International Journal of Medical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

Archetype-based data warehouse environment to enable the reuse of electronic health record data

Luis Marco-Ruiz[a,b,*], David Moner[c], José A. Maldonado[c,d], Nils Kolstrup[a,e], Johan G. Bellika[a,b]

[a] Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway
[b] Department of Clinical Medicine, Faculty of Health Sciences, UIT The Arctic University of Norway, Norway
[c] Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas, Universitat Politècnica de València, Valencia, Spain
[d] VeraTech for Health SL, Valencia, Spain
[e] General Practice Research Unit, UIT The Arctic University of Norway, Norway

ARTICLE INFO

Article history:
Received 24 November 2014
Received in revised form 26 May 2015
Accepted 28 May 2015
Available online xxx

Keywords:
Data reuse
Semantic interoperability
Electronic health record
openEHR
Data warehouse

ABSTRACT

Background: The reuse of data captured during health care delivery is essential to satisfy the demands of clinical research and clinical decision support systems. A main barrier for the reuse is the existence of legacy formats of data and the high granularity of it when stored in an electronic health record (EHR) system. Thus, we need mechanisms to standardize, aggregate, and query data concealed in the EHRs, to allow their reuse whenever they are needed.
Objective: To create a data warehouse infrastructure using archetype-based technologies, standards and query languages to enable the interoperability needed for data reuse.
Materials and methods: The work presented makes use of best of breed archetype-based data transformation and storage technologies to create a workflow for the modeling, extraction, transformation and load of EHR proprietary data into standardized data repositories. We converted legacy data and performed patient-centered aggregations via archetype-based transformations. Later, specific purpose aggregations were performed at a query level for particular use cases.
Results: Laboratory test results of a population of 230,000 patients belonging to Troms and Finnmark counties in Norway requested between January 2013 and November 2014 have been standardized. Test records normalization has been performed by defining transformation and aggregation functions between the laboratory records and an archetype. These mappings were used to automatically generate open EHR compliant data. These data were loaded into an archetype-based data warehouse. Once loaded, we defined indicators linked to the data in the warehouse to monitor test activity of Salmonella and Pertussis using the archetype query language.
Discussion: Archetype-based standards and technologies can be used to create a data warehouse environment that enables data from EHR systems to be reused in clinical research and decision support systems. With this approach, existing EHR data becomes available in a standardized and interoperable format, thus opening a world of possibilities toward semantic or concept-based reuse, query and communication of clinical data.

1. Background

The reuse of data inside the electronic health record (EHR) is considered of paramount importance to support clinical research [1,2] and clinical decision support (CDS) [2,3] systems. This reuse needs clinical data to flow from the systems where it was captured to a specialized system that processes it with different objectives, other than clinical care. Agile aggregation during this flow of information can enable the creation of large repositories containing samples to perform inferences over populations [4]. Many different actors like physicians, hospitals, pharmaceutical and biotech companies can benefit from this data flow to deliver their services more effectively [5].

* Corresponding author at: Norwegian Centre for Integrated Care and Telemedicine, University hospital of North Norway, P.O. Box 35, N-9038 Tromsø, Norway. Tel.: +47 41441113.
E-mail address: Luis.Marco.Ruiz@telemed.no (L. Marco-Ruiz).

http://dx.doi.org/10.1016/j.ijmedinf.2015.05.016
1386-5056/

Mechanisms to improve clinical accessibility, better integration of care across settings and advances in information exchange are needed to reuse data effectively [6]. Several initiatives have contributed with advances in those directions [7–9]. However, challenges in semantic interoperability and technical infrastructure are still present [10].

1.1. Semantic interoperability

On the semantic interoperability side the challenges are: (a) the integration and harmonization of the formats among different sources [2]; and (b) the definition of shared clinical information models, and their terminological binding. To harmonize and integrate EHR data, the adoption of EHR standards such as HL7CDA [11], openEHR [12] or EN ISO 13606 [13] is needed to exchange health information extracts. The use of such standards provides a common syntax and representation of clinical data. In addition, the definition of clinical information models that represent domain-oriented data structures allows the implementation of solutions for specific clinical use-cases. Finally, the terminology binding of clinical data and clinical information models provides semantics to the standardized information. The use of terminologies such as SNOMED-CT, LOINC or ICD delivers a common vocabulary used to specify the exact meaning of clinical data despite cultural or language differences. These actions are recognized as essential steps towards the semantic interoperability of health information [14]. In this sense, the memorandum of understanding signed between the Department of Health and Human Services and the European Commission stated the immediate importance in adopting international standards and interoperability specifications for EHRs. Also, the definition of harmonized clinical information models is of importance to achieve semantic interoperability across health information systems [15]. Internationally, the clinical information modelling initiative (CIMI) [16] is a not-for-profit community of users and stakeholders working in the definition of shared clinical information models. At the European level, the epSOS project [17] has defined the minimal infrastructure and clinical information models to communicate the patient summary, prescription and dispensation documents across Europe using the HL7 CDA standard. National initiatives like the NHS interoperability toolkit [18] or the meaningful use regulations [19,20] are powering the adoption of standards providing interoperability specifications and establishing payment for health professionals who adopt certified EHRs [4].

1.2. Linkage of the EHR with inference models

On the infrastructure side, the "impedance mismatch" [21] present between the information model, (where EHRs lay) and the inference model (where CDS systems and data mining are located) [22] can jeopardize clinical data reuse. The reason is that clinical information services are not structured to support ad hoc queries to allow reuse; therefore the use of data warehouses (DW), registers and repositories is needed [23]. To overcome the mismatch between the EHR and inference models several approaches have been proposed with application both in CDS and clinical research. The Arden syntax [24] was the first standard to create independent medical logic modules (MLM). Its syntax separated knowledge expressions from data access to EHR. However, it encapsulated data access between curly braces inside the MLM leading to the "curly braces problem". The former can be explained as the necessity to adapt the data access sections when moving the system among different production environments. GELLO [25] advanced in the direction of decoupling the EHR data access from the CDS system allowing to work with object models. Other computer interpretable guidelines formalisms like SAGE or GLIF used a summary view of the EHR, a.k.a. virtual medical record (VMR) using HL7 RIM [26]. Peleg et al. [27], following the former approach, defined a RIM VMR but added a mapping ontology to transform EHR granular data from the VMR into highly aggregated concepts for CDS [27]. Kawamoto et al. contributed by identifying the features needed to create a CDS VMR specific standard [28] to solve some limitations of the HL7 RIM based views. With regard to archetypes, Marcos et al. [29] and Fernandez-Breis et al. [30] relied on archetypes to generate an EHR view. Over that view they used additional layers of archetypes to increase the level of abstraction to generate the concepts needed by a patient cohort identification system for clinical research. Hybrid approaches to take advantage of the best feature of each standard have also been proposed. For example, in MobiGuide [31], Gonzalez-Ferrer et al. [32] propose to use HL7 RIM for backend systems and define a layer of archetypes compliant with the HL7 VMR for integration with the CDS system [32]. More oriented to the data reuse in research the SHARPn consortium has defined a complete pipeline to extract data from the EHR, normalize it into clinical element models and allow its reuse through health quality measures format (HQMF) [33].

1.3. Data reuse for CDS in North Norway primary care

Several studies have documented the benefits of CDS systems in primary care [3,34]. When applied to laboratory interventions, CDS systems have been documented to reduce costs avoiding redundant tests [35]. However, population information for general practitioners (GPs) is usually limited by the patients they are assigned and their personal communications with colleagues. They seldom have access to real time population test results or colleagues requests. Access to anonymized and aggregated population data about laboratory interventions of other colleagues and laboratory personnel can empower their environmental awareness of communicable infectious diseases and help them to determine which set of tests should be ordered when suspecting certain conditions.

In the North Norway health region tests are ordered by a GP who usually sends a request with a sample to perform several tests to confirm or discard a set of infectious agents to the microbiology service. At the microbiology service, the staff can add new tests for other infectious agents to the ordered request based on their knowledge; e.g., if the beginning of an epidemic is suspected, an outbreak of a communicable disease etc.

Conceptually, a request is a composition of some demographical data related to the patient and requester (GP/microbiology service) which contains a battery of individual tests, each aimed to detect an infectious agent. Test results are stored in the laboratory information system (LIS) in a proprietary format. Currently, reports of the tests performed in a geographical area in a period of time are generated using the SNOW system [36,37]. The system uses an agent-based architecture to extract the data related to tests and aggregate it per municipality and date to monitor the number of confirmed cases of several diseases. It applies transformation and aggregation rules to generate derived data from the individual results. For example, the infectious agent and its category (common cold viruses, gastrointestinal bacteria etc.) are derived from the type of test (DNA, culture, enzyme immunoassay etc). The aggregation of tests by request is done by defining grouping rules using the metadata of each result related to fields as patient id, request data, requester identifier etc.

In the case of transformation rules, the SNOW system implements a total of 25 rules to derive the values of some fields from other fields. These rules are implemented using the Drools business rule management system [38] as result of collaboration with the laboratory personnel. As no standard terminologies such as LOINC are implemented, such rules are based on the internal codes and names. As an example, the rule displayed in Fig. 1 is used to infer the infectious agent, the family of symptoms and the subcategory

Please cite this article in press as: L. Marco-Ruiz, et al., Archetype-based data warehouse environment to enable the reuse of electronic health record data, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.05.016