A clinical data warehouse based on OMOP and i2b2 for Austrian health claims data

Christoph RINNER[a,1], Deniz GEZGIN[a], Christopher WENDL[a] and Walter GALL[a]
[a]*Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna*

Abstract. Background: To develop simulation models for healthcare related questions clinical data can be reused. **Objectives**: Develop a clinical data warehouse to harmonize different data sources in a standardized manner and get a reproducible interface for clinical data reuse. **Methods**: The Kimball life cycle for the development of data warehouse was used. The development is split into the technical, the data and the business intelligence pathway. **Results**: Sample data was persisted in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). The i2b2 clinical data warehouse tools were used to query the OMOP CDM by applying the new i2b2 multi-fact table feature. **Conclusion**: A clinical data warehouse was set up and sample data, data dimensions and ontologies for Austrian health claims data were created. The ability of the standardized data access layer to create and apply simulation models will be evaluated next.

Keywords. Secondary use, standardized health data, clinical data warehousing.

1. Introduction

Clinical data reuse is defined as “non-direct care use of personal health information including but not limited to research” [1]. In [2] it is concluded that clinical data reuse can constitute an important pillar to increase the quality in medical research and efficiency of clinical research. Reused data can come from various sources, documented for various purposes in varying granularities and levels of structure. To successfully reuse these data, the source data have to be cleaned, harmonized and pre-processed to allow meaningful research.

Various efforts exist to unify clinical data reuse by storing data in harmonized data models. The Observational Health Data Sciences and Informatics (OHDSI) program [3] supports the community in the development and adaptation of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Informatics for Integrating Biology and the Bedside (i2b2) [4] offers a generic data model for clinical data and an open source clinical data warehouse framework with advanced access and querying mechanisms and good extensibility using so called cells. In [5] the CDISC Operational Data Model (ODM) was implemented using i2b2 as front-end.

[1] Corresponding Author: Christoph Rinner, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria, E-Mail: christoph.rinner@meduniwien.ac.at.

A clinical data warehouse offers life science businesses easier access to stored information and gives insights into their health care data. As part of the imProve project[2] we want to use simulations of the Austrian health care landscape to innovate the health care industry. Simulation models are models of reality which are created and parameterized by the model builders using information resources describing reality. The expected benefit of a clinical data warehouse with a standardized interface between the developed simulation models and the source data is easier deployment of the models and reproducible results.

This article describes the process of implementing a local clinical data warehouse infrastructure based on the OMOP CDM and the i2b2 data warehouse infrastructure as front-end to analyze Austrian health claims data.

2. Methods

Based on the Kimball Lifecycle diagram [6] a local clinical data warehouse infrastructure was created. The business requirements are specified first, and then the lifecycle splits up into three pathways. In the technical pathway strategic and technical directions are planned based on the business requirements. The data pathway reflects the categorization of data into measurement facts and descriptive dimensions. Dimensions can be seen as structures that categorize the measurement facts to enable users to answer business questions. First the relevant dimensions based on the available data (i.e. measurement facts) are analyzed and formalized, then the data models to store the measurement facts are created and finally the source data are loaded into the data warehouse. In the business intelligence pathway the information needs are analyzed and tools to access and query the data are created.

3. Results

3.1. Business Requirements

As a first use case Austrian health claims data were selected. Health claims data are documented in a highly structured form needed for reimbursement purposes and existing know-how from previous projects with this type of data was available. To prevent data privacy concerns when using real claims data sample data was created. The clinical data warehouse should be extendible for future use cases in respect to the clinical data stored but also in respect to technical interfaces like future distributed query possibilities. The data stored in the data warehouse should be accessible during the building of simulations models to parameterize the model and as input during the execution of simulations models.

3.2. Technical path

As part of a previous project [7], an i2b2 clinical data warehouse was deployed for Austrian health claims data. A special focus was put on technical aspects like the handling of large data sources and compartmentalization using Docker containers. I2b2

[2] http://www.dwh.at/de/expertise/projekte/improve/

implements various strategies to maintain patient privacy [8] and offers a generic database design based on the star schema. Observations and the clinical data are stored in five main tables. This generic structure allows for a high degree of flexibility yet increases complexity. In the current version of i2b2 the multi-fact table concept is introduced to enable i2b2 to also access data from other data sources beside the i2b2 observation fact table[3]. This feature allows for example OMOP CDM as source for observational data in i2b2 using database entries without changes to the i2b2 source code. OMOP CDM is optimized for claims data with separate tables for drug costs, visits costs etc. resulting in an easier to understand column-based database design.

Based on the business requirements the OMOP CDM (v5.0, October 2014) and i2b2 (v1.7.09b) accessing OMOP CDM via the new multi-fact table feature was selected. For the deployment a virtual machine based on VMWare ESXI (v5.5), with 1 CPU (3 GHz), 12 GB RAM and 130 GB of disk space was created. As operating system Ubuntu server (v16.04.2) was selected to host i2b2 and the Microsoft SQL Server (v14.0).

3.3. Data path: Analysis of data dimensions

The OHDSI Athena[4] allows the distribution of standardized vocabularies, which already use the correct data format for the OMOP CDM. The two dimensions for the Anatomical Therapeutic Chemical (ATC) Classification System used to document dispensed drugs and the International Statistical Classification of Diseases and Related Health Problems 10th revision (ICD-10) used to document hospital diagnoses in Austrian health claims data could be reused directly.

Additionally to the ATC codes, pharmaceutical products sold in Austria are also identified using the unique Austrian pharmaceutical registration number (PRN). The 16,500 medications represented by unique PRN are modeled as separate *concepts* (with distinct entries for *vocabulary* and *concept_class*) in the OMOP CDM. Using *concept_relationship* the PRN are linked to the existing ATC codes.

The catalogue of 1,951 individual medical services (i.e. “MEL codes”) was hierarchically structured into chapter, subchapter, unit, anatomy (coarse and fine), access and source. Each MEL code and structure was modeled as new *concepts* and linked using the *concept_relationships*.

Austria can be divided into more than 17.000 cities (i.e. “Orte”) with distinct postal codes located in 2,100 counties (i.e. “Gemeinden”) located in in 32 supply regions (i.e. “Versorgungsregionen”) located in 9 states (i.e. “Bundesländer”) and finally four supply zones (i.e. “Versorgungszonen”). The *location* table in OMOP CDM has fields for all entities except the supply zones and regions. Therefore these two dimensions were documented in the *address_2* field instead of creating separate *observations* for each visit.

Claims data in Austria originates from the 19 insurance carriers which were modelled as *concepts*. For each *visit_occurence* one of 281 *care_sites* was assigned including the unique hospital id, state, supply region and supply zone.

[3] https://community.i2b2.org/wiki/display/MFT/Multi-fact+Table+Home
[4] https://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/

3.4. Data path: Physical design of OMOP Schema

From the 36 tables available in the OMOP CDM only 15 were used to document the Austrian health claims data. The *person, death* and *location* tables are used to store information about the patients. The *provider* table (including *care_site* and *location*) are used to store information about the health care provider. Hospital visits were modelled as *visit_occurences* with *observations*, *measurements* and *condition_occurence* to document information about the visit. Prescribed and dispensed medications are stored in the *drug_exposure* table. The previously described data dimensions are modelled using the tables *concept, concept_class, domain, vocabulary* and *concept_relationship*.

3.5. Data path: Integration of the claims data

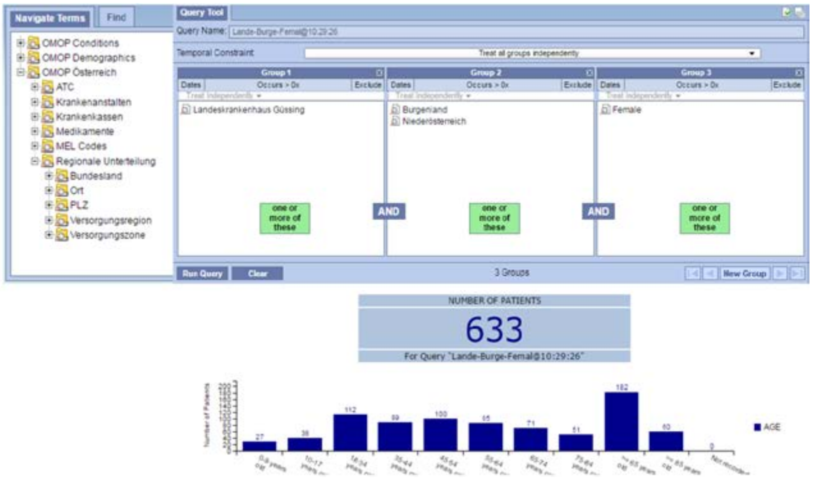
To evaluate the created clinical data warehouse infrastructure and to prevent data privacy issues we created sample data for female Austrian patients base on the distribution of age, discharge diagnoses and medication distribution in the Austrian population. The final sample data consists of 4.3 million patients with 460.000 hospital stays and 10.3 million medications representing 16.500 different pharmaceutical products and covered a time period of 3 years. We initially evaluated the Observational Medical Dataset Simulator (OSIM)[5] for OMOP. Due to version incompatibility of OSIM we could not apply it directly and decided to created comma separated value (csv) files using a simple JAVA routine with the distributions as input. All csv files were imported into the database and all transformations were performed using SQL scripts. This approach, where the data wrangling and transformations are performed directly in the database instead of separate tools is called extract, load and transform (ELT) process [9].

3.6. Business Intelligence

To get a generic overview of the imported data in the OMOP CDM the Achilles[6] framework was used. To allow more in-depth analysis and to control access on a per user base, the i2b2 web interface was used. In Figure 1 an overview of the query interface with the various dimensions is depicted. To access OMOP CDM with i2b2 the i2b2 multi-fact tables were enabled and the OMOP CDM data was linked to i2b2 by creating a dedicated i2b2 ontology for our data. This OMOP_AUSTRIA ontology references the OMOP CDM tables directly. For complex queries (e.g. “Patients with more than 3 hospital stays”) we created separate database views also accessible using the multi-fact feature. The final goal will be to create simulation models based on the data stored in the data warehouse and offer access to the simulations models to the end users.

[5] ftp://ftp.ohdsi.org/osim2/

[6] https://www.ohdsi.org/analytic-tools/achilles-for-data-characterization/



Query Name: Lande-Burge-Femal@10:29:26

Temporal Constraint: Treat all groups independently

Group 1: Landeskrankenhaus Güssing

Group 2: Burgenland, Niederösterreich

Group 3: Female

3 Groups

NUMBER OF PATIENTS

633

For Query "Lande-Burge-Femal@10:29:26"

Figure 1: Overview of the i2b2 query editor with the created dimension on the left hand side, a sample query in the center and the results at the bottom.

4. Discussion and Conclusion

We implemented a clinical data warehouse based on i2b2 using the new i2b2 multi-fact tables to access health claims data stored in the OMOP CDM. Austrian dimensions for pharmaceutical registration numbers, regions, health care providers, insurance carriers and medical services were created. Sample data of more than 4 million patients was created and imported using SQL scripts. Using generated data and health claims data in particular, the import was quite straight forward since data were already structured and no data cleaning, data extraction or linking was needed. The first cycle of the Kimball life cycle for data warehouses was finished.

The created data warehouse can be deployed at different data holders using the virtual image created. To deploy the data warehouse extract, load and transformation steps are still needed to import the source data. To create simulation models, the Achilles tool is a good starting point to gain an overview of the stored data and for quality assessment. Depending on the use case specific reports (e.g. regional distribution) should be developed. The i2b2 interface allows direct access to cohort estimations on a very fine grained level (e.g. number of patients with a specific diagnosis in a specific time frame followed by a specific medication). If the data holders are not willing to offer direct access to their data due to data privacy regulations and privacy concerns, Achilles and i2b2 enable the data holder to extract needed parameters themselves without advanced technical background.

As a next step we plan to quality assess the developed dimensions and distribute them as standardized vocabularies with the Athena tool from the OHDSI community. In a similar fashion the i2b2 ontology we created to query OMOP CDM data representing Austrian health claims data with i2b2 should be made available to other researchers. We plan to refine the import scripts to automate the import process further and to allow all Austrian health care providers documenting health claims data or insurance carriers receiving health claims data to easily reuse the created dimension and store their data in their locally deployed OMOP CDM. This is a first step to prepare a small subset of dimensions used in the Austrian health care landscape for reuse purposes.

The OMOP CDM and the i2b2 ontology offer a good framework to formalize dimensions. Reusing these dimensions can help to facilitate future data preparation tasks and make research easier to compare. At the time of the first deployment, the multi-fact tables in i2b2 were only optimized for Microsoft SQL servers. Since these are now also available for PostgreSQL we are currently migrating the dimensions, ontologies, import scripts to PostgreSQL. The OMOP CDM allows an easier selection of stored data using SQL, compared to the generic i2b2 data model. I2b2 only distinguishes between patients, visits, observations and observers, all other nuances in the data are hidden behind the concept dimension. OMOP offers the Achilles tools to get a quick overview of the data in the data warehouse and with the i2b2 query editor a simple interface for end users is available.

When analyzing claims data OMOP CDM already offers specific tables for medications services. With harmonized ontologies in i2b2 the same effect can be achieved. Combining the data model of OMOP CDM and the query interface and data management of i2b2 is relatively easy to achieve and enhances both products. Our proposed data warehouse implementation can build a starting point for other projects to enable reproducible and comparable research and facilitate the development and application of simulations models.

5. Acknowledgement

This research was supported by Vienna Business Agency. We thank all our collaborators in the imProve project especially Georg Duftschmid, Nicolas Popper and Barbara Glock.

6. References

[1] Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association : JAMIA. 2007;14(1):1-9.

[2] Meystre SM, Lovis C, Burkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearbook of medical informatics. 2017;26(1):38-52.

[3] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in health technology and informatics. 2015;216:574-8.

[4] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association : JAMIA. 2010;17(2):124-30.

[5] Meineke FA, Staubert S, Lobe M, Winter A. A comprehensive clinical research database based on CDISC ODM and i2b2. Studies in health technology and informatics. 2014;205:1115-9.

[6] Kimball R. The data warehouse lifecycle toolkit: John Wiley & Sons; 2008.

[7] Endel F, Duftschmid G. Secondary Use of Claims Data from the Austrian Health Insurance System with i2b2: A Pilot Study. Studies in health technology and informatics. 2016;223:245-52.

[8] Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy in i2b2. Journal of the American Medical Informatics Association : JAMIA. 2011;18 Suppl 1:i103-8.

[9] Soutier M. Von ETL zu ELT in Big Data-Systemen 2015. Available from: http://www.soutier.de/blog/2015/03/01/von-etl-zu-elt-big-data/.