additional clinical trials (CTE use case). Each new version of the EHR4CR CIM has an extended scope and improved quality.



Workspace: EHR4CR
EHR4CR Meta Data Repository
Payers
Advance directives
Support
Functional status
Demographics
Problem list
Diagnosis
Finding
Ability to comply with treatment
Best corrected visual acuity
Cardiac investigative observable
Cellular AND/OR subcellular abnormality
Chosen not to use contraception
Contraceptive history
Dead
ECOG performance status finding
code
ECOG performance status finding
value
ECOG 0
ECOG 1
ECOG 2
ECOG 3
ECOG 4
ECOG 5
Glomerular filtration rate (ml/min/1.73 m2)
Hoehn and Yahr grades
Intraocular pressure
Kurtzke multiple sclerosis rating scale
Left ventricular ejection fraction

ecog performance status finding
Indexation :
code 424122007
Labels :
ECOG performance status finding
Context
Ancestors :
Children :
422512005 422894000 423053003 423237006 423409001 425389002
Annotations (Notes, Definitions, Scope Notes, History Notes, Change Notes, Examples)

Figure 1: 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 e.g. SCT/423740007/ECOG performance status. *The data element: "value" (DataType=ConceptDescriptor (C*D)) 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).

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 1 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 two 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, admitting diagnosis, etc.). 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 four 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=5,655 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).

Template (nb. of data elements)	Template scope	Specialized template scope	Data element	Terminlogy 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	SCT gender types	4
			birthTime		
			deceasedInd		
			deceasedTime		
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	SCT encounter types	6
			dischargeDispositionCode		
			effectiveTime		
			lengthOfStayQuantity		
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	SCT condition types	4
			code	Subset of SCT findings	16
		diseaseCondition: are inferred from medical claims data, textual clinical document, collected via forms (e.g. from a problem list), etc.	category	SCT diagnostic types	25
			code	diseases (ICD10+subset of SCT diseases)	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	subset of SCT observable entities	26
			value	value sets specific to each categorical observable entity	95
		vitalSignObservation: refer to blood pressure, body temperature, pulse rate and respiratory rate.	name	subset of SCT vital signs	5
			value		
		laboratoryObservation: refer to laboratory tests.	name	subset of LOINC codes (Top 2000)	2000
			value	value sets specific to each categorical observable entity	>500
		anatomicPathologyObservation: records of measurements performed by a pathologist analyzing tissues/cells with a microscope (including scores, grades, stages, etc).	name	subset of LOINC codes (Top 80)	80
			value	value sets specific to each categorical observable entity (e.g. ICD-O, TNM, etc)	>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	A medication statement is inferred from clinical		administrationUnitCode		

Statement (n=2)	events associated with orders, prescriptions written, pharmacy dispensing, procedural administrations, and other patient-reported information. Medication includes medicines, vaccines, and large-molecule biologic therapies.	consumableCode	ATC codes	6000

Table 1: Description and structure of the six core FHIR-templates of the EHR4CR mediation model.

- Format (Req. 10)

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

- EHR4CR Collaborative editing tool (Req.11)

A tool was developed for authoring and maintaining the shared semantic resources of the mediation model. The EHR4CR CIM Editor allows to:

- Browse/search the repository of EHR4CR semantic resources (Common Element Templates (e.g. observations, procedures, substance administrations, etc.), Common Data Elements, Value Sets and Terminologies)
- Import semantic resources from external providers (e.g. UMLS, BioPortal, HL7, IHTSDO, etc.)
- Export any type of EHR4CR semantic resources in standard formats (e.g. SKOS)
- Create/modify the model of the EHR4CR semantic resources

- Semantic Interoperability Services (SIS) (Req.12)

Semantic services have been designed and developed to provide standardized interfaces to semantic resources to different types of applications. Application developers – such as developers of the EHR4CR end-user services for PRS, PRS & CTE - shall develop semantically enabled applications which can use standardized web services for accessing and consuming semantic resources. 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.

3.2 Mapping tools or data standardization in hospital sites

Once hospital clinical data repositories (CDRs) are connected to the EHR4CR platform, 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. Supporting tools are still under development. The current version of the Terminology Mapping Editor (TME) 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.

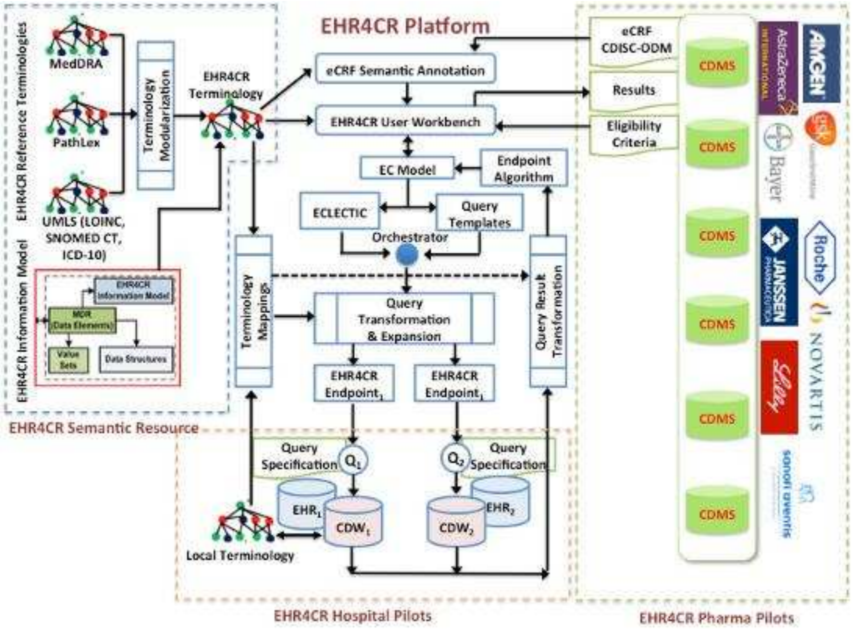


Figure 2. EHR4CR Semantic Interoperability platform: a set of EHR4CR Semantic Resources and Semantic Interoperability Services (SIS) are used during the setup and execution phases of the EHR4CR use case.

4 Discussion

With the development of platforms enabling the use of routinely collected clinical data in the context of international clinical research, scalable solutions for cross border and cross domain semantic interoperability need to be developed. There is currently an intense focus directed to the issue of developing and maintaining shareable, multipurpose, high-quality mediation models.

4.1 Contribution

The EHR4CR semantic interoperability platform fulfills most of the initial requirements initially identified based on the authors' experience with attempts to adopt models and vocabularies [24] and gleaned from the reported experiences of others.

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 four languages spoken by 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 **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 e.g. 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 (TME), 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 fulfil the expected requirements (such as use case driven support for prioritizing the mapping effort, contextual terminology mapping, automatic mapping algorithms addressing multi lingual 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.

4.2 Limits, related works and perspectives

Our current **mediation model** does not fully fulfil some of the ten requirements. We are considering, in the future, to integrate terminology mappings between reference terminologies (e.g. mappings between SNOMEDCT and MedDRA, NCI-T, ICD-9, ICD-10, ICD-O) in order to fully support multi-terminology binding. We still are working to represent multiple granularities, multiple consistent views, context representation. We plan to evaluate the FHIR resources currently being developed in order to represent multi-scale clinical data including molecular findings such as genomics information.

We still need to define complex templates allowing the combination of basic templates. Developing a smart user interface for searching and/or browsing within complex semantic resources remains problematic. We also plan to improve the collaborative editing of these resources by medical experts using the GUI and/or CSV files. We are also working on an improved distribution model (with three modes: full, snapshots and/or deltas).

Regarding the **data standardization process in hospitals**, the Terminology Mapping Editor is still at its infancy and does not yet fulfil the expected requirements (such as use case driven support for prioritizing the mapping effort, contextual terminology mapping, automatic mapping algorithms addressing multi lingual issues)

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** [1,21,34]. 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 top-down 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 [29], FDA Mini-Sentinel [5], I2B2-SHRINE [15,22], STRIDE [17], eMERGE [23,25], SHARPn [26,28] and other initiatives [14,31,33,34]. 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 [2]. CDISC SHARE CDEs need to be considered for enriching the EHR4CR mediation model. In the SALUS project, Sinaci et al. also applied a comprehensive set of semantic web technologies with the commonly adopted MDR standard – ISO/IEC 11179. In addition, they built a federated semantic MDR framework and demonstrated that it was possible to semantically link disparate CDE definition efforts by different organizations [32].

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. Since 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 set up 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.

5 Conclusion

Clinical research is on the threshold of a new era in which electronic health records (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 eleven participating hospitals located in five European countries and ten pharmaceutical companies [Coorevits13, Moor14]. This paper described the strengths and limitations of the EHR4CR semantic interoperability platform.

What was already known on the topic?

- Semantic interoperability is one of the main challenge to address to enable the reuse of hospital EHR data to support clinical research studies.
- 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 a "high-quality" semantic interoperability platform can be defined
- The EHR4CR mediation model fulfill most of the requirements, but some remain problematic
    - The scope of the mediation model needs to be continuously adapted to the user's needs. Since the update can hardly be fully automatized (e.g. 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.

Acknowledgements: We thank the members of EHR4CR WPG2 for their contribution to the design and development of the EHR4CR semantic interoperability platform. **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) **Authorship and Contributorship:** DO, SH, CD designed the semantic interoperability platform with substantial input from ES, EZ, NP, MM. DO and ES developed the platform with substantial input from EZ. CD, SH and MJ developed the literature review method. CD and SH drafted the original paper with the support of DK and MJ. DO, NP, KF and DK revised the

paper for submission. **Competing interests**: None. **Provenance and peer review**: Not commissioned; externally peer reviewed.

6 References

1. Ahn S, Huff SM, Kim Y, Kalra D. Quality metrics for detailed clinical models. Int J Med Inform. 2013 May;82(5):408-17.
2. CDISC SHARE. <http://www.cdisc.org/cdisc-share> [accessed 09/24/15]
3. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med. déc 2013;274(6):547-560.
4. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. Int J Med Inform 2011;80:371–88.
5. Curtis LH et al. Design considerations, architecture, and use of the MiniSentinel distributed data system. Pharmacoepidem Drug Saf 2012;21:23–31
6. 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 Jan 10;15:18.
7. 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.
8. European Commission. Semantic Interoperabiloity for Better Health and Safer Healthcare. 2009.http://ec.europa.eu/information_society/activities/health/docs/publications/2009/2009semantic-health-report.pdf
9. Hammond WE, Jaffe C, Kush RD. Healthcare standards development. The value of nurturing collaboration. J AHIMA 2009;80:44–50; quiz 51–52.
10. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. août 2013;51(8 Suppl 3):S30-37.
11. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 1 janv 2013;20(1):117-121.
12. IHE_QRPH_Data Element Exchange (DEX). http://www.ihe.net/Technical_Framework/upload/IHE_QRPH_Suppl_DEX_Rev1-0_PC_2013-06-03.pdf [accessed 09/24/15]
13. ISO/IEC 11179. Home Page for ISO/IEC 11179 Information Technology -- Metadata registries. ISO/IEC 11179, Information Technology -- Metadata registries (MDR). 2013. http://metadata-standards.org/11179/ [accessed 09/24/15]
14. Jiang G, Evans J, Oniki TA, Coyle JF, Bain L, Huff SM, Kush RD, Chute CG. Harmonization of detailed clinical models with clinical study data standards. Methods Inf Med. 2015 Jan 12;54(1):65-74.
15. 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;19:181–5.
16. Kush R, Alschuler L, Ruggeri R, et al. Implementing Single Source: the STARBRITE proof-of-concept study. J Am Med Inform Assoc 2007;14:662–73.
17. Lowe HJ et al. STRIDE – an integrated standards-based translational research informatics platform. AMIA Annu Symp Proc 2009;2009:391–5.
18. Ouagne D, Hussain S, Sadou E, et al. The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. Stud Health Technol Inform 2012;180:534–8.
19. Pathak J et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc 2011;18:376–86.
20. Tao C et al.Towards semantic-web based representation and harmonization of standard meta-data models for clinical studies. AMIA Summits Transl Sci Proc 2011;2011:59–63
21. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS ONE. 2013;8(3):e55811.
22. Mead CN. Data interchange standards in healthcare IT--computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? J Healthc Inf Manag 2006;20:71–8.
23. Moor GD, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, Karakoyun T, Ohmann C, Lastic PY, Ammour N, Kush R, Dupont D, Cuggia M, Daniel C, Thienpont G, Coorevits P. Using Electronic Health Records for Clinical Research: the Case of the EHR4CR Project. J Biomed Inform. 2014 Oct 18.
24. Moreno-Conde, A., Moner, D., da Cruz, W.D., Santos, M.R., Maldonado, J.A., Robles, M., Kalra, D. Clinical Information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J. Am. Med. Inform. Assoc. 2015;
25. Murphy SN, Dubey A, Embi PJ, et al. Current State of Information Technologies for the Clinical Research Enterprise across Academic Medical Centers. Clinical and translational science 2012;5:281–4.
26. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. juin 2013;20(e1):e147-154.
27. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc. déc 2013;20(e2):e341-348.
28. Prokosch H-U, Ries M, Beyer A, et al. IT infrastructure components to support clinical care and translational research projects in a comprehensive cancer center. Stud Health Technol Inform 2011;169:892–6.
29. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J. Biomed. Inform. 2012 Aug;45(4):763–71.
30. Reisinger SJ et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc 2010;17:652–62.
31. Schreiweis B, Trinczek B, Köpcke F, Leusch T, Majeed RW, Wenk J, Bergh B, Ohmann C, Röhrig R, Dugas M, Prokosch HU. Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. Int J Med Inform. 2014 Nov;83(11):860-8.
32. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014 Mar-Apr;21(2):221-30.
33. Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. J Biomed Inform. oct 2013;46(5):784-794.
34. Weng C, Tu SW, Sim I, et al. Formal representation of eligibility criteria: a literature review. J Biomed Inform 2010;43:451–67.