User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-857

A Metadata-Based Patient Register for Cooperative Clinical Research: A Case Study in Acute Myeloid Leukemia

Anja S. FISCHER[a,1], Ulrich MANSMANN[a]
[a] *Institute for Medical Informatics, Biometry and Epidemiology (IBE), Ludwig-Maximilians-University Munich, Germany*

Abstract. In many medical indications clinical research is organized within study groups which provide and maintain the clinical infrastructure for their randomized clinical trials. Each group also manages a data center where high quality databases store the study specific individual patient data. Sharing this data between study groups is not straightforward. Therefore, a concept is needed which allows to represent a detailed overview on the information available across the cooperating groups. We propose a metadata based patient register and describe a first prototype. It provides information about available patient data sets to interested research partners while the typical register approach only collects a predefined limited core data set. This register implementation enables cooperative groups to allocate clinical data for future research projects in distributed data sources beyond the restrictions of core data sets. Additionally, it supports the research network in communication and data standardization and complies with a governance structure which is compatible with ethical aspects, privacy protection, and patient rights.

Keywords. metadata, patient register, CDISC ODM, data integration, networked clinical research

1. Introduction

Academic clinical research is organized by study groups which provide and maintain the infrastructure to run large randomized clinical trials. Typically, there a several national or international study groups working on the same medical indication. Each study group also manages a data center which performs the data management for ongoing studies but also manages large databases from completed clinical trials.

Warehouse techniques can be used within those data center to explore relevant clinical information across different studies of the study group. Relevant issues which need well documented patient data are for example: meta analyses, prognostic factor research, biomarker research, subgroup analyses, simulation of future trials, health economic research, or determination of surrogate endpoints. Since those activities are mostly from exploratory character, one also needs extensive data sets to validate findings of interest. Often, data repositories of single study groups are not large enough to manage exploration and validation of specific clinical aspects. This is an incentive to

[1] Corresponding author: Anja S. Fischer, Institute for Medical Informatics, Biometry and Epidemiology (IBE), Marchioninistr. 15, 81377 Muenchen, Germany; E-mail: anja.fischer@ibe.med.uni-muenchen.de. This work was supported by the German José Carreras Leukaemia Foundation (DJCLS H06/04V).

establish an infrastructure for cooperation between academic study groups with clinical research in a specific indication.

Since the research questions in cooperative clinical research are quite broad, it is not helpful to establish a classical patient registry between the cooperating study groups which contains a uniform core data set restricting the question of interest. Whereas many different definitions of patient registers exist [1-6] and various implementations of this concept are found [7, 8], the uniform standardized data set of every patient is common to all of them. Its size can vary; a survey on 14 German disease registers [7] found an average number of about 200 collected items per patient.

Furthermore, it may be problematic to share patient information (even in a pseudomized or anonymized form) between the clinical study groups. Partners may be ready to share project specific data, but may be reluctant to provide extensive patient profiles for a central registry. Partners may be less reluctant to share information on patient information available in their repositories. This can be done by sharing study specific data dictionaries which define the data items of the study and the ways they are measured. Even, it may be easy to disclose which item is measured with good or bad quality for which patient.

Consequently it is required to collect syntactic and semantic meta information about a data item in a specific study. This comprises metadata about the data item representation and stored values as well as contextual metadata concerning the data capture process. Representation and contextual metadata can be elevated from certain study documents (i.e.: Data dictionary as central information on the structure of a specific study database, data validation plan as the document which defines data quality, and a study protocol to explain the logic which sets the variables of a study in their specific logical context. Content metadata (i.e. patient wise availability and quality of item values collected in the study) has to be compiled directly from the study database and must be a updated regularly as the study data collection progresses.

The metadata provides a reliable planning basis for cooperative research projects. It simplifies communication between collaborating partners and supports and accelerates the development process of a feasible common research protocol.

We will present an IT infrastructure based on modern technical components and internationally accepted data standards for extraction, transformation and loading of metadata into a metadata-based patient register.

The developed technical infrastructure has to be implemented into a governance infrastructure which assures data safety, privacy rights, and a transparent cooperative work. We also show that the implementation of the concept allows improving standardization of data management in clinical studies between the cooperating study groups.

With our concept we follow the general principles of caBIG™ [9] of opening and implementing the cross-communication between distributed and federated data sources in oncology. Our approach is a deviation from the fully federated model of caBIG™ by establishing the metadata-based patient register as a central component. It offers a central link to available clinical research data of a patient in the research community.

As a case study we consider a metadata-based patient register for four German study groups on Acute Myeloid Leukemia (AML) which is a rare disease characterized by a high mortality rate [10]. In Germany, investigator-driven multicenter treatment optimization trials are the main instrument in clinical leukemia research [11]. In the course of the trial a broad range of clinical data is collected providing the basis for evidence-based evaluation of the trial objectives. All trials together offer a rich

information basis to perform meta-analyses, sub-group analyses, discovery and validation studies for biomarkers and surrogate endpoints, and diagnostic as well as prognostic rules.

The heterogeneity in clinical documentation in AML studies (i.e. therapies and therapy outcome, concurrent diseases, etc.) is a recurring challenge in cooperative research projects. Therefore this is an interesting and significant field for evaluation of the concept of a metadata-based patient register.

2. Methods

The problem of collating AML clinical data from multiple centers for meta-analysis: The classical patient data registry was discarded because of the severe restrictions implied by a uniform data set. The warehouse concept can not be applied because the partners did not agree on a permanent sharing of full patient data. The metadata based approach offers sufficient flexibility for the design of research projects by maximal protection of the individual patient data.

The design of the processes for collation and for the management of metadata, the approach taken for requirements elicitation: For requirements compilation as well as documentation of available sources of clinical data semi-structured interviews with selected staff of study groups were conducted. The project stakeholders discussed and assessed the approach on a regular basis.

Details of the system design: Discussions, requirement engineering and decision-making were supported by modeling of core data processes with the Business Process Modeling Notation (BPMN), i.e. (1) process of metadata extraction from data source (2) the load process of metadata into the register (3) the extraction and forwarding process of clinical data.

Tools and techniques used for building the system: Various metadata standards (ISO/IEC 11179 [12], CDISC ODM [13], Resource Description Framework [14]) have been assessed regarding their ability to transmit extracted meta information from clinical data sources to the meta data oriented patient register. An important demand put on an appropriate metadata format is its power to convert data from legacy study databases with various technological back-ends (e.g. MS Access, MS SQL Server) to an international accepted format. The assessment resulted in the choice of CDISC ODM to act as model for implementation of metadata standardisation, extraction, transmission and storage.

Software interfaces and tools were modeled with UML 2.0 and implemented with Java, JAXB, XML, Hibernate, Lauch4J, Ant, Maven. A PostgreSQL database acts as back-end for central metadata storage.

3. Results

An evaluation of possible meta information about a clinical data source to be extracted and loaded into the metadata register was conducted and resulted in the following definition on which meta information will be collected about a clinical data source: (1) Attributes of the research project (e.g. project type, research plan synopsis, etc.), (2) status of data management processes (e.g. data capture, data validation, database closure, etc.), (3) Description, structure and content of (electronic) case report forms

(e.g. scheme of study visits and forms), (4) Description of data items (e.g. item description, data type and precision, location of item in case report form, etc.) , (5) Data validation plan, (6) Pseudonyms of included patients, and (7) "Captured/Missing-Flag" (i.e. a True/False flag, indicating on the data item level, if clinical information about a single patient was captured (True) or is missing (False))

Since the CDISC ODM format isn't able to document the Captured/Missing flag an extension of the ODM standard was required. The ODM extension was documented in an amended XML schema.

Software for fully automatic metadata and clinical data extraction from distributed data sources under different ownership was implemented. It allows the data owning study group to control the transferred data. On one hand it can be configured to extract patient pseudonyms and Captured/Missing information. This conversion of clinical information to the metadata format is conducted on basis of mapping information. The mapping instructions are documented in XML format defined by an XML schema. The so called 'DB2ODMMapping' allows the specification of mapping constraints between a relational database and ODM data items as well as constraints on interpreting the Captured/Missing-Status of data values.

Second the software is able to extract clinical data from a relational database on request of a cooperative research project. The clinical data to be extracted can be configured in the 'DB2ODMMapping' file.

Further software tools for processing of collected meta information have been implemented, i.e. for loading metadata into the central database and for creation of meta data documentation in PDF format.

All project related software has been implemented in Java 6. A modular concept of three Java APIs (core, dataaccess, odm) support software maintenance and enable software re-usability.

At present meta information about three clinical trials from two AML study groups has been integrated into the central meta-data based patient register. Together these three data sources contain clinical information about 4115 leukemia patients.

Automatic extraction of clinical data from the study databases on basis of available meta information has been tested. Clinical evidence concerning the status and classification of AML (i.e. French-American-British classification, WHO classification) from 4102 patient data sets has been extracted and provided for statistical analysis. This process disclosed classification inconsistencies between the trials and allowed to start a process to standardize between both study groups.

The prototype allows straightforward extension to the full set of available clinical trial in several study groups.

4. Discussion

The challenges of clinical research ask for a cooperative efficient use of high-quality data. Such data is in general available in databases of clinical trials, especially of randomized controlled studies. Sharing the data of such studies has to be done with care and within a transparent and regulated setting to protect patient rights as well as integrity of the clinical data. The concept and prototype for a general cooperative infrastructure in clinical research is presented which complies with legal, ethical and technical requirements. It supports cooperative initiatives in consolidation of available clinical evidence for evaluation of open research questions. Potential cooperative

projects are: (1) Discovery and validation studies for prognostic and predictive models, biomarkers and surrogate endpoints, (2) planning data capture for future trials, and meta-analyses using individual patient data (surrogate endpoints, treatment effects, subgroup analyses).

Processes for metadata extraction and loading into the central register facility have been implemented and are highly supported by comfortable software tools. In addition, the metadata-based patient register acts as a platform for network communication and data standardization activities. Besides the ongoing integration of metadata from clinical study databases future work will concentrate on modeling and implementation of a web-based register platform and of data transformation processes for harmonizing clinical data from different sources.

References

[1] Dreyer NA, Garner S. Registries for robust evidence. *JAMA.* 2009 Aug 19;302(7):790-1.

[2] Gliklich RE, Dreyer NA, editors. *Registries for Evaluating Patient Outcomes: A User's Guide*. 2nd edition. Rockville (MD): Agency for Healthcare Research and Quality (US); 2010 Sep.

[3] Drolet BC, Johnson KB. Categorizing the world of registries. *J Biomed Inform*. 2008 Dec;41(6):1009-20. Epub 2008 Feb 5.

[4] Brooke EM. The current and future use of registers in health information systems. *WHO Offset Publ* No. 8 1974 pp. ii + 43 pp.

[5] Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc*. 2002 Nov-Dec;9(6):600-11.

[6] Gladman D, Menter A. Introduction/overview on clinical registries. *Ann Rheum Dis*. 2005 Mar;64 Suppl 2:ii101-2. Review.

[7] Stausberg J, Altmann U, Antony G, Drepper J, Sax U, Schuett A. Registers for networked medical research in Germany: Situation and prospects. *Appl Clin Inf*, 2010. 1: p. 408-418.

[8] Newton J, Garner S. *Disease Registers in England.* Institute of Health Sciences, University of Oxford, 2002. ISBN 1 8407 50286.

[9] National Institutes of Health, National Center for Research Resources. CaBIG™ overview. 2006. [cited at 2011 Apr 29]. Available from http://www.ncrr.nih.gov/publications/informatics/caBIG.pdf.

[10] European Medicines Agency, Committee for Orphan Medicinal Products. Public summary of opinion on orphan designation, EMA/COMP/804144/2009. London, 2010.

[11] Hehlmann R, Berger U, Aul C, Büchner T, Döhner H, Ehninger G, et al. The German competence network 'Acute and chronic leukemias'. *Leukemia.* 2004 Apr;18(4):665-9.

[12] ISO/IEC 11179-3+COR1 (2003) Information Technology - Metadata Registries (MDR) Part 3: *Registry Metamodel and Basic Attributes*. Second edition 2003-02-15 Incorporating COR1. Available from http://jtc1sc32.org/doc/N1151-1200/32N1168-ISO-IEC11179-3-2003COR1.zip.

[13] http://www.cdisc.org/models/odm/v1.3/index.html. [cited 2011 Mar 06].

[14] http://www.w3.org/standards/techs/rdf#w3c_all. [cited 2011 Mar 06]