Clinical Data Models at University Hospitals of Geneva

Vishnyakova, Dina; Gaudet-Blavignac, Christophe; Baumann, Philippe; Lovis, Christian

How to cite

VISHNYAKOVA, Dina et al. Clinical Data Models at University Hospitals of Geneva. In: Studies in health technology and informatics, 2016, vol. 221, p. 97–101. doi: 10.3233/978-1-61499-633-0-97

This publication URL: https://archive-ouverte.unige.ch/unige:88481
Publication DOI: 10.3233/978-1-61499-633-0-97

Transforming Healthcare with the Internet of Things
J. Hofdijk et al. (Eds.)

doi:10.3233/978-1-61499-633-0-97

Clinical Data Models at University Hospitals of Geneva

Dina VISHNYAKOVA [a,1], Christophe GAUDET-BLAVIGNAC [a], Philippe BAUMANN [a], and Christian LOVIS [a,b]
[a] *Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland*
[b] *University of Geneva, Geneva, Switzerland*

Abstract. In order to reuse data for clinical research it is then necessary to overcome two main challenges – to formalize data sources and to increase the portability. Once the challenge is resolved, it then will allow research applications to reuse clinical data. In this paper, three data models such as entity-attribute-value, ontological and data-driven are described. Their further implementation at University Hospitals of Geneva (HUG) in the data integration methodologies for operational healthcare data sources of the European projects such as DebugIT and EHR4CR and national project the Swiss Transplant Cohort Study are explained. In these methodologies the clinical data are either aligned according to standardised terminologies using different processing techniques or transformed and loaded directly to data models. Then these models are compared and discussed based on the quality criteria. The comparison shows that the described data models are strongly dependent on the objectives of the projects.

Keywords. Clinical data model, Data Integration, Data Ontology, EHR

Introduction

The research of eHealth usually faces the problem of data reuse and consequently the semantic data interoperability. Solution of this problem is essential for the sustainable use of information. Electronic health records (EHR) are the growing part of the eHealth, ranging from clinical findings to genome structures where continuous electronic processes improve coordination and rapid exchange of information among stakeholders. However, the up cycling (secondary utilisation of data) of clinical data to improve healthcare quality and patient safety are very limited. Therefore, the need to integrate heterogeneous data from multiple sources and sharing information in a distributed and collaborative environment are highly challenging. The data integration process for the research environment is a complex task, which has to take into account the following questions: 1) How the system will be used? 2) Who will use the system?

The data model could influence the research facilities starting from defining what kind of data can be stored to how the information will be queried and extracted. To date, there are some models, which became widely used in research domains, for instance OMOP Data Model of OHDSI [1], archetypes of openEHR [2] and the dimensional modelling of i2b2. Since there is no solution to identify the generic data model able to

[1] Corresponding Author.

be efficient for any research need we propose various data models selected for the projects conducted at University Hospitals of Geneva: 1) the cohort project aims to acquire high quality data about patients with chronic diseases. The Swiss Transplant Cohort Study (STCS), launched in 2008, was the first module implemented in this project. This module is daily used in Switzerland to follow the complete population of transplanted patients since 2008. Other active cohorts have since joined the project to allow follow-up and statistical studies over more than 5000 patients; 2) In a context where the emergence and increase of antimicrobial resistance is problematic, the European FP7 DebugIT project [3], [4] aimed to improve and monitor prescriptions of antibiotics and thus reduce antimicrobial resistance; 3) EHR4CR - provides adaptable, reusable and scalable solutions for data reuse systems of EHRs for clinical research. The project addresses four main scenarios: the feasibility of clinical protocols, identification and patient recruitment, execution of clinical trials and reporting of side effects [5].

We should notice that this paper doesn't aim the description of the legal dimension of medical data reuse.

1. Methods

We present the data integration methodologies implemented by University Hospitals of Geneva (HUG), which promote technical and semantic interoperability for operational healthcare data sources. The selection of the data model is based on the needs of the projects. We present 3 different data models: 1) Entity-attribute-value evolved from the legacy system 2) ontology-driven data model which is built from scratch in order to extract antibiotic-relevant information and 3) data-driven model, the existing data model (i2b2 star schema) is adapted to the needs of the project.

1.1. Entity-attribute-value model

The STCS project and as well as the 4 other cohort projects of HUG are based on the entity-attribute-value (EAV) information model. Since at the beginning it was an institutional project the information model of the project is based on the legacy clinical system. Every row in EAV model is composed of three fields: 1) an entity representing a described item (a consultation with a patient); 2) an attribute describing the entity (e.g. cardiac frequency) and 3) the value of the attribute (e.g. 64 beats/min), which can be of any type. For each cohort project an ad-hoc web application was developed to allow an interaction with the clinical data. This interaction is based on the object-relational mapping tool. The data-model is patient oriented. The semantics of data is based on the terms, specially developed by clinicians to match the needs of the project. Since the number of patients per institution was not large, the population of data model is done manually by data-managers of the defined medical centres.

The data is usually extracted from EAV model on demand and sent to a central-data-manager, it is then cleaned and processed to be easily usable for analysis. There are currently more than 80 research projects that have passed the ethical and scientific committees and use these data for research.

1.2. Ontological model

The DebugIT project performed a bottom-up approach for the data integration. The data model is done from scratch. Initially, clinical data from different hospitals were collected and organized in virtualized clinical data directories. Thus, each pilot site has a relational database fulfilled by information on microbiology, medicaments and patient administration. This information was standardised by terminologies such as NEWT and WHO-ATC. The goal of a pilot site – to formalise the data so that it can be ubiquitously accessed using a formal query language. The queries constructed in a manner to answer a specific question related to antibiotic resistance, e.g. "What is the evolution of bacteria resistance to antibiotic during period at location?" For this purpose, the underlying database content is transformed into a formal language representation. This is achieved by defining the elements, their classes, properties, instances and relationships, using a formal ontology language [6]. In this case, the formalized data model (FDM) is a direct map of the original database schema to an ontological model. The FDM vocabulary directly reflects the table and column names of the source model. The FDM is then connected to the underlying non-formal database so that it can provide data in the RDF format and be accessed through a SPARQL query protocol.

1.3. Data-driven model

In the context of EHR4CR project, the HUG as a pilot site has chosen a model-driven approach for data integration. The pilot sites agreed to expose their data in a form of Clinical Data Warehouse (CDW). HUG's CDW is based on the platform of Informatics for Integrating Biology and the Bedside (i2b2) [7]. Since the scenarios of the project are patient-centric. The choice of i2b2 data model is rather rational, it fits the patient-centric scenarios and is represented as a five-axis star: Patient, Visit, Observation, Concept and Provider [7]. The clinical data of HUG corresponding to such axis as Patient, Visit and Provider were integrated directly to the i2b2 schema. Since clinical data of laboratory analysis (Labs) is not using the international standardization code such as LOINC we have created local concepts of Labs for the CONCEPT dimension. The semantic interoperability with the research environment is implemented outside of the data model. In order to federate all heterogeneous clinical data sources the project had addressed some semantic interoperability aspects and the clinical data storage model through: 1) the terminology mapping service for dynamical translation of concepts of the central to local terminologies [8]; 2) the query system developed to transform eligibility criteria into queries that interrogate heterogeneous local data warehouses [9] .

2. Results

In order to compare the three different data models we have used the data quality criteria defined in [10], see Table 1. In this table the criteria defined in [10], is adapted for the clinical data models as following: 1) completeness – does the data coverage fit the project needs ; 2) integration – does the model link all data dimensions (Patient, observation, visit, laboratory analysis) correctly? ; 3) understandability – do the data structure and concept make sense to all end users (data investigators, data managers, clinicians, etc)? ; 4) simplicity – is it easy to transform data elements to the model? 5) flexibility – is it

possible to extend the project scope within the data model (e.g. to add new elements without changing the schema)?

Table 1. Evaluation of data models according to quality dimensions

Dimension	Ontology-driven	Data-driven	EAV
Completeness	yes	yes	yes
Integration	yes	yes	yes
Understandability	partly	yes	yes
Simplicity	no	yes	no
Flexibility	no	no	yes

Since the models were maintained in the framework of the projects we have excluded the criterion as *implementability* from the list of dimensions. This criterion depends on some parameters such as time, cost and technical facilities which vary from project to project.

3. Discussion and Conclusions

Since medical science is constantly evolving, the need to modify the database is permanent and the capacity to update the database without affecting the general model is important. Flexibility or extensibility is a major advantage for the data model. For instance the EAV model by adding new fields in the database doesn't require changes in schema. Adding new rows in a specific table can easily do it. But, the disadvantage of this model concerns mostly performance issues (e.g. time of performance, query complexity).

The data coverage of research questions together with data integration are the major parameters for model selection. Thus, the achieved results of the ontology-driven model [11-14] showed the adequacy of the semantic integration methods developed by the project consortium. One of the main benefits demonstrated by this approach is increased portability of the semantically formalized data sources. Though the ontological data integration was achieved, local semantic formalization was not fully interoperable. In the ontology-based integration system, the automatic mapping from global to local ontologies using first-order logic hinders logical consistency [15]. Consequently, various local ontologies were not completely resolved in the global model. However, regarding the flexibility criterion, in DebugIT, query templates must be defined centrally for each new data source.

The choice of the data model based on i2b2 [16-19] for the EHR4CR environment is rather rational [20]. The data model of i2b2 was easily adapted to EHR4CR needs since the patient-centric model matched the patient-centric scenarios of the project. The understandability made an i2b2 to be an admired tool for data integration and querying by clinical users. The semantic interoperability was achieved through external EHR4CR software such as terminology services. It is also worth to notice one of the differences between the DebugIT and EHR4CR environments is that in the latter one the query templates are defined locally. Moreover, the clinical data warehouse based on i2b2 data model relies on the relational database mechanism where the query construction to access data is less complex than for the ontology-driven model.

The data integration approaches such as ontology-driven and data-driven are able to homogenize the distinct data sources. In the EAV model due to the specificity of the project, the homogenization is done on demand.

Conclusively, the described projects have different goals and to define a suitable data integration model which would fit the needs of every project is not realistic.

References

[1] Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. Journal of the American Medical Informatics Association. 2015:ocu023.

[2] Garde S, Knaup P, Hovenga EJ, Heard S. Towards Semantic Interoperability for Electronic Health Records--Domain Knowledge Governance for open EHR Archetypes. Methods of information in medicine. 2007;46(3):332-43.

[3] Lovis C, Colaert D, Stroetmann VN. DebugIT for patient safety-improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Studies in health technology and informatics. 2008;136:641.

[4] Schulz S, Karlsson D. Records and situations. Integrating contextual aspects in clinical ontologies. Bio-Ontologies 2011. 2011.

[5] De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: The case of the EHR4CR project. Journal of biomedical informatics. 2014.

[6] Teodoro DH, Choquet R, Schober D, Mels G, Pasche E, Ruch P, et al. Interoperability driven integration of biomedical data sources. Studies in health technology and informatics. 2011;169:185-9.

[7] Murphy S, Gainer V. Working Assumptions of i2b2 Data.

[8] Ouagne D, Hussain S, Sadou E, Jaulent M-C, Daniel C. The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. Studies in health technology and informatics. 2012(180):534-8.

[9] Bache R, Taweel A, Miles S, Delaney B. An Eligibility Criteria Query Language for Heterogeneous Data Warehouses. Methods Inf Med. 2015;54(1):41-4.

[10] Moody DL, Shanks GG. Improving the quality of data models: empirical validation of a quality management framework. Information systems. 2003;28(6):619-50.

[11] Daniel C, Choquet R, Assele A, Enders F, Daumke P, Jaulent M-C, editors. Comparing the DebugIT dashboards to national surveillance systems. BMC Proceedings; 2011: BioMed Central Ltd.

[12] Schobera D, Choquetb R, Depraeterec K, Endersd F, Daumked P, Jaulentb M-C, et al. DebugIT: Ontology-mediated layered Data Integration for real-time Antibiotics Resistance Surveillance.

[13] Pasche E, Ruch P, Teodoro D, Huttner A, Harbarth S, Gobeill J, et al. Assisted knowledge discovery for the maintenance of clinical guidelines. PloS one. 2013;8(4):e62874.

[14] Choquet R, Daniel C, Grohs P, Douali N, Jaulent M-C, editors. Monitoring the emergence of antibiotic resistance using the technology ot the DebugIT platform in the HEGP context. BMC Proceedings; 2011: BioMed Central Ltd.

[15] Schulz S, Stenzhorn H, Boeker M, Smith B. Strengths and limitations of formal ontologies in the biomedical domain. Revista electronica de comunicacao, informacao & inovacao em saude: RECIIS. 2009;3(1):31.

[16] Boussadi A, Caruba T, Zapletal E, Sabatier B, Durieux P, Degoulet P. A clinical data warehouse-based process for refining medication orders alerts. Journal of the American Medical Informatics Association. 2012:amiajnl-2012-000850.

[17] Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent J-F, Garin E, et al., editors. Roogle: an information retrieval engine for clinical data warehouse. MIE; 2011.

[18] Mate S, Bürkle T, Köpcke F, Breil B, Wullich B, Dugas M, et al. Populating the i2b2 database with heterogeneous EMR data: a semantic network approach. Studies in health technology and informatics. 2010;169:502-6.

[19] Segagni D, Gabetta M, Tibollo V, Zambelli A, Priori SG, Bellazzi R, editors. ONCO-i2b2: improve patients selection through case-based information retrieval techniques. Data Integration in the Life Sciences; 2012: Springer.

[20] Vishnyakova D, Bottone S, Pasche E, Lovis C. Practical Implementation of a Bridge between Legacy EHR System and a Clinical Research Environment. Studies in health technology and informatics. 2014;197:29-33.