e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-1060

A Framework for Integrating Heterogeneous Clinical Data for a Disease Area into a Central Data Warehouse

Christian KARMEN[a,1],Matthias GANZINGER[a], Christian D. KOHL[a], Daniel FIRNKORN[a], Petra KNAUP-GREGORI[a]
[a]*Institute of Medical Biometry and Informatics, Heidelberg University, Germany*

Abstract. Structured collection of clinical facts is a common approach in clinical research. Especially in the analysis of rare diseases it is often necessary to aggregate study data from several sites in order to achieve a statistically significant cohort size. In this paper we describe a framework how to approach an integration of heterogeneous clinical data into a central register. This enables site-spanning queries for the occurrence of specific clinical facts and thus supports clinical research. The framework consists of three sequential steps, starting from a formal data harmonization process, to the data transformation methods and finally the integration into a proper data warehouse. We implemented reusable software templates that are based on our best practices in several projects in integrating heterogeneous clinical data. Our methods potentially increase the efficiency and quality for future data integration projects by reducing the implementation effort as well as the project management effort by usage of our approaches as a guideline.

Keywords. Data Warehouse,Data Integration,Data Harmonization, Clinical Research

Introduction

Nowadays, the usage of digital data acquisition systems in health care is increasingly growing. In the United States, the percentage of office-based physicians with Electronic Medical Records (EMR) systems reached a peak in 2012 with estimated 72 % [1]. Furthermore, open source electronic data capture systems, such as *REDCap* [2] are used to collect data in clinical trials. Both types of data, patient records and data from trials, represent facts that can be used for research purposes beyond clinical trials. It is assumable that the structured collection of relevant clinical data for a specific disease area has the potential to improve research efficiency in its specific field.

In order to have a significant number of cases to analyze a specific research question, it is often necessary to accumulate patient data from several institutions. Each hospital might focus on different aspects of clinical facts, having a varying data granularity. They might furthermore use varying data structures and storage methods which results in a highly heterogeneous collection of clinical data lacking semantic interoperability. There also might be connectivity constraints or privacy concerns in

[1]Corresponding Author.

multi-centered databases. Thus, a central registry, e.g. in the form of a dedicated data warehouse (DW) is required to integrate data from several sites. However, the necessary integration into a unified structure, which can be accessed efficiently is challenging. The complexity of these tasks often leads to individual solutions for each data integration project. In order to define a general approach with the help of our best practice methods, we developed a framework to use as a guideline for a threefold approach: (1) harmonizing digitally recorded medical data from multiple sites to a harmonized target structure; (2) implementing a transformation from the clinical source data to the target structure; (3) setting up a central data warehouse system that is suitable for the storage of clinical facts.

1. Methods

We established integrated data platforms in three different disease areas and in three different national and international research networks:

- EURenOmics, international research consortium for rare kidney diseases [3]
- German Center for Lung Research (DZL) that focus on prevention, diagnosis and therapy of serious lung diseases [4]
- CLIOMMICS doing advanced research on system medicine for multiple myeloma with the help of *omics* based data [5].

Continuous improvement of our methods in analyzing, processing and implementation finally resulted in this framework. We divided it into three sequential parts and developed methods and tools for each part.

1.1. Harmonization of multi-centered clinical data

Each site that is collecting clinical data about a specific disease area might have a different purpose why data are captured. Often the amount and granularity of the acquired data, the laboratory focus as well as the research question might be completely different. Because of this diversity we need to *harmonize* all collected data from all locations. A detailed analysis of each data source is necessary to identify the meaning, data structure, and granularity for each clinical fact. Then the domain experts need to come to an agreement on *what* data are relevant to answer the specific research question and *how* they are organized. As obvious as this might sound, it is the most critical step in the whole process. As a result, a list of research relevant items, eventually grouped by elementary categories, should be available. This list, from now on denoted as ontology, is supposed to make sense to a domain expert. Finally, the ontology needs to be mapped with both, the documented syntax as well as the semantics of each individual data set.

This approach is supposed to be used as a process of continuous improvement. This means that the initial version of the ontology is allowed to have a low granularity whereas future versions are evolving continuously since the domain experts will come to agreements that affect the quantity and quality of the ontology.

The process of data harmonization usually involves many experts, mostly physicians and biometricians, but also study nurses as well as a medical informatician.

The latter focuses on the technical practicability while the discussion of the central ontology structure is ongoing. This target data structure might include sections such as

- Patient's demographics
- Laboratory results
- Common clinical data / health values
- Family history
- Treatment data (prescription / acceptance)

1.2. Transformation from Clinical Data Sources into the Target Structure

After a careful identification of the relevant clinical data, each clinical site needs to run through a process of three steps: (1) extract its clinical data from its site specific source; (2) transform its data into the format of the target ontology structure; and (3) load the results into the DW. For step (3) we act on the assumption that a DW is used as a common central data repository. This process is described as Extract-Transform-Load (ETL) and will be applied with an advanced open source tool. Within this tool we developed advanced and highly flexible methods enabling us to accelerate the initial ETL implementation.

1.3. Setting up the Data Warehouse

The general purpose of a DW is to collect data from several locations and in different formats in order to analyze its data as fast, convenient, and extensive as possible. Further helpful requirements are a data export function for user-defined queries on specific clinical facts as well as high usability on both, installation and the user-interface. For economic reasons an open source solution is preferred.

2. Results

2.1. A spreadsheet for managing data harmonization

As a helpful utility we created a comprehensive data harmonization table in Excel in order to provide a semantic matching between the several data sources and the target structure. In this table the previously mentioned target ontology marks the main structure and is extended with fields for the data type, e.g. String or Integer, a short description and, if appropriate, the accepted range of values. A *comments* field is also recommended. In case of ontology elements that have a limited number of values, e.g. categories or bivalent statements, it is important to state each of them in a separate row. This way, an individual handling for each element can be provided. This table can be used independently of a disease area.

Now, the harmonization table is expanded with the data sources by adding one data column for each participating hospital. Each data row, containing one specific ontology element, must be analyzed by the respective participant to identify the level of compatibility to the target ontology. There a three levels to differentiate: (1) complete, (2) sufficient and (3) no coverage. While level 1 compatibility can be easily mapped to the target ontology and level 3 is a lack of data, the most challenging part is sufficient

data coverage (2). Here, the data is basically available but cannot be matched directly and thus, some kind of data transformation is necessary. This might involve a recalculation of numerical values, e.g. for unit conversions, re-interpretation of area-specific date formats or a value re-interpretation due to a coding mechanism.

In General, the alignment of clinical data from each hospital site to the target ontology must be validated completely manually by one or more local domain experts. An automatic alignment is hard to realize and thus not recommended because of two reasons: (1) a computational semantic analysis of the meta data description cannot guarantee an 100 per cent match and (2) the context and purpose of the data records might be in a completely different focus and thus, might lead to misleading conclusions.

2.2. Talend Open Studio as a flexible open source ETL tool

The task to transform the source into the target data structure might be challenging. Therefore, there is a need for a powerful ETL tool. We created a well-structured and documented workflow in the sophisticated open-source ETL tool Talend Open Studio [6]. In order to enhance reutilization we implemented templates for the ETL tool in a way that it is capable to use a set of Excel files to describe different aspects of the target ontology, like descriptions for data types, dynamic and static categories and black-listed items.

This tool reduces massively the effort to implement filters for a specific hospital and is highly optimized on the database structure of the selected DW. Once the filters are set, the ETL tool can generate a stand-alone program to run the data import anywhere, without installing the complete ETL tool. This is especially useful for a physician with missing technical abilities.

2.3. An open source data warehouse as the central clinical data repository

The last element in the process chain of our framework is the central data repository for the clinical facts corresponding to the ontology. Our requirements, mentioned in the section Methods, were satisfied best by i2b2 [7], an advanced open source software framework for clinical researchers that provides a database structure as well as a web based interface especially for integration of clinical facts.

We took advantage of the ontology editor provided by an i2b2 plug-in to describe the target ontology structure. The i2b2 tailor-suited ETL jobs (section 2.2) are not only providing the correct transforming from the source into target data formats, but are also linking the clinical fact correctly to the corresponding ontology elements, based on a concept coding system provided by i2b2.

We also benefit from the well-designed usability. I2b2 provides a graphical user interface (GUI) that allows ontology queries by dragging and dropping clinical facts from the ontology tree into the filter widgets. Finally, our last requirement is fulfilled by a plug-in that enables the export of the clinical facts based on user-defined queries.

3. Discussion

Integrating heterogeneous clinical data into a central data repository is considered a necessary step for clinical research. We developed a best practice framework that breaks the complexity down into three consecutive basic steps consisting of (1) creating a harmonization table, (2) setting up an ETL process and finally (3) putting the resulting data structure into a central repository that enables custom queries. Furthermore, we provided spreadsheets and ETL templates [8] to support an individual implementation, based on the software tools of our choice. This may decrease the work load and improves the understanding of the complexity behind data integration.

The Integrated Data Repository Toolkit (IDRT) [9] project has intersections with our methods, but lacks in support of early steps of data harmonization. We focused specifically on the individual process of integrating heterogeneous data from multi-centered sites.

The framework we introduced in this paper still has potential for improvements in several directions. Data privacy is a critical element in each patient data based analysis. In our work we assumed that this issue has already been taken care of. This may not apply in all cases; therefore the framework can be enhanced by privacy issues, like patient's consents and pseudo nomination.

The implemented ETL solutions we described are depending on i2b2. Another solution might be more appropriate, e.g. when professional support or extended features of a DW are needed. An adaption of the ETL implementation to common commercial systems is possible.

References

[1] Hsiao C, Hing E. Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001-2012. NCHS Data Brief 2012; (111):1–8

[2] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics 2009; 42(2):377–81. Available from: URL:http://www.sciencedirect.com/science/article/pii/S1532046408001226

[3] European Union 7th Framework Programme. EURenOmics [cited 2014 Jan 30]. Available from: URL:http://www.eurenomics.eu

[4] Deutsches Zentrum für Lungenforschung. DZL: The German Center for Lung Research [cited 2014 Jan 30]. Available from: URL:http://www.dzl.de

[5] Bundesministerium für Bildung und Forschung (BMBF). e:Med: Maßnahmen zur Etablierung der Systemmedizin [cited 2014 Jan 30]. Available from: URL:http://www.gesundheitsforschung-bmbf.de/de/5111.php

[6] Talend Open Studio: Talend Inc.; 2014. Available from: URL:http://www.talend.com

[7] Partners HealthCare System in Boston, Mass. i2b2: Informatics for Integrating Biology & the Bedside [cited 2014 Jan 30]. Available from: URL:http://www.i2b2.org

[8] Institute of Medical Biometry and Informatics, University of Heidelberg. Integration of Heterogeneous Phenotype Data [cited 2014 Feb 4]. Available from: URL:http://www.klinikum.uni-heidelberg.de/index.php?id=136562&L=1

[9] Goltz U, Magnor MA, Appelrath H, Matthies HK, Balke, Wolf-Tilo & Wolf, Lars C., editors. Integrated Data Repository Toolkit: Werkzeuge zur Nachnutzung medizinischer Daten für die Forschung: GI; 2012. (LNI)