German Medical Data Sciences: Visions and Bridges
R. Röhrig et al. (Eds.)

doi:10.3233/978-1-61499-808-2-85

Implementing a Data Management Platform for Longitudinal Health Research

Jan-Patrick WEIẞ[a,1], Ursula HÜBNER[a], Jens RAUCH[a], Jens HÜSERS[a],
Frank TEUTEBERG[b], Moritz ESDAR[a], Jan-David LIEBE[a]
[a] *Health Informatics Research Group, Osnabrück University AS, Germany*
[b] *Research Group Accounting & Information Systems, University Osnabrück, Germany*

Abstract. Health IT adoption research is rooted in Rogers' Diffusion of Innovation theory, which is based on longitudinal analyses. However, many studies in this field use cross-sectional designs. The aim of this study therefore was to design and implement a system to (i) consolidate survey data sets originating from different years (ii) integrate additional secondary data and (iii) query and statistically analyse these longitudinal data. Our system design comprises a 5-tier-architecture that embraces tiers for data capture, data representation, logics, presentation and integration. In order to historicize data properly and to separate data storage from data analytics a data vault schema was implemented. This approach allows the flexible integration of heterogeneous data sets and the selection of comparable items. Data analysis is prepared by compiling data in data marts and performed by R and related tools. IT Report Healthcare data from 2011, 2013 and 2017 could be loaded, analysed and combined with secondary longitudinal data.

Keywords. Survey data, longitudinal analyses, health IT adoption research

1. Introduction

Health IT adoption research is rooted in the work about the adoption and diffusion of innovation and draws on Rogers' Diffusion of Innovation (DOI) theory [1]. Health IT adoption studies often make use of surveying techniques to obtain the necessary information about IT adoption rates from samples of healthcare organisations and hereby, design their work as cross-sectional studies [2]. However, cross-sectional studies are only snapshots in time and therefore not suitable to study trends as the DOI theory requests. At the same time, longitudinal studies, which are more appropriate to answer these questions, demand more resources and are therefore underrepresented in the health IT adoption research [2,3]. IT Report Healthcare is a regularly conducted survey with a focus on measuring IT adoption in Germany, but also in Austria [4], the Netherlands and Switzerland that was developed in accordance with the OECD eHealth benchmark. IT Report Healthcare surveys like most other repetitive surveys, which capture information in a highly agile environment, face the same kind of issues: 1) new IT developments, 2) changing context, e.g. regulations, 3) changes of the statistical unit, e.g. merging of organisations, 4) availability of new data sources, 5) new research questions that are evolving from previous research findings. Thus, there is the need to

[1] Corresponding Author, Jan-Patrick Weiß, Osnabrück University of AS, Health Informatics Research Group, PO Box 1940, 49009 Osnabrück, Germany; E-Mail: j.p.weiss@hs-osnabrueck.de

encapsulate data from each point of time, to easily identify comparable items and to link only these data across time. It is the overall goal of the project to develop, implement and continuously improve an integrated platform for the management, analysis and visualisation of research data in IT adoption studies, but also in health services research. This part of the project focusses on the overall architecture and data management for longitudinal research. The aim of this study therefore was to design and implement a system built on open source components to (i) consolidate different items of one survey which was conducted in different years into one database, (ii) integrate additional secondary data sources and (iii) query and statistically analyse data over multiple years for longitudinal analyses.

2. State of the art

Data warehouse systems face the challenge of integrating several isolated information repositories into one single logical repository [5]. A literature and internet search, which was performed prior to the developments, resulted in no publications of a system that fulfilled the requirements as stated above. There were a few approaches to design and implement data warehouse systems for using survey-based data [6,7]. These approaches have a fixed concept in which the survey file formats, the survey structure and the user requirements stay the same and therefore are not flexible enough for iterative and agile research processes. IT adoption studies [8] often do not refer to any issues of data management but focus on data analytics primarily.

3. Concept

In order to meet these requirements, a hybrid approach for designing and implementing the system was chosen [9]. We combined a supply driven approach [5], in which we identified and analysed the data available, with a demand driven approach [10], in which we determined the requested information from users according to previous IT adoptions studies [11,12].

Our system design comprises a 5-tier-architecture that embraces a *data capture tier*, a *data tier*, a *logic tier* and a *presentation tier* (Fig. 1). They are connected via a fifth tier, the *data integration bus*. The *data capture tier* describes instruments from which the relevant data originates. The *data tier* is represented by a data warehouse for data consolidation and storage consisting of three layers: in the *source layer,* each of the data sources provided by the *data capture tier* is mirrored as a relational database table to ensure compliance with the defined data schema and data constraints. All extraction, transformation and loading (ETL) processes are implemented and executed through the *data integration bus*.

Primary data from different survey datasets and secondary data (e.g. hospital quality reports) are loaded into a consolidated form within the *core layer*. The core layer within the data tier constitutes the centre of data management and lays the foundation for flexible and longitudinal analyses. This is achieved via a schema that is based on the data vault model [13]. This schema is uncoupled from the model of the source layer to ensure flexibility. The data vault schema consists of three types of tables (hub, link, satellite). A hub represents a real-world or abstract object (e.g. survey item, site, quality indicator), which can be uniquely identified by its natural key. Each hub object receives a technical primary key and load timestamp for historicizing within the ETL process.

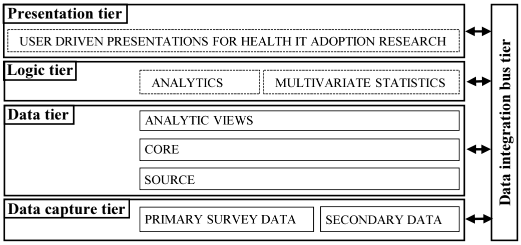

Figure 1. Overview of the system architecture

Links model the relationships between hubs (e.g. item – response, survey – site, site – response – quality indicator). A satellite contains attributes of one hub or one link. To historicize multiple temporal versions of attribute values its primary key is the foreign key of one hub/link with a timestamp. By separating natural keys (e.g. item codes, site ID) and the relationships between them from their attributes, a data schema is created that can, in comparison to the classical relational model [14], historicize various objects from different years and is able to combine these various objects in a flexible way, and to provide data for different, frequently changing query requirements.

Data marts, which constitute the analytics view layer in the data tier, provide optimised views with automatic aggregated data for predefined analyses. To perform more complex calculations, the data is loaded into further tools of the *logic tier* and the relevant results are stored back in additional data marts. This exchange service is represented via the *data integration bus*. In the *presentation tier* the data is displayed for standardised regular reports or for further research.

In contrast to other approaches, this concept does not focus on the integration of some specific types of data for certain use cases (e.g. clinical data [15], patient data [16]) but rather aims at the scalability through the data vault model approach for persistent, historicized storage of multiple surveys from different years.

4. Implementations

Pentaho Data Integration 7 served as the central *data integration bus* for all ETL jobs. Primary survey data were extracted from LimeSurvey 2.54.3 or legacy SPSS survey files from previous years. Secondary data was extracted from publicly available data sources (hospital quality reports, demographic hospital data). All data sources were loaded into relational database tables into the source layer (PostgreSQL 9.6 on Ubuntu Server 16.04). Then the data sources were transformed to one unified, consolidated, physical data vault schema (Fig. 2).

The data warehouse contains surveys of IT Report Healthcare from 2011, 2013 and 2017 (2011: 339 datasets, 203 items: 2013: 259 datasets, 521 items; 2017: 283 datasets, 226 items), historical demographic data of German hospitals from 2003 to 2014 (2883 datasets, 63 attributes) and hospital quality reports from 2012 to 2014 (381 quality indicators). In the design process, synonymously named items are mapped onto one item entity and stored in the hub *h_item*. Descriptive properties from the data sources, from which these items originated, are historicized in the satellites. Data marts were created to provide aggregated data e.g. denormalised demographic data of survey respondents or item frequency tables for longitudinal analysis. Data marts were accessed by the statistical software R 3.3.2 for data analysis and visualisation. Same survey items over the years 2011, 2013 and 2017 are provided by one data mart as frequency tables and were then further processed and visualised in R using the package *ggplot2 2.2.1* [17] (Fig. 3).

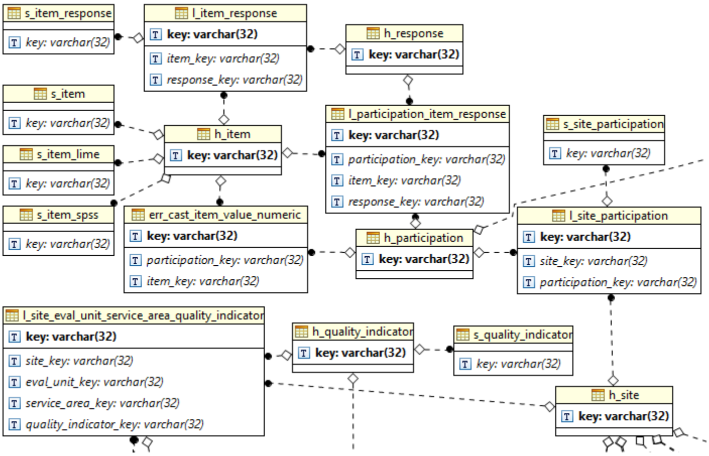

Figure 2. Excerpt of the deployed data vault model centered on the survey structure

5. Lessons learned

The proposed system for collecting, processing, storing and analysing data is composed of separate components, which are connected by ETL jobs. Each component can be replaced without affecting the rest of the system – except for the need of designing new ETL jobs. All tools used are open source and freely accessible. All relevant datasets could be loaded into the source layer. Using the proposed data vault model changes in the structure of the data source will lead to no changes in the structure of the model. Existing data can thus be easily extended by adding additional satellites, hubs or links.

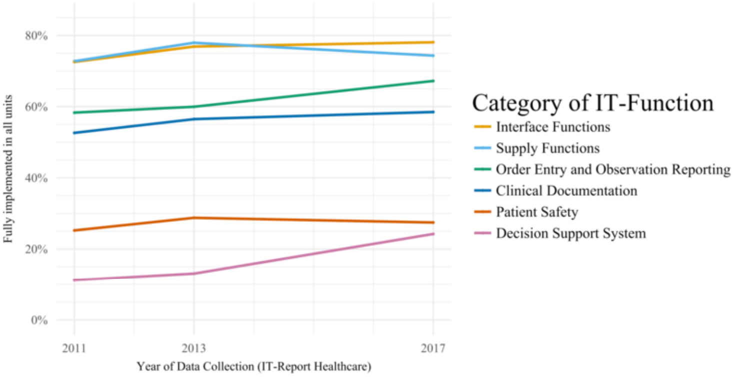

Figure 3. Display of IT adoption rates of six types IT applications for the years 2011, 2013 and 2017

6. Conclusion

We designed and implemented a system that improves data management for longitudinal analyses via the data vault model approach for persistent, historicized storage of datasets. It thus allows for agile, longitudinal analyses of heterogeneous datasets and lays the foundation of more rigorous and theory oriented studies of health IT adoption.

7. Conflict of Interest

The authors state that they have no conflict of interests.

Acknowledgment

This study is funded by Lower Saxony Germany (grants: ZN 3103, ZN 3062).

References

[1] Rogers EM. Diffusion of innovations. 5th ed. New York, London, Toronto, Sydney: Free Press, 2003.
[2] Agarwal R, Gao G, DesRoches C, Jha AK. Research Commentary —The Digital Transformation of Healthcare: Current Status and the Road Ahead. Inf. Syst. Res. **21** (2010), 796–809.
[3] Jones SS, Rudin RS, Perry T, Shekelle PG. Health information technology: an updated systematic review with a focus on meaningful use. Ann. Intern. Med. **160** (2014) 48–54.
[4] Hüsers J, Hübner U, Esdar M, Ammenwerth E, Hackl WO, Naumann L, Liebe JD. Innovative Power of Health Care Organisations Affects IT Adoption: A bi-National Health IT Benchmark Comparing Austria and Germany. J Med Syst. **41** (2017), 33. DOI:10.1007/s10916-016-0671-6.
[5] Inmon WH. Building the data warehouse. 4th ed. Indianapolis Ind.: Wiley, 2005.
[6] Seah BK, Ezam Selan N. Design and Implementation of Data Warehouse with Data Model using Survey-based Services Data. In: INTECH 2014. Piscataway, NJ: IEEE; (2014), 58–64.
[7] Yost M, Nealon J. Using a dimensional data warehouse to standardize survey and census metadata. National Agricultural Statistics Service, U.S. Department of Agriculture, 1999.
[8] Buntin MB, Burke MF, Hoaglin MC, Blumenthal D. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health aff. **30** (2011), 464–71.
[9] Romero O, Abelló A. A Survey of Multidimensional Modeling Methodologies. IJDWM **5** (2009), 1–23.
[10] Kimball R, Ross M. The data warehouse toolkit: The complete guide to dimensional modeling. 2nd ed. New York: Wiley, 2002.
[11] Liebe J, Hüsers J, Hübner U. Investigating the roots of successful IT adoption processes - an empirical study exploring the shared awareness-knowledge of Directors of Nursing and Chief Information Officers. BMC Med. Inform. Decis. Mak. **16** (2016), 10.
[12] Liebe JD, Hübner U, Straede MC, Thye J. Developing a Workflow Composite Score to Measure Clinical Information Logistics. Methods Inf. Med. **54** (2015), 424–33.
[13] Golfarelli M, Graziani S, Rizzi S. Starry Vault: Automating Multidimensional Modeling from Data Vaults. In: Pokorný J, Ivanović M, Thalheim B, Šaloun P, editors. Advances in Databases and Information Systems. Lecture Notes in Comput. Sci. Cham: Springer Int. Publishing; (2016), 137–51.
[14] Codd EF. A relational model of data for large shared data banks. Commun. ACM **13** (1970), 377–87.
[15] Gui H, Zheng R, Ma C, Fan H, Xu L. An Architecture for Healthcare Big Data Management and Analysis. In: Yin X et al. editors. HIS 2016, Shanghai, China, November 5-7, 2016, Proceedings. Lecture Notes in Comput. Sci. Vol 10038. Cham, s.l.: Springer International Publishing; (2016), 154–60.
[16] Kaspar M, Fette G, Ertl M, Dietrich G, Nagler N, Störk S, et al. Extraktion und Transfer patientenbezogener Daten aus klinischen Informationssystemen in Studiendatenbanken – effektive Unterstützung klinisch-epidemiologischer Forschung durch ein Data Warehouse: GMS Publishing House, 2015.
[17] Wickham H, Chang W. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics, 2016.