Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-534

The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology

David OUAGNE[a], Sajjad HUSSAIN[a], Eric SADOU[a], Marie-Christine JAULENT[a] and Christel DANIEL[a,b,c,1]
[a]UMR_S 872, Eq. 20, Paris, F-75006 France; [b]Université Paris Descartes, Paris, F-75006 France; [c]AP-HP, Hôpital George Pompidou, Département d'Informatique Hospitalière, Paris, F-75015 France

Abstract. A major barrier to repurposing routinely collected data for clinical research is the heterogeneity of healthcare information systems. Electronic Healthcare Record for Clinical Research (EHR4CR) is a European platform designed to improve the efficiency of conducting clinical trials. In this paper, we propose an initial architecture of the EHR4CR Semantic Interoperability Framework. We used a model-driven engineering approach to build a reference HL7-based multidimensional model bound to a set of reference clinical terminologies acting as a global as view model. We then conducted an evaluation of its expressiveness for patient eligibility. The EHR4CR information model consists in one fact table dedicated to clinical statement and 4 dimensions. The EHR4CR terminology integrates reference terminologies used in patient care (e.g LOINC, ICD-10, SNOMED CT, etc). We used the Object Constraint Language (OCL) to represent patterns of eligibility criteria as constraints on the EHR4CR model to be further transformed in SQL statements executed on different clinical data warehouses.

Keywords. Clinical research, semantic interoperability, electronic health record, clinical data warehouses, eligibility criteria, controlled vocabulary

Introduction

The EHR4CR (Electronic Health Records for Clinical Research) project aims to improve the efficiency and reduce the cost of conducting clinical trials, through better leveraging of routinely collected clinical data at key points in the trial design and execution life-cycle [1]. The EHR4CR platform will implement 4 use cases – clinical protocol feasibility, patient identification and recruitment, clinical trial execution and adverse event reporting – to be demonstrated by 10 pilots in 5 European countries. A major barrier to repurposing clinical data of Electronic Healthcare Records (EHRs) or Clinical Data Warehouses (CDWs) during clinical trial design and execution is that information systems in both domain–patient care and clinical research use different schemas and terminology systems. The collective international efforts of multiple

[1] Corresponding Author. Christel Daniel, E-mail: christel.daniel@crc.jussieu.fr

organizations (such as ISO, HL7, CDISC, etc) currently focuses on defining the various standards required to achieve computable semantic interoperability and to bridge the gap between clinical research and patient care. The international initiative *Integrating the Healthcare Enterprise (IHE)* has defined, as part of pre-population data templates, some mappings between patient care templates and clinical research templates for clinical trial execution and adverse event reporting scenarios [2]. However, the limitation is that once the pre-population templates are modified due to emerging requirements, new mappings are needed. We addressed this shortcoming in a previous work proposing a dynamic mapping mechanism supported by the use of SNOMED CT as the "pivot terminology" to facilitate mappings [3]. We argue that integrating patient care and clinical research domains requires a standard-based expressive and scalable semantic interoperability framework, allowing dynamic mappings between data structures and semantics of varying data sources. There have been various attempts for solving the semantics gap between medical terminologies, ontologies and information models [4, 5], and also generating a networked knowledge-base from available medical ontologies using Semantic Web technologies [6]. With regards to eligibility determination an additional issue is the definition of a formal representation of free-text eligibility criteria [7, 8].

1. Methods: EHR4CR Semantic Interoperability Framework

In this paper we propose the EHR4CR Semantic Interoperability Framework for consistent interpretation of clinical data accessed from varying sources, and demonstrate the expressiveness and computability of the EHR4CR framework for eligibility determination. The core of the EHR4CR semantic interoperability framework is a shared conceptual reference model (**EHR4CR information model**) acting as a global as view model to correlate the schemas and concepts from different sources. The EHR4CR information model is a HL7-based UML model annotated with the concepts of a shared terminology (**EHR4CR terminology**). The EHR4CR terminology is available in the EHR4CR portal, based on LexEVS, a terminology server designed to support terminology services (see Figure 1).

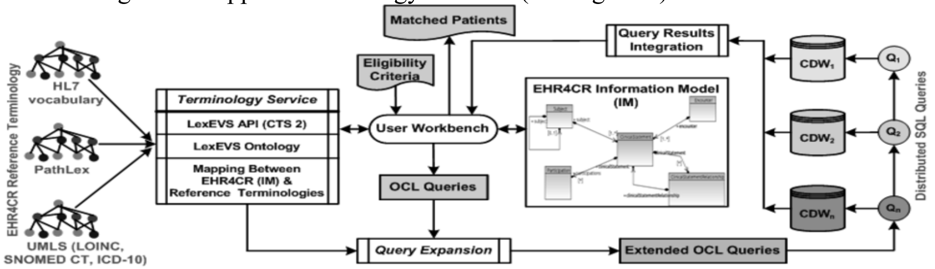

Figure 1. Overview of the EHR4CR Semantic Interoperability Framework: Semantic mediation between various Clinical Data Warehouses (CDWs) in pilot sites based on the EHR4CR information model (global as view model) bound to the EHR4CR terminology.

1.1. Representing Eligibility Criteria using OCL

Accordingly to [7] we distinguish two different possible use of formal representation of eligibility criteria: i) searching for already defined eligibility criteria to be re-used in the design of a new clinical trial and ii) patient eligibility determination. In both use cases, the expression language serves to formally model the relationships between multiple concepts embedded within eligibility criteria statements. But in the context of eligibility determination, criteria shall be represented using a formal query language designed to operate on any given model of patient data ("clinical information model") in order to build queries running on EHRs or CDWs. Languages of varying expressiveness have been used to represent the logic of eligibility criteria, including ad hoc expressions, Arden Syntax, variants of logic-based and description logic languages, Structured Query Language (SQL) and object-oriented query languages [8]. In the EHR4CR project, we use Object Constraint Language (OCL) to distribute queries over our EHR4CR UML-based information model, which provides a standard interface to heterogeneous EHRs or CDWs [9].

1.2. EHR4CR Information Model

We used Open Medical Development Framework (OMDF) [10]–an extended UML modeler–to transform HL7 v3 models in UML models and adapt these models to the purpose and scope of the EHR4CR project. We considered the HL7 version 3 model A_SupportingClinicalStatementUniversal, a component of the StudyDesign proposed by the HL7 Regulated Clinical Research Information Model (RCRIM) Work Group for clinical research protocol representation[2]. We transformed this model into a multidimensional model that contains one central class *ClinicalStatement*, 4 dimension classes and 82 attributes. The central class *ClinicalStatement* is the result of the merge of the *Act* classes of the «A_SupportingClinicalStatementUniversal» model. The 4 dimensions attached to the central class are: i) *Subject*, represents the information related to the subject of the clinical statement; ii) *Encounter*, represents the information related to the administrative context of the clinical statement; iii) *Participation*, represents additional information related to the clinical statement; iv) *ClinicalStatementRelationship*, represents the relationships between clinical statements.

1.3. EHR4CR Terminology

Representing criteria for eligibility determination requires a network of terminologies for clinical findings, test results, labs, or medications, etc. We use LexEVS to build the shared terminology (EHR4CR terminology). The current version of the EHR4CR terminology contains various concepts of reference terminologies/ontologies that are uploaded from UMLS (SNOMED CT, LOINC, ICD-10 codes, etc.) or other sources.

2. Evaluation & Results

We selected 10 clinical trials promoted by pharmaceutical companies involved in the EHR4CR project and running in more than one of the 10 pilot sites. These clinical trials are both interventional and non interventional studies related to different domain areas (cardiovascular, oncology, nervous system disorders, etc). From the 269 free-text eligibility criteria of the 10 clinical trials, 99 have been manually pre-processed and translated into 186 elementary queries. These queries were represented using 17 query

[2] Study Design Model, HL7 Std. www.hl7.org/v3ballot/html/domains/uvrt/editable/PORT_RM100002UV.html

templates which were formally represented as constraints on the EHR4CR information model using OCL[3]. Medical concepts of the queries were encoded using the EHR4CR terminology. We assessed the extent to which OCL rules capture the semantics of the eligibility criteria. Table 1 shows an example of a free text eligibility criteria and its corresponding formal OCL query. The OCL queries were designed to be distributed to endpoints in pilot sites and transformed into SQL statements to be executed on heterogeneous information models of legacy CDWs in order to screen patients for potential eligibility for the selected clinical trials.

Table 1. An example of a free text eligibility criteria and its corresponding formal OCL query

<table>
<tr><th>Initial criteria</th><th>OCL Query</th></tr>
<tr><td>**female**, be either **post-menopausal** for at **least 2 years**, **surgically sterilized** or have undergone **hysterectomy** or, if of child bearing potential, be willing to avoid pregnancy by using an **adequate method of contraception** for four weeks prior to, during and four weeks after the last dose of trial medication.</td><td>**def**: getCountECT02(date : TS): Integer =
Subject.allInstances()->select(sbj: Subject |
sbj.entityClassCode.code = 'PAT'
and sbj.determinerCode.code = 'INSTANCE'
and sbj.administrativeGenderCode.code = 'Female'
and (
sbj.clinicalStatements->exists(cs: ClinicalStatement |
cs.classCode.code = 'OBS'
and cs.code.code = 'Menopause'
and cs.effectiveTime.terms->exists(ts | ts.oclAsType(TS).lessThan(date).value))
or sbj.clinicalStatements->exists(cs: ClinicalStatement |
cs.classCode.code = 'PROC'
and cs.moodCode.code = 'EVN'
and cs.code.code = 'Hysterectomy')
or sbj.clinicalStatements->exists(cs: ClinicalStatement |
cs.classCode.code = 'OBS'
and cs.moodCode.code = 'EVN'
and cs.code.code = 'Adequate method of contraception')
)
)->size()</td></tr>
</table>

3. Discussion & Conclusion

The use of EHRs or CDWs for eligibility determination requires semantic matching between representations of eligibility criteria and representations of patient data in heterogeneous clinical systems. It is an active research area with challenges such as the semantic gap between eligibility criteria and both structured and free-text patient data. In EHR4CR, the representation of eligibility criteria requires (i) an expressive language to define executable eligibility rules, (ii) a patient information model, and (iii) an appropriate clinical terminology to facilitate mapping from eligibility concepts to patient data. Recent systems developed for eligibility determination have largely adopted sophisticated patient information models, providing an abstraction layer for EHRs or CDWs. Some of these models are based on the HL7 Reference Information Model (RIM), with varying degrees of adoption (including for instance, only one *Observations* class [11-13]. In EHR4CR, we propose a simplified information model based on the HL7 «A_SupportingClinicalStatementUniversal» model. Since our information model is multidimensional, it is well suited for querying CDWs.

Similar to other recent systems [7,8], in EHR4CR, we face issues related to (i) labor-intensive manual task of transforming free-text eligibility criteria in formal queries, (ii) expressiveness of the query language (including the possibility of query expansion), and (iii) creation and maintenance of the shared controlled terminology as

[3] Using Eclipse OCL, an implementation of the Object Constraint Language (OCL) OMG standard (http://www.eclipse.org/modeling/mdt/?project=ocl)

well as of the mapping between the EHR4CR/local information models and terminologies. In our approach, we deal with the above mentioned issues. We plan to extract key eligibility concepts and support flexible mappings to a range of terminologies in using Natural Language Processing (NLP) techniques [7]. We also plan to extend OCL in order to better represent temporal information and clinical context [14]. Then we plan to develop specific loaders for LexEVS in order to enrich the current version of the EHR4CR terminology with terminologies that are not yet in UMLS (such as HL7, IHE, CDA and PathLex vocabularies, etc), and to define mappings between the EHR4CR/local information models and terminologies in LexEVS. The main objective for defining these mappings is to exploit them for extending the user-defined eligibility criteria and to generate more comprehensive and extended queries. Based on the given eligibility criteria defined in an OCL query and the defined mappings in the terminology server, we aim to apply ontology-based query expansion techniques and distribute extended queries across different EHRs or CDWs.

References

[1] Electronic Healthcare Record for Clinical Research (EHR4CR) [Online]. Available: http://www.ehr4cr.eu/

[2] IHE Quality, Research and Public Health (QRPH) Technical Framework Supplement Clinical Research Document (CRD), August 2010. [Online]. Available: http://www.ihe.net/Technical Framework/upload/IHE QRPH Suppl CRD Rev2-2 TI 2010-08-30.pdf

[3] El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, Jaulent MC, Daniel C. Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. J Biomed Inform. 2011;44 Suppl 1:S94-S102.

[4] Schulz S, Schober D, Daniel C, Jaulent M. Bridging the semantics gap between terminologies, ontologies, and information models. Stud Health Technol Inform 2010; 160(Pt 2: 1000–1004.

[5] Sahay R, Zimmermann A, Fox R, Polleres A, Hauswirth M. Interoperability in Healthcare Information Systems: Standards, Management, and Technology, chapter A Formal Investigation of Semantic Interoperability of HCLS Systems. IGI Global; 2011.

[6] Ghazvinian A, Noy NF, Jonquet C, Shah N, Musen MA. What four million mappings can tell you about two hundred ontologies. In Proceedings of the 8th International Semantic Web Conference, pages 229–242. Springer-Verlag; 2009.

[7] Tu SW, Peleg M, Carini S, Bobak M, Ross J, Rubin D, Sim I. A practical method for transforming free-text eligibility criteria into computable criteria. J Biomed Inform. 2011;44(2):239-50.

[8] Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform. 2010;43(3):451-67.

[9] Object Constraint Language, Version 2.0. [Online]. Available: http://www.omg.org/spec/OCL/2.0/PDF/

[10] Ouagne D, Nadah N, Schober D, Choquet R, Teodoro D, Colaert D, Schulz S, Jaulent MC, Daniel C. Ensuring HL7-based information model requirements within an ontology framework. Stud Health Technol Inform. 2010;160(Pt 2):912-6.

[11] Johnson P, Tu S, Musen M. A virtual medical record for guideline-based decision support. In: Proceedings of the AMIA annual fall symposium; 2001. p.294–8.

[12] 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. p. 308–12.

[13] Lonsdale DW, Tustison C, Parker CG, Embley DW. Assessing clinical trial eligibility with logic expression queries. Data Knowl Eng 2007;66(1):3–17.

[14] Sordo M, Boxwala A, Ogunyemi O, Greenes R. Description and status update on GELLO: a proposed standardized object-oriented expression language for clinical decision support. Stud Health Technol Inform 2004; 107:164–8.