JAMIA Open, 1(2), 2018, 172–177
doi: 10.1093/jamiaopen/ooy039
Advance Access Publication Date: 20 September 2018
Case Report

Case Report

Secondary use of routine data in hospitals: description of a scalable analytical platform based on a business intelligence system

Jan A. Roth,[1,2] Nicole Goebel,[2,3,4] Thomas Sakoparnig,[2,5,6] Simon Neubauer,[2,3,4] Eleonore Kuenzel-Pawlik,[2,3,4] Martin Gerber,[2,4] Andreas F. Widmer,[1,2] Christian Abshagen,[2,4] Rakesh Padiyath,[2,4] and Balthasar L. Hug[2,7]; the PATREC Study Group[†]

[1]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland, [2]University of Basel, Basel, Switzerland, [3]Analytics Unit, Department of Finance, University Hospital Basel, Basel, Switzerland, [4]Department of Finance, University Hospital Basel, Basel, Switzerland, [5]Focal Area of Computational and Systems Biology, Biozentrum, University of Basel, Basel, Switzerland, [6]Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Basel, Switzerland and [7]Department of Internal Medicine, Kantonsspital Luzern, Lucerne, Switzerland

[†]Members are listed in the Acknowledgements section.

Corresponding Author: Balthasar L. Hug, MD, MBA, MPH, Department of Internal Medicine, Kantonsspital Luzern, Spitalstrasse, 6000 Luzern 16, Switzerland (balthasar.hug@luks.ch)

Received 14 December 2017; Revised 11 May 2018; Editorial Decision 7 August 2018; Accepted 31 August 2018

ABSTRACT

We describe a scalable platform for research-oriented analyses of routine data in hospitals, which evolved from a state-of-the-art business intelligence architecture for enterprise resource planning. This platform involves an in-memory database management system for data modeling and analytics and a high-performance cluster for more computing-intensive analytical tasks. Setting up platforms for research-oriented analyses is a highly dynamic, time-consuming, and costly process. In some health care institutions, effective research platforms may be derived from existing business intelligence systems.

Key words: health services research; database management systems; health information systems; high performance analytic appliance (HANA); machine learning

INTRODUCTION

The amount and variety of real-world hospital data (HD)—that is routinely generated and collected data in the course of health care delivery—is steadily increasing and offers a unique opportunity to inform and support clinicians, researchers, and hospital administrators by creating novel hypotheses, predictions, and evidence as part of pragmatic studies.[1,2] However, as the amount of electronic HD is growing, demands on data storage and integration, computing power and data analytics have also increased.[3] Implementing and optimizing large-scale health information technology is an ongoing and challenging process, especially when used for research purposes.[4,5]

With the vision to enable innovative intra-institutional research on large HD sets and to improve health care quality at our institution, we expanded the capabilities for modeling and analyzing HD, and we implemented a standard process for analyzing large sets of HD. We considered large data(sets) as HD whose volume, velocity, and/or variety make it difficult to manage and analyze by use of conventional systems and methods.

As little is known about widespread proprietary database systems and for greater transparency, we sought to describe and discuss our systems and processes involved in modeling and analyzing large HD for intra-institutional research purposes, which evolved from an existing business intelligence architecture.

Integrating various data within a hospital into a database or warehouse may be an important step for subsequent cross-institutional collaborations—for instance via common data models provided by the "Observational Health Data Science and Informatics" collaboration and other organizations[6–8]; Respective applications and experiences have been described previously and are not covered in this report.[6–11]

METHODS

The University Hospital Basel (UHB) is the 850-bed tertiary referral center of Northwestern Switzerland with more than 1 000 000 ambulatory patient contacts and over 36 000 inpatients per year—pulling its HD from more than 100 source systems. At UHB, we formed a multidisciplinary team combining expertise in data management, business analytics, informatics, machine learning, epidemiology, and clinical domains in 2016 as part of an ongoing data mining study. In the framework of this single-center study project, we (1) expanded the capabilities for modeling and analyzing HD, and we (2) implemented a standard process for research-oriented analyses of large HD sets.

The Ethics Committee of Northwestern and Central Switzerland (EKNZ) approved this project (number 2016-02128).

Analytical platform

Based on experiences with our business intelligence architecture and challenged by restricted computing power and limited data models as obstacles for advanced analyses (eg machine learning), we have developed a scalable platform for research-oriented analyses of HD (Figure 1). This platform consists of a multipurpose state-of-the-art database management system, which can be used both for administrative and research purposes without affecting the system performance. The goal was to facilitate the following research tasks:

1. Selecting relevant HD variables of predefined study populations via an anonymized data interface.
2. Modeling and linking structured data originating from the hospital's main source systems (patient demographics, medication, laboratory and other diagnostic parameters, and administrative data [eg International Classification of Diseases diagnosis codes, procedure codes, etc.]).
3. Generating deidentified target datasets with appropriate data representations and data formats.
4. Analyzing large HD sets with sufficient processing speed.

In the light of precision medicine, linking individual's clinical, genomic and molecular information has become a key research priority for large health care institutions.[12] With the main aim to predict the response to different treatments, some health care organizations such as the Mayo Clinic began to integrate clinical and genomic data; these processes however may be challenging due to different data sharing policies, the rapid pace of innovation in the field of bioinformatics and IT, the need for sufficient metadata, and the highly diverse requirements of end-users.[12] At our institution, a project is ongoing to specifically address the requirements of researchers in the field of precision medicine; for some requirements, it is likely that our current analytical platform may play a central role—for instance to integrate genomic data.[7] At the moment, our platform is not used to analyze omics data.

At the core of our analytical platform is a SAP® High Performance Analytic Appliance database (HANA; SAP AG, Walldorf, Germany), which is an in-memory relational database management system.[13,14] It was introduced at the UHB in 2014 in order to handle large HD sets. Since then, the HANA database was in use for financial controlling, logistic, and business reporting purposes.

By means of recent upgrades, the database was complemented with proprietary modules for clinical data. Although other existing databases could have been used at our institution to set up such an analytical platform, we chose to use the already existing HANA database to enable research-oriented analyses on HD sets. This decision was mainly based on the advantages listed in Table 1. The approach of a combined clinical and research in-memory data management system was pursued to satisfy both business and research needs, to benefit from the already existing experience with HANA functionalities and to readily implement a platform for research-oriented analyses.[15]

Structured data and metadata from the main hospital source systems (administrative systems, electronic medical records, laboratory, and other diagnostic information systems) are replicated 1:1 via SAP® Landscape Transformation servers into the HANA database. In doing so, data are replicated without affecting the performance of the source systems. Routine data from the hospital's radiology and pathology information systems (including high-throughput sequencing data) are not yet incorporated into the HANA database. However, integration of the 2 systems is planned with replication of HD originating from these systems into the HANA database. This will allow researchers to study complex multi-level relationships between administrative, clinical, radiological, pathological, laboratory, and microbiological data.[13,14] Furthermore, our HANA architecture will be complemented with a data lake to store large amounts of structured, semi-structured and unstructured HD at low cost, mainly for research purposes (eg medical imaging data); these data will be made accessible via the HANA database.

Within the HANA database, initial analyses are performed by experts of the analysis technology team. At this level, the models are still 1:1 system-true representations and contain basic key figures.

To make HD utilizable for analytical research purposes, Qlik Sense® (QlikTech GmbH, Düsseldorf, Germany) is primarily used as frontend tool. Qlik Sense® enables advanced data algorithm development and flexible data structure output. On this level, deidentified and more complex data models and reports are generated that may be composed of several systems. The aim is to connect all source systems and to model them in a meaningful way for authorized users (administration and research). Access authorization is controlled with a specific authorization scheme; up to now, research data queries and initial analyses are solely performed by experts of the analysis technology team. An ethical approval by the local institutional review board is required for research-oriented analyses.

Data management and analysis process

For research-oriented analyses of large deidentified datasets using the HANA database and respective frontends, we established a standard iterative data management process consisting of the following main steps adapted from the knowledge discovery in databases process (Figure 2).[16] The main data modeling steps are performed with

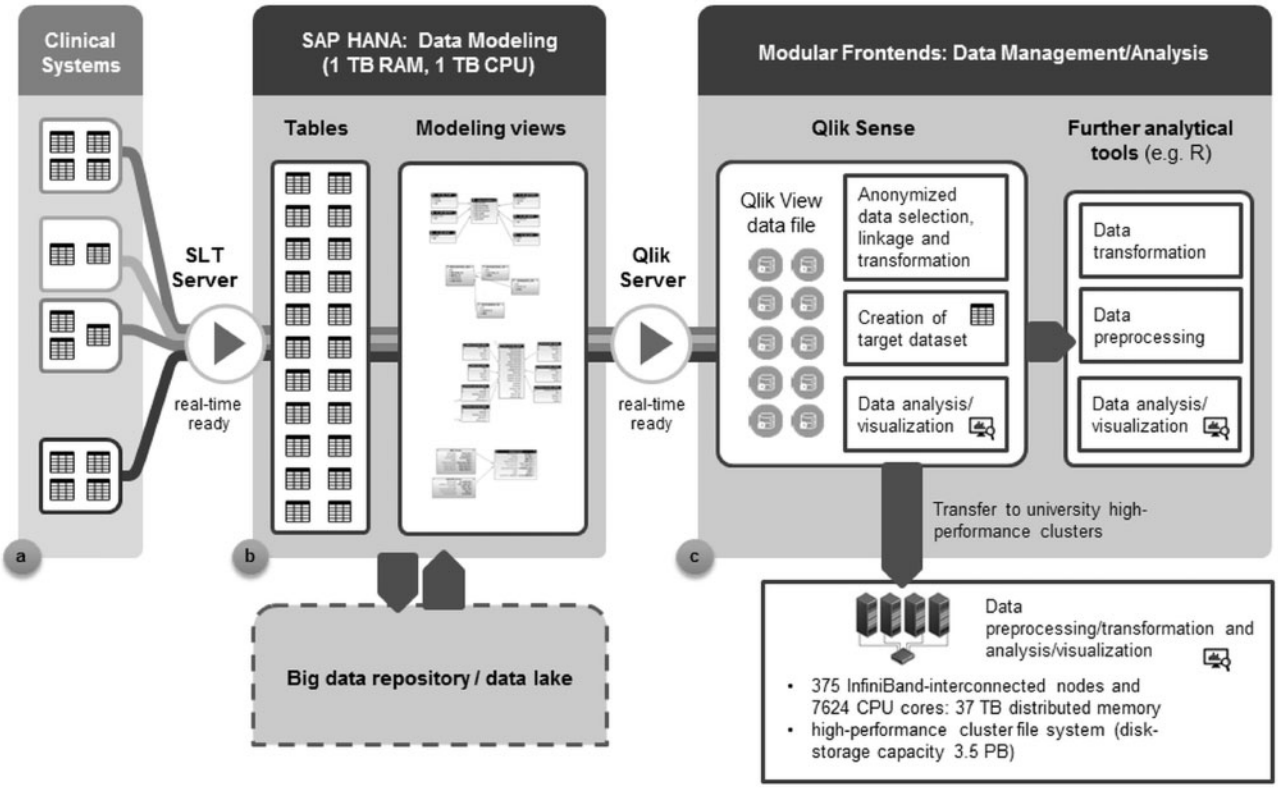

Figure 1. Analytical platform at the University Hospital Basel. [a]Main hospital source systems are connected to the HANA database via SLT servers. Tables are replicated in real-time to HANA 1:1 and are not extracted. [b]SAP® HANA uses a table-based relational database model. Within the HANA database, initial analyses are performed via views. These procedures do not require any data storage and are therefore fast and scalable. At this level, the models are still 1:1 representations and contain basic key figures. Larger amounts of data are projected to be stored in a data lake (big data repository) and can be made accessible via the HANA database. At our institution, only experts in analysis technology have access to SAP® HANA. [c]Via Qlik servers, data are loaded into our frontend tool Qlik Sense®. On this level, more complex data models and reports are generated that may be composed of several systems. The aim is to connect all source systems and model them in a meaningful way for authorized users. Access authorization is controlled with a specific authorization scheme. The data are deidentified at this level. For research purposes, deidentified files can be exported, if needed. CPU, central processing unit; HANA, High Performance Analytic Appliance; PB, petabyte; RAM, random access memory; SLT, SAP Landscape Transformation; TB, terabyte.

Qlik Sense® as HANA frontend application. Unlike our usual business data models, which contain various tables, Qlik Sense® can run advanced data algorithms to create flat tables with each column representing a variable and each row signifying a patient or case.

Analyses of large deidentified datasets are performed with Qlik Sense® or with standard statistical software packages (mainly "R") located on a local Qlik Sense® server or within the secure high-performance computing core facility at the University of Basel (http://scicore.unibas.ch).[17] This research facility maintains a cluster file system (disk-storage capacity; 3.5 petabyte) and a high-performance computing infrastructure providing 37 terabyte of distributed memory.

DISCUSSION

On the basis of a HANA data management system, we described a scalable platform for research-oriented analyses of large HD sets. This platform involves an in-memory database, which enables rapid data linkage within HANA and Qlik Sense® and subsequent data management and analysis steps by use of flexible frontend tools. Furthermore, in specific cases, large HD sets can be analyzed within the university high-performance computing core facility to speed up computationally intensive analyses.

Up to now, little has been published about the architecture of research databases and analytical platforms of health care institutions, which may often rely—at least partly—on proprietary database and source systems for intra-institutional research and not on open-source data models and systems.[18,19] The analytical platform and architecture described here fulfill key elements of clinical research platforms and data warehouses, that is, data protection by deidentification, an effective query interface for cohort discovery, anonymized chart reviews and rapid data extraction and export, if necessary.[15] In contrast to our platform, which includes mainly proprietary components (eg HANA, Qlik Sense®), open-source database systems with underlying common data models (eg "Informatics for Integrating Biology and the Bedside") may be less expensive and more flexible for collaborative health research projects.[6,7,9,20–22] As there is a rapid evolution of common data models and a multitude of accompanying tools,[6,8,9] we did not intend to compare our analytical platform with open-source systems and technologies for data integration and analysis.

Ethical approval is mandatory for all researchers working with HD on our analytical platform. At our institution, only experts in analysis technology have access to the HANA database and the frontend access is centralized due to security and data protection reasons. HD linkage, extraction, and export are performed by the in-house analysis technology team only.

Table 1. HANA data management system, key features, and resources at the University Hospital Basel

	Advantages	Disadvantages
Resources		
Presence of HANA infrastructure	• HANA already available at our institution • No additional acquisition and consulting cost • Handling of large data volumes possible	• Novel technologies and analytical tools may not be compatible with HANA
Presence of an experienced team of developers and data analysts (HANA, Qlik, R)	• Fast development and modeling • Less development and consulting cost • Internal knowledge building and expansion	• N/A
Features		
Source-agnostic data access and integration	• Ability to index and access external data from across the hospital (if needed in real-time)	• N/A
Flexible column- and/or row-based data modeling	• Flexible data modeling • Fast data access and parallel processing (columns) • Efficient data compression (columns) • Generic algorithm pattern to enable column based data structure	• N/A
In-memory computing	• Fast data access • Fast data processing from any data source	• Volatile memory • Expensive additional data storage
Efficient data deidentification layer	• All data can be automatically deidentified (256-bit hash encryption) within Qlik Sense® • Enables research-oriented analyses of large datasets	• N/A
Hybrid approach possible	• HANA may be used together with data lakes (big data repositories) • Lower costs for data storage with hybrid approach compared with HANA only	• High maintenance cost

Abbreviations: HANA: High Performance Analytic Appliance; N/A: not applicable.

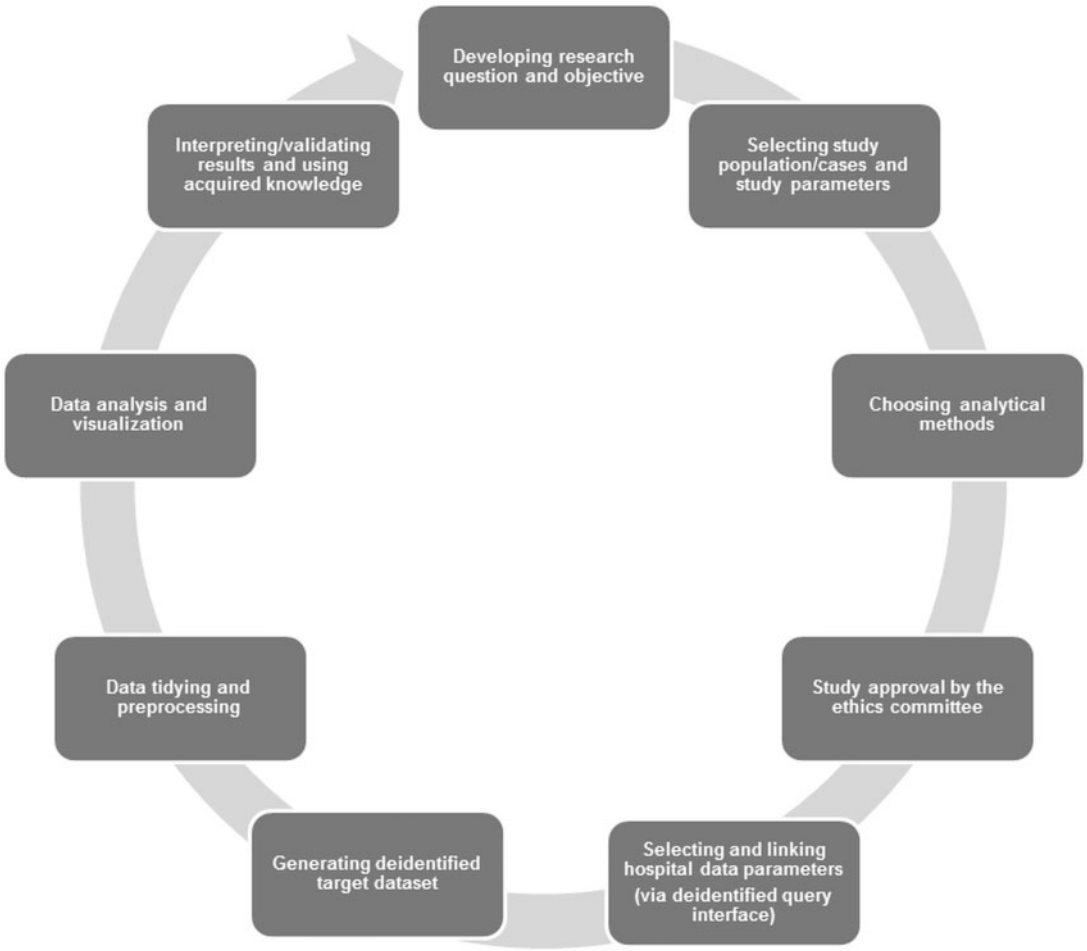

Figure 2. Data management and analysis process at the University Hospital Basel.

With rising expectations and needs for research-oriented analyses on large HD, our analytical platform is continually evolving on the HANA as well as on the frontend level. In regard to data linkage and representation, various data models have been implemented for HANA and Qlik Sense®, since selection of appropriate data models and representations depends largely on the specific research question and the required analysis method. For instance, some machine learning tasks may require plain data files, whereas for other analyses, direct access of the relational database may be more appropriate. In the latter case, HANA may be directly accessed via R on HANA to parallelize and speed up computing-intensive operations. In general, we observed that the data linkage and generation of large target datasets can be readily achieved via Qlik Sense® and that most research-oriented analyses do not require direct access to HANA modules. If required, we export deidentified datasets for subsequent analysis; in this case, the chosen data format depends on the size of the dataset and the specific tools used for further analyses.

Developing and maintaining an analytical platform, which may involve database systems, data warehouses, data lakes, in-memory technologies—or combinations of it—is an ongoing and highly dynamic process.[12,14] Although the importance of unstructured data is increasing in biomedical research, fundamental principles of HD for research purposes will most likely remain[23]:

1. Subject oriented: data are mostly represented on a case/subject level.
2. Integrated: data are gathered and merged from a variety of sources.
3. Time-variant: data are tagged with a "time stamp"; this permits exact re-execution of data queries made at a specific point in time.
4. Non-volatile: data are stable; more data are added but is not removed on a regular basis.

In the framework of an ongoing personalized health initiative in Switzerland, our multipurpose analytical platform with a deidentification layer may serve on the frontend level as a local platform for the nationwide anonymized health-related data exchange for research purposes as mandated by the Swiss government; however, this may also be achieved with other integrated systems.

In conclusion, we describe an advanced analytical platform for research-oriented analyses of HD derived from an innovative business intelligence architecture. This platform involves an in-memory database management system for data modeling as well as an external high-performance computing cluster for computing-intensive analytical tasks. Setting up platforms for research-oriented analyses is a highly dynamic, time-consuming and costly process and database systems are evolving rapidly. In some health care institutions, research platforms may be derived from existing business intelligence systems.

ACKNOWLEDGMENTS

PATREC Study Group: Christian Abshagen (University Hospital Basel, Basel, Switzerland); Geoffrey Fucile (University of Basel, Basel, Switzerland); Martin Gerber (University Hospital Basel, Basel, Switzerland); Nicole Goebel (University Hospital Basel, Basel, Switzerland); Balthasar L Hug (Kantonsspital Luzern, Lucerne, Switzerland); Bernd Jaegle (University Hospital Basel, Basel, Switzerland); Eleonore Kuenzel-Pawlik (University Hospital Basel, Basel, Switzerland); Simon Neubauer (University Hospital Basel, Basel, Switzerland); Rakesh Padiyath (University Hospital Basel, Basel, Switzerland); Jan A Roth (University Hospital Basel, Basel, Switzerland); Thomas Sakoparnig (University of Basel, Basel, Switzerland); Thierry Sengstag (University of Basel, Basel, Switzerland); Damian Spyra (University Hospital Basel, Basel, Switzerland); and Andreas F. Widmer (University Hospital Basel, Basel, Switzerland).

AUTHOR CONTRIBUTION

J.A.R., N.G., and B.L.H. contributed to the case report conception. All authors contributed to drafting the manuscript. All authors reviewed and approved the final manuscript.

REFERENCES

1. Jarow JP, LaVange L, Woodcock J. Multidimensional evidence generation and FDA regulatory decision making: defining and using "real-world" data. *JAMA* 2017; 318 (8): 703–4.
2. Beeler PE, Bates DW, Hug BL. Clinical decision support systems. *Swiss Med Wkly* 2014; 144: w14073.
3. Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: the dawn of precision delivery. *JAMA* 2016; 315 (7): 651–2.
4. Cresswell KM, Bates DW, Sheikh A. Ten key considerations for the successful optimization of large-scale health information technology. *J Am Med Inform Assoc* 2017; 24 (1): 182–7.
5. Amarasingham R, Audet AM, Bates DW, *et al*. Consensus statement on electronic health predictive analytics: a guiding framework to address challenges. *EGEMS (Wash DC)* 2016; 4 (1): 3.
6. Hripcsak G, Duke JD, Shah NH, *et al*. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–8.
7. Bouzille G, Jouhet V, Turlin B, *et al*. Integrating biobank data into a clinical data research network: the IBCB project. *Stud Health Technol Inform* 2018; 247: 16–20.
8. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. *J Am Med Inform Assoc* 2016; 23 (5): 909–15.
9. Maier C, Lang L, Storf H, *et al*. Towards implementation of OMOP in a German university hospital consortium. *Appl Clin Inform* 2018; 9 (1): 54–61.
10. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012; 19 (1): 54–60.
11. Gaye A, Marcon Y, Isaeva J, *et al*. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014; 43 (6): 1929–44.
12. Horton I, Lin Y, Reed G, Wiepert M, Hart S. Empowering Mayo Clinic individualized medicine with genomic data warehousing. *J Pers Med* 2017; 7 (3): 7.
13. Kreuzthaler M, Martinez-Costa C, Kaiser P, Schulz S. Semantic technologies for re-use of clinical routine data. *Stud Health Technol Inform* 2017; 236: 24–31.
14. Firnkorn D, Knaup-Gregori P, Lorenzo Bermejo J, Ganzinger M. Alignment of high-throughput sequencing data inside in-memory databases. *Stud Health Technol Inform* 2014; 205: 476–80.
15. Shin SY, Kim WS, Lee JH. Characteristics desired in clinical data warehouse for biomedical research. *Healthc Inform Res* 2014; 20 (2): 109–16.
16. Bate A, Lindquist M, Edwards IR. The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database. *Fundam Clin Pharmacol* 2008; 22 (2): 127–40.
17. Stockinger H, Altenhoff AM, Arnold K, *et al*. Fifteen years SIB Swiss Institute of Bioinformatics: life science databases, tools and support. *Nucleic Acids Res* 2014; 42 (W1): W436–41.
18. Kamal J, Liu J, Ostrander M, *et al*. Information warehouse—a comprehensive informatics platform for business, clinical, and research applications. *AMIA Annu Symp Proc* 2010; 2010: 452–6.
19. Danciu I, Cowan JD, Basford M, *et al*. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform* 2014; 52: 28–35.

20. Jiang G, Kiefer RC, Sharma DK, Prud'hommeaux E, Solbrig HR. A consensus-based approach for harmonizing the OHDSI common data model with HL7 FHIR. *Stud Health Technol Inform* 2017; 245: 887–91.
21. Turley CB, Obeid J, Larsen R, *et al*. Leveraging a statewide clinical data warehouse to expand boundaries of the learning health system. *EGEMS (Wash DC)* 2016; 4 (1): 1245.
22. Ross TR, Ng D, Brown JS, *et al*. The HMO research network virtual data warehouse: a public data model to support collaboration. *EGEMS (Wash DC)* 2014; 2 (1): 1049.
23. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *J Am Med Inform Assoc* 2010; 17 (2): 131–5.