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

doi:10.3233/978-1-61499-432-9-1115

A Comprehensive Clinical Research Database based on CDISC ODM and i2b2

Frank A. MEINEKE[a,b,1], Sebastian STÄUBERT[a,c], Matthias LÖBE[a], Alfred WINTER[a]
[a] *Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University*
[b] *IFB Adiposity Diseases, Leipzig University*
[c] *Centre for Clinical Trials Leipzig, Leipzig University*

Abstract. We present a working approach for a clinical research database as part of an archival information system. The CDISC ODM standard is target for clinical study and research relevant routine data, thus decoupling the data ingest process from the access layer. The presented research database is comprehensive as it covers annotating, mapping and curation of poorly annotated source data. Besides a conventional relational database the medical data warehouse i2b2 serves as main frontend for end-users. The system we developed is suitable to support patient recruitment, cohort identification and quality assurance in daily routine.

Keywords. Medical Informatics, Health Information Systems, Information Storage and Retrieval, Clinical Trials, Hospital Information Systems, CDISC ODM

Introduction

Clinical informatics provides information technology (IT) and services for health care and research. The data centre project of the Integrated Research and Treatment Center Adiposity Diseases (IFB) [1] develops methods which supports clinical research projects in the phases of planning (including cohort identification, feasibility, data reports for hypothesis generation), implementation (e.g. data-entry-systems, recruitment management) quality assurance (e.g. data monitoring, benchmarking) and information translation back into clinical routine.

To do so, we have built up a clinical research database (RDB) which integrates as many relevant data sources from the hospital information system (HIS) as possible (laboratory, biobank registry, health record, secondary sources) and data from clinical trials of the IFB, held in the clinical data management system (CDMS). When the IFB started in 2010 and the outpatient clinic opened, no structured obesity-specific documentation existed. Even weight and height of a patient, needed to calculate the Body Mass Index – the single most important measure for obesity diagnosis – were missing. A detailed entry form (PMD, parameterized medical document) for structured obesity anamnesis and documentation for the clinical workplace SAP/Siemens i.s.h.med was devised and implemented. Every patient of the obesity outpatient clinic was documented, although the database behind was a black box for physicians and researchers.

[1] Corresponding Author: frank.meineke@imise.uni-leipzig.de

We decided to develop an easy to use system based on standard technologies and open source software to establish a RDB at the University Medical Center Leipzig.

1. Methods

1.1. Clinical Research Database (RDB) – an Archival Storage Information System

The RDB presented here is not a single perfectly trimmed database, but is instead the core of an Archival Information System (AIS) in the sense of OAIS [2]. The methods we describe can be interpreted as part of the functional entities (or layers), ingest (ETL with ODM target), archival storage (ODM based) and access (i2b2 [3], SQL database). We conceived the RDB knowing that the most time consuming challenge is not setting up an appropriate, secure IT research infrastructure, but to support the early steps of the data lifecycle from conceptualisation to the ingest-phase [4]. To be of practical use, a comprehensive RDB has to cover storage and access functionalities and at least in a modest way aspects of data curation, normalization and standardization.

Instead of using transformations from every proprietary data source to an evolving proprietary RDB and then implement transformations to all potential access systems, we use a standardized format for medical data representation as stable target for all data transformations from external source to the access layer, thus completely decoupling ingest and access layer. Only transformations to and from this standard are necessary. Changes at the producing or consuming systems do not affect the middle, storage tier.

1.2. CDISC ODM

"The Operational Data Model (ODM) is designed to facilitate the archive and interchange of the metadata and data for clinical research" [5]. Various CDMS can write, some statistical software can read ODM. An ODM file contains metadata (<Study>) and a data part (<ClinicalData>). The metadata defines the structure of a study in terms of events, forms, item groups and items with data types and code lists. The clinical data section refers to these definitions and contains the actual data for each subject (patient / proband). This structure is oriented toward clinical trials [6], but as we discuss here, it can be mapped to represent hospital data as well.

1.3. Access Layer

Typical users of a RDB are biostatisticians and physicians or PIs (principal investigator of a clinical trial), often with very different needs and qualifications.

We built up a classical relational database for the first group: a scheme for each data source, tables for every group of items, tables for code-lists and so on. Access rights can be given in very fine granularity. This database is suited for statistical analysis of data with well-known endpoints, but not for ad hoc queries since scheme overarching queries grow complex especially as multiple joins and patient-id mapping tables become too large.

In contrast, the medical data warehouse i2b2 uses a user friendly web-based query tool to search for specific patient groups in heaps of data coming from different sources [7]. It operates on a star-scheme, holding facts in a single EAV (Entity-Attribute-

Value) table. Queries are assembled in groups by drag and drop and refined by operators (e.g. and, or, exclude) or constraints (data range, occurrence counts, same encounter).

2. Results

We installed the RDB inside the clinic as part of the HIS. In Saxony, clinical data may be used freely for research within the institution. However, to contact an outpatient patient at home, e.g. because he/she qualifies for a clinical trial, or to use data outside the clinic, written consent is obligatory. More than 50% of our patients gave informed consents to do both.

2.1. Data Integration – Ingest Layer

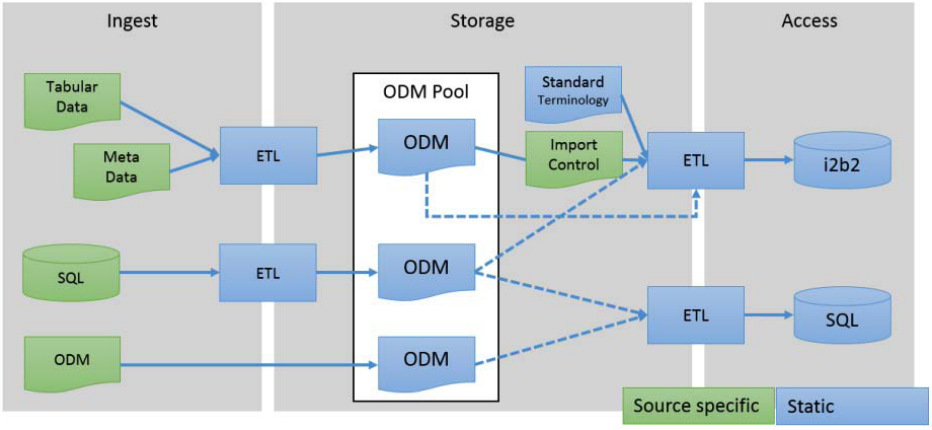

Figure 1. Data integration pipeline

Data coming in tabular format is transformed to ODM by a generic, highly customizable job written in a dedicated data integration tool in three steps:

- de-identification of data (e.g. names, addresses, free text)
- curation (e.g. remove test cases, respect special missing codes, correct known data-entry errors and so on)
- annotating (e.g. comments, grouping labels) and normalization (e.g. use of UCUM (Unified Code for Units of Measure) from a separate metadata definition file

The metadata description is derived from structural data analysis, interviews and existing documentation. This applies to all data coming from HIS. For non-tabular sources, namely CDMS systems, the above steps are not necessary as data is already quality assured. If the CDMS cannot provide ODM on its own, the data is transformed directly from the source database.

All ODM files are archived in a pool. Data subsets are assigned to user group specific projects, e.g. for the paediatric or adult departments. The process pipeline from storage to i2b2 is controlled and configured by a source specific control file, defining superfluous structure levels, mapping local to generics concept (e.g. gender, age), inserting standard terminology trees (ICD, OPS), mapping data items to i2b2-metadata

(e.g. subject key, timestamp of observation) and more. Other control elements manage database connections, define the destination ontology path or include nested import jobs. The import from ODM to SQL database is based on XSL Transformation.

2.2. Mapping hospital data to ODM

We started with two specific obesity documentation forms (PMD) used in the children and adult outpatient clinics and in bariatric surgery with about 300 / 200 defined items. A standardized dataset from the HIS (§21 KHEntgG [8], e.g. diagnoses and procedures), biobank registry data and trial structure data was added for testing. Record linkage is based on hospital patient-id. The definition of the metadata and data cleansing process is troublesome, as exports from the SAP system neither contain metadata nor inner structure. Table 1 shows the mapping used. The top level nodes of the i2b2 ontology are fixed as documents, diagnoses, studies etc.

Table 1. HIS to ODM to i2b2 Mapping

Source	ODM	I2b2
SAP encounter	StudyEvent	
SAP document	FormEvent	2nd level
Metadata module definition	ItemGroup	3rd level
SAP data columns + Metadata meas. units	Item	4th level
Metadata codes/labels	CodeList	5th level
ICD-10-GM / OPS/ICPM catalog		5th to nth level

Additional information (e.g. case-id, date of observation, patient age or sex) was retrieved by XPath expressions directly from corresponding ODM item values, entities or attributes – whatever is available. ODM itself does not determine these.

2.3. User Feedback and usage Scenarios

I2b2 kept its promise to be easy to use. Even people having only a vague knowledge of IT and database internals needed only 1-2 hours of training. We encountered two typical usage scenarios:

1. Immediate help in recruitment phase, e.g. *Find patient with BMI > 40 kg/m² without diabetes*. I2b2 exports patient-id, candidates are looked up and double checked against their health record, double checked for their written consent, and finally contacted.
2. Quality assurance for the outpatient clinics, e.g. check for patients not fulfilling outpatient inclusion criteria or concerning data quality.

In both cases help was appreciated as a true time and labour saving option; the alternative would have been to scan health records or to send unspecific invitations. However, temporal questions (e.g. X occurs not more than 3 weeks after Y) are still difficult. Usual queries took less than 10 seconds.

The system itself is not easy to setup. Data import and maintenance of the more elaborate features are not well supported, although some third party effort is currently developed [9], from which we used the installation and administration tool.

3. Discussion

Much care has been taken to make i2b2 usage as comfortable as possible by using an intuitive concise ontology and readable item labels, integrating clear labeled code lists and standard-terminologies. It took time and some persuasion to implement the system into research routine, but clinicians and researchers are beginning to appreciate the new possibilities. The black-box RDB became more transparent, especially physicians working in the clinic and for a clinical trial see the immediate benefit of the structured documentation. The possibility to explore their *own data by themselves* provides confidence, trust and fosters acceptance. We expect the RDB will be used for trial planning (estimate sample sizes, develop hypothesis) in the future.

Our ODM centric approach seemed like an artificial complication at first, but in fact reduced the complexity of the whole system by managing rather simple ETL jobs reading or writing ODM. ODM proved to be flexible enough to handle all data sources and could be easily processed and understood.

We are working on a more general approach outside the clinical context, which allows true trial and HIS data overarching queries. This solution demands additional components and processes, such as a pseudonymization service, trustee managed patient lists. Data exports for retrospective analysis will play a bigger role, making a second level of pseudonymization and elaborated reporting tools necessary. Finally this concept will lead to the establishment of an Archival Informations System.

Data quality of primary care is often criticized, analysts and physicians know this. Particular financially motivated coding is problematic in research. [10] Nevertheless this is not a big issue in the scenarios considered here, as the patient set is big enough to minimize the effect of a few false negatives. The best way to assure data quality is to work with it.

A**cknowledgement:** This work was supported by the Federal Ministry of Education and Research (BMBF), Germany, FKZ: 01EO1001 and 01KN1102

References

[1] IFB, "Integrated Research and Treatment Center AdiposityDiseases," [Online]. Available: http://www.ifb-adipositas.de/en. [Accessed 07 02 2014].
[2] OAIS, Reference Model for an Open Archival Information System (OAIS), ISO 14721:2012.
[3] i2b2, "Informatics for Integrating Biology and the Bedside," [Online]. Available: https://www.i2b2.org/. [Accessed 07 02 2014].
[4] S. Higgins, „The DCC Curation Lifecycle Model," The International Journal of Digital Curation, Bd. 3, Nr. 1, pp. 134-140, 2008.
[5] CDISC, "Operational Data Model," [Online]. Available: http://www.cdisc.org/odm. [Accessed 07 02 2014].
[6] M. Löbe, Einsatzmöglichkeiten von CDISC ODM in der klinischen Forschung, Proc. 57th GMDS Annual Meeting, Braunschweig 2012, 1294 – 1304.
[7] T. Ganslandt, Mate S, Helbing K, Sax U und H. U. Prokosch, „Unlocking Data for Clinical Research – The German i2b2 Experience," Applied Clinical Informatics, Bd. 2, Nr. 1, pp. 116-127, 2011.
[8] KHEntgG, §21 Transmission and use of data, German Hospital Financing Act, 1972/2013.
[9] T. Ganslandt, Integrated Data Repository Toolkit: Werkzeuge zur Nachnutzung medizinischer Daten für die Forschung, GI, Lecture Notes in Informatics, Bonn 2012, 1252 – 1259.
[10] T. Botsis, G. Hartvigsen, F. Chen und C. Weng, „Secondary Use of EHR: Data Quality Issues and Informatics Opportunities," AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science, Bd. 2010, pp. 1-5, 2010.