Computers in Biology and Medicine 87 (2017) 179–186

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A semantic interoperability approach to support integration of gene expression and clinical data in breast cancer

Raul Alonso-Calvo [a,*], Sergio Paraiso-Medina [a], David Perez-Rey [a], Enrique Alonso-Oset [a], Ruud van Stiphout [b], Sheng Yu [b], Marian Taylor [b], Francesca Buffa [b], Carlos Fernandez-Lozano [c], Alejandro Pazos [c], Victor Maojo [a]

[a] Biomedical Informatics Group, DIA & DLSIIS, ETSI Informáticos, Universidad Politécnica de Madrid, Spain
[b] Department of Oncology, Old Road Campus Research Building, Oxford, OX3 7DQ, United Kingdom
[c] Department of Information and Communication Technologies, Faculty of Computer Science, University of A Coruna, 15071, A Coruña, Spain

ARTICLE INFO

Keywords:
Clinical research informatics
Semantic interoperability
Data integration
Diagnostic classifier
Gene expressions
Biomedical terminologies

ABSTRACT

Introduction: The introduction of omics data and advances in technologies involved in clinical treatment has led to a broad range of approaches to represent clinical information. Within this context, patient stratification across health institutions due to omic profiling presents a complex scenario to carry out multi-center clinical trials.
Methods: This paper presents a standards-based approach to ensure semantic integration required to facilitate the analysis of clinico-genomic clinical trials. To ensure interoperability across different institutions, we have developed a Semantic Interoperability Layer (SIL) to facilitate homogeneous access to clinical and genetic information, based on different well-established biomedical standards and following International Health (IHE) recommendations.
Results: The SIL has shown suitability for integrating biomedical knowledge and technologies to match the latest clinical advances in healthcare and the use of genomic information. This genomic data integration in the SIL has been tested with a diagnostic classifier tool that takes advantage of harmonized multi-center clinico-genomic data for training statistical predictive models.
Conclusions: The SIL has been adopted in national and international research initiatives, such as the EURECA-EU research project and the CIMED collaborative Spanish project, where the proposed solution has been applied and evaluated by clinical experts focused on clinico-genomic studies.

1. Introduction

Clinical trial complexity is dramatically increasing as new genetic and molecular variables are gathered in clinical settings [1]. Due to the costs of such clinical studies and challenges for recruiting trial cohorts, they often involve multiple clinical institutions [2]. New data management methods are therefore required by clinical users and investigators from institutions involved in multi-center clinical research [3]. In most cases, researchers need to know the different data representations of the institutions participating in the study and significant manual data management is required [4]. To facilitate certain processes required to achieve semantic integration from heterogeneous sources in the area (e.g., clinical trial management systems, electronic health records or laboratory systems, among others) (semi-) automatic methods have been recently addressed by international initiatives [5].

Several efforts have recently focused on facilitating communication and exchange of information between clinical systems by using biomedical standards [6]. In general, interoperability initiatives provide an underlying data model for different areas. Examples of these initiatives are, to mention a few relevant examples, the Observational Medical Outcomes Partnership (OMOP) [7], Integrating Biology and the Bedside (i2b2) [8], the HL7 Reference Information Model (RIM) [9], Fast Healthcare Interoperability Resources (FHIR) [10], Integrating the Healthcare Enterprise (IHE) [11] or PCORnet [12]. These initiatives have been developed with the objective of obtaining valuable results in the clinical research area. Few translational research platforms have actually

* Corresponding author. DLSIIS, ETSI Informáticos, Universidad Politécnica de Madrid Campus de Montegancedo S/N, 28,660, Boadilla del Monte, Spain.
E-mail addresses: ralonso@infomed.dia.fi.upm.es (R. Alonso-Calvo), sparaiso@infomed.dia.fi.upm.es (S. Paraiso-Medina), dperez@infomed.dia.fi.upm.es (D. Perez-Rey), enriquealonso@infomed.dia.fi.upm.es (E. Alonso-Oset), ruud.vanstiphout@oncology.ox.ac.uk (R. van Stiphout), sheng.yu@oncology.ox.ac.uk (S. Yu), marian.taylor@oncology.ox.ac.uk (M. Taylor), francesca.buffa@oncology.ox.ac.uk (F. Buffa), carlos.fernandez@udc.es (C. Fernandez-Lozano), apazos@udc.es (A. Pazos), vmaojo@infomed.dia.fi.upm.es (V. Maojo).

http://dx.doi.org/10.1016/j.compbiomed.2017.06.005
Received 19 January 2017; Received in revised form 30 May 2017; Accepted 2 June 2017

0010-4825/© 2017 Elsevier Ltd. All rights reserved.

exploited the benefits of the analysis and interaction of interoperability models with genetic information and related terminologies [13], i.e. tranSMART platform that is based on i2b2 [14].

Clinical terminologies have been historically used in medicine to classify and categorize diseases. One of the most relevant terminologies is SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) [15]. SNOMED-CT is a general purpose clinical vocabulary distributed by The International Health Terminology Standards Development Organization (IHTSDO), with over 400 thousands concepts, 1 million of descriptors and more than 1 million of relationships between them. While SNOMED-CT provides broad coverage, there are other terminologies oriented to more specific clinical areas. Logical Observation Identifiers Names and Codes (LOINC) [16], developed by the Regenstrief Institute in Indiana, USA, is a clinical terminology for identifying laboratory and clinical test results.

In the context of breast cancer research, recent studies show that more than 5% of breast cancer patients might be hereditary [17], caused by gene information inherited from their families' relatives. "All-purpose" terminologies such as SNOMED-CT frequently do not provide the highest coverage for this specific domain. In this area, terminologies such as the HUGO Gene Nomenclature Committee (HGNC) [18] contain only genetic concepts. HUGO is an international classification of the human gene nomenclature, and an open access database containing more than 33,000 gene names and symbols at the time of writing. The majority of these items are protein-coding genes, but they also contain pseudogenes, non-coding RNAs, phenotypes and genomic features.

With an increasing focus on genomics, in last years the number of translational biomedicine solutions has significantly increased. Different approaches intend to exploit the availability of omic data correlated with clinical data to enhance prevention, diagnosis, and therapies [19] [20]. Standardization initiatives in biomedicine such as tranSMART [14], HL7 in standard v3 [21], HL7 FHIR [22] and CDISC [23] are actively working in translational biomedicine. I.e. CDISC has delivered the Study Data Tabulation Model (STDM) [24] for representing the clinical domain; CDISC also propose an implementation guide for pharmacogenomics and pharmacogenetics (STDMIG-PGx) [25], defining relations of biospecimen and genetics-related data. Research projects such as the cancer translational research informatics platform (caTrip) [26] or BioShare [27], have proposed the exploitation of BioBank data together with electronic health records (EHR) data on breast cancer, providing insights on the viability of implementing translational platforms. Electronic Medical Records and Genomics Network (eMERGE) has been created to plan the integration of genomic data into the next-generation of EHRs [28] in a project from 2007 to 2019. eMERGE consider three different strategies to genomic and clinical data. One approach is to store each laboratory genetic result into the EHR system, introducing significant storage requirements for multiple tests looking at a broad range of polymorphisms. The second approach is to generate interpretation of the genomic information at a single point and store it in the EHR assuming the degree of information loss vs. performance improvement. And the third one is linking the original data with an external genetic resource through the EHR system without any loss of genetic information.

We describe our proposed Semantic Interoperability Layer (SIL), and selected examples of its applications within international research projects: EURECA (Enabling information re-use by linking clinical Research and Care) [29] and CIMED (Collaborative Project on Medical Informatics). The objective is to investigate if such standards-based approach can be used to integrate all the genomic information support (similar to eMERGE Project) for the analysis of its interactions in breast cancer studies and diagnostic classifier analysis. This Semantic Interoperability Layer uses standard terminologies as a vehicle for addressing two main challenges in multi-centric interoperability: harmonizing heterogeneities from different data sources as well as for integrating omic and clinical data.

2. Materials and methods

To homogenize common information across different clinical settings, such as clinical trial management (CTMS) systems, electronic health records (EHR), laboratory information management systems (LIMS) and others, in this work we propose a standard-based SIL including one common information model (CIM) and a set of services as homogenous endpoints to access data. As shown in Fig. 1, the proposed SIL is defined by the interaction between the CIM and services for data access. The CIM is composed of three main components: (i) the common data model (CDM), (ii) the core dataset (terminologies) and (iii) the linking between them (terminology binding). The SIL was designed as the basis for software services and tools developed within the project, which are focused on enhancing clinical research with genetic information.

To analyze the interaction of breast cancer gene expressions with clinical data, a set of services for data retrieval were defined within the SIL. These services provided uniform access to data stored in the SIL, exploiting semantic and abstraction capabilities of the CIM. The core dataset integrates terminologies such as SNOMED-CT, HGNC and LOINC for covering the clinical scenario domain [30]. The CDM is a HL7 RIM-based structure required to homogenize data models of information systems from different institutions. Finally, a binding solution for linking the concepts from clinical terminologies to the corresponding CDM

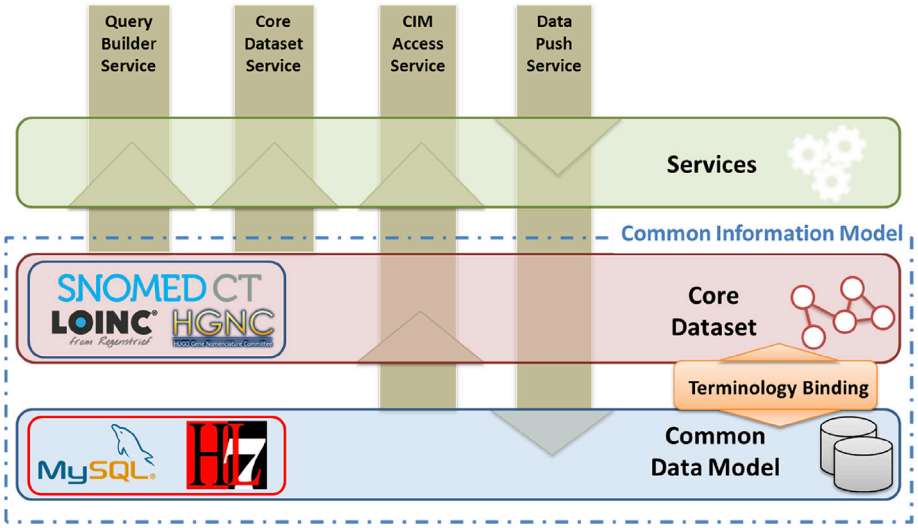

Fig. 1. Interaction diagram of SIL components.

Download English Version:

https://daneshyari.com/en/article/4964810

Download Persian Version:

https://daneshyari.com/article/4964810

Daneshyari.com