dHealth 2023
B. Pfeifer et al. (Eds.)

doi:10.3233/SHTI230036

Making a Virtue of Necessity - A Highly Structured Clinical Data Warehouse as the Source of Assured Truth in a Hospital

Werner O. HACKL[a,1], Sabrina B NEURURER[a,b], Marco SCHWEITZER[a] and Bernhard PFEIFER[a,b]
[a] *Division for Digital Medicine and Telehealth, UMIT TIROL - Private University for Health Sciences and Health Technology, Hall in Tirol, Austria*
[b] *Tyrolean Federal Institute for Integrated Care, Tirol Kliniken GmbH, Innsbruck, Austria*

Abstract. Data-driven decision-making in health care is becoming increasingly important in daily clinical use. A data warehouse, storing all the clinically relevant information in a highly structured way, is a primary basis for achieving this goal. We are developing a clinical data warehouse where more than 20 years of clinical data can be persisted, and newly generated data from different sources can be integrated. A back room was created to store all hospital information system data in a PostgreSQL database. Due to the enormous number of diverse forms in the hospital information system, a broker service was developed that integrates the individual data sources into the data warehouse as soon as they are released for storage. The front room represents the interface from the infrastructure to the targeted analysis. Database query and visualization tools or business intelligence tools can display and analyze processed and interleaved data. In all areas of business and medicine, structured and quality-adjusted data is of major importance. With the help of a clinical data warehouse system, it is possible to perform patient-centered analyses and thus realize optimal therapy. Furthermore, it is possible to provide staff and management with dashboards for control purposes.

Keywords. Clinical data warehouse, hospital information system, business intelligence, digitalization

1. Introduction

Many hospitals face the major challenge of replacing their central hospital information system (HIS), which may be due to various technological and organizational reasons. The current systems are already several years or even decades in place. However, over time, the market for HIS changed significantly. Some major vendors disappeared, were bought up or merged, and whole product lines were taken from the market or merged with other products. There have been various technological innovations, including advances in terminologies and communication as well as interoperability standards.

In addition to these external drivers of change, the requirements for a HIS have also changed fundamentally. Whereas documentation tasks, information transfer and

[1] Corresponding Author: Werner Hackl, UMIT TIROL, Hall in Tirol, Austria, E-Mail: werner.hackl@umit-tirol.at

exchange, service documentation, or legal protection used to be the leading interest for physicians and hospital operators in the past, today, the patient and the patient's needs are placed at the center. The trend is clearly moving towards patient-centered knowledge generation, trans-institutional information sharing, intelligent clinical data analytics capabilities, artificial intelligence, machine learning, and support of informed decision-making [1,2]. In addition, special patient portals that enable patients to enter data into or retrieve information from clinical documentation systems themselves are finding their way into everyday clinical practice [3].

Since hospital information systems are the backbone of a complex information logistics process, exchanging a central application system for another product is complicated. One particularly important aspect that must be ensured in this context is preserving data integrity and preventing information loss. Because complete data transfers from old to new systems are particularly complex and time-consuming, they are often avoided.

However, in any case, it must be ensured that the statutory retention periods for clinical data and documents are complied with. In Austria, for example, clinical records for inpatient stays must be kept for at least 30 years, and for outpatient visits for at least ten years. An easy way - and often used - is to save the data and documents from the legacy system in PDF format and store them in a long-term archive. So, data is preserved even if the legacy system is not available anymore, but this causes a loss of structure in the sense of machine-readable data and a substantial loss of metadata.

An excellent way to get around these problems is to export the data from the legacy systems, get as much metadata as possible, and then store everything in a structured way in an integrated clinical data repository. The concept of data warehousing [4] is ideally suited for this purpose [5]. A clinical data warehouse (CDW) is a centralized repository of patient data to support clinical research and decision-making [6]. It is designed to collect, store, and manage large amounts of data from multiple sources, such as electronic health records (EHRs), laboratory information systems, and medical imaging modalities. The data needs to be organized and integrated in a manner that supports professionals, researchers as well as administrative personnel. By providing a single source of truth and a unified view of clinical and patient data, warehouses help improve patient care and support evidence-based medicine [6].

A CDW usually includes demographic information, medical history, lab results, medication information, and other clinical data from different sources. The data is cleaned, transformed, and integrated into a standardized format. Thereafter, the data is organized into logical data models, such as patient encounters, lab results, or medication administration, which reflect the clinical domain [7,8].

Utilizing the data stored in a CDW offers numerous benefits. One key advantage is its ability to facilitate clinical research by providing a large and diverse dataset for analysis. Another advantage is the improvement of patient care, which is accomplished by providing health professionals with a thorough understanding of a patient's medical history, lab findings, and medication information. The data stored in a CDW can also support population health management by revealing patterns and trends in patient data that inform healthcare policy and decision-making. Next, using clinical data in a data warehouse can also contribute to improving financial performance by providing administrators with real-time data on utilization, costs, and outcomes. Finally, the data in a CDW can also be used to generate reports, dashboards, and alerts. These can be used to monitor clinical quality, identify areas for improvement, and track performance against benchmarks.

Overall, a CDW can be a powerful tool for healthcare organizations, providing the ability to extract insights from data and make data-driven decisions to improve patient care, clinical research, and healthcare operations. Nevertheless, at least in German-speaking countries, few hospitals have implemented their own clinical data warehouses. This may have several reasons. For example, such projects have high complexity and require many resources. Often, clinical data are also not readily available for secondary use purposes. And in many cases, related data warehouse projects are motivated more by business and less by clinical, medical, or nursing considerations.

In the present case in the Tyrolean federal hospitals, a highly developed business warehouse with highly structured data warehousing processes has existed for many years. While numerous analyses are generated from the clinical information system for secondary use purposes, no formal data warehouse process is defined for clinical data, nor is there – except for the nursing sector, which has operated its own nursing data mart – a comprehensive CDW.

It was, therefore, obvious to seize the opportunity, to make a virtue of necessity and develop – as the objective of the present work – a clinical data warehouse during the replacement of the hospital information system.

2. Materials and Methods

The Tyrolean federal hospitals (Tirol Kliniken GmbH), an association of several hospitals and the biggest healthcare provider in Tyrol/Austria, are currently working on replacing their "old" HIS (Cerner Millennium®) system with a new one (Meona®). As there are more than 20 years of clinical data persisted in the Cerner system [9], it seems impossible to migrate all data into the new HIS. The PDF approach described above was chosen to comply with the legal retention obligation. To keep legacy data accessible in a highly structured format, to address clinical questions, and to enable clinical research for future questions, a data warehousing approach according to the SPIRIT framework [5] was chosen, and a corresponding project was launched in January 2022.

This project aims to develop a data warehouse system that consolidates all structured data from the legacy Cerner system by June 2023 and will be capable of retrieving data from the new Meona HIS afteronce the migration is complete. Furthermore, integration with other source systems (e.g., patient administration and patient data management systems, laboratory information systems, and other specialized data systems) will be realized.

A requirement for the project was to rely mainly on open-source software. For the extraction, transformation, cleansing, and loading (ETL / ETcL) operations, the Talend Open Studio software package implemented in Java was used.

Since direct access to the databases and the underlying data model of the Cerner system's databases was not feasible for the project team, it was decided to use the existing data export options, export the clinical data as flat CSV files from the Cerner system and then transform them accordingly and load them into the data warehouse. For that, a metadata model based on the clinical documentation in the Cerner system (e.g., the structure of the documentation forms) was implemented to avoid losing structural information.

To ensure a comprehensive view of the clinical context, it was planned to also transfer data from the existing SAP® business warehouse, such as patient master data, information on inpatient stays or day-care stays, outpatient cases and movements, diagnoses, procedures, orders, and services, to the clinical data warehouse. It was also planned to map the chronological course and performance of diagnostics and therapy to transfer data from scheduling and tracking data into the CDW. The import of documents was waived, but it was planned to integrate corresponding links to the storage locations in the multimedia archives into the CDW. However, to ensure fast and direct evaluations, it was designed to store texts from documents in the CDW. In addition, it was also planned to transfer data from the existing nursing data mart to be able to ensure appropriate continuity.

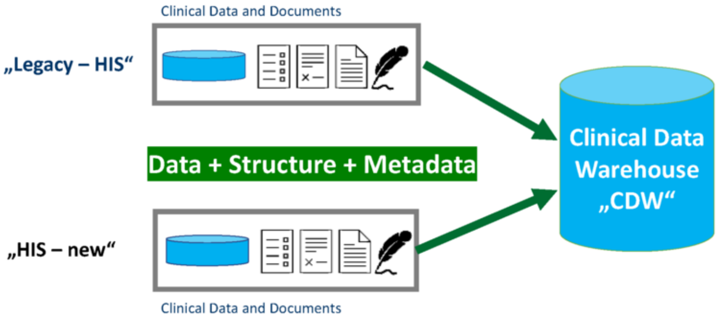
Clinical Data and Documents
„Legacy – HIS"
Data + Structure + Metadata
„HIS – new"
Clinical Data and Documents
Clinical Data Warehouse „CDW"

Figure 1: Overall structure describing the data flow from the legacy and new hospital information system to the clinical data warehouse.

Figure 1 illustrates how the two hospital information systems are mapped into the clinical data warehouse. An essential part is integrating the data and the underlying structure, as well as the descriptive information and metadata.

3. Results

The developed data warehouse system is based on the data warehouse meta-model according to [10] and can be divided into three parts (cf. Figure 2). The first part, data supply, contains the data sources, which include the hospital information systems, the billing system (SAP®), the laboratory information system, and other legacy systems.

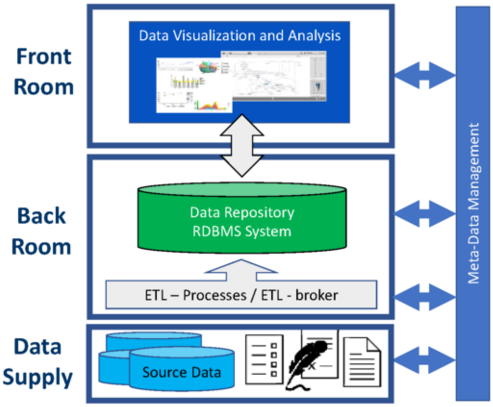
Front Room
Data Visualization and Analysis
Back Room
Data Repository RDBMS System
ETL – Processes / ETL - broker
Data Supply
Source Data
Meta-Data Management

Figure 2: Overall structure of the data warehouse system

The second part, or the so-called backroom, consists of a software component called ETL Broker, which ensures that the data integration jobs are started as soon as new information is available. The data sources of any type must ensure that the information to be integrated into the data warehouse is moved to the broker service. After being received by the broker, the relevant and associated data integration processes are started. The integration of the available data in the broker is facilitated by using ETL processes generated using Talend Open Studio software. These processes are generated in the form of specific integration modules written in Java programming language and are deployed to the broker for executing their assigned task.

While the ETL processes are executed in the broker service, a log file is created, and data that fails to import is stored in a designated area for later review and correction by an expert. The object-relational PostgreSQL version 15 is used as the base database system (RDBMS). PostgreSQL is deemed particularly suitable for creating the data warehouse in this case, as it is both open-source and capable of handling high volumes of data.

The third area is the so-called front room. This area is meant for the users to access the data warehouse's data and perform analyses. This process starts with a query submitted to the database system in SQL, followed by processing the result set with a business intelligence tool or a statistics program. Tableau® was used in the first version, but other software tools, like Qlik Sense® or Microsoft Power BI®, are currently being tested and evaluated.

Metadata management ensures that all data elements are described accordingly and can be found in the data warehouse. Since not every user can access all data in the system, the data warehouse administrators can create specific data marts for the respective areas. Furthermore, user-specific access can be managed through an authorization scheme.

4. Discussion and Outlook

With the increasing use of electronic health records and other clinical data systems, healthcare organizations face the challenge of managing massive amounts of data. The need to replace key components of clinical information systems poses a significant risk of information loss for organizations because data transfer from one system to another is a highly complex matter. An excellent way to circumvent this is using data warehousing, as we have been able to show using the example of the Tyrolean federal hospitals.

Clinical data warehouses will further gain importance, as they are powerful tools for improving patient care, advancing medical research, and supporting quality improvement initiatives by providing structured long-term data. Advances in technology, such as cloud computing and artificial intelligence, are expected to play a vital role in the evolution of clinical data warehouses, making it possible to store, manage, and analyze even larger amounts of data to improve patient outcomes and support the overall healthcare system.

However, it is also important to remember that all data stored in the various systems was collected for specific purposes and that data quality is a critical quality condition for all reuse purposes. Another potential issue is the existence and handling of unstructured data. In many cases, extracting and restoring the data in the desired structure is difficult or impossible, which may lead to the loss of information. Therefore, such systems must focus on keeping the stored data quality high to maximize the benefits of evaluations and analyses.

In conclusion, a clinical data warehouse is critical to a modern healthcare organization's infrastructure. It enables healthcare providers to make informed decisions about patient care and organizations to improve the health of their patients through data-driven insights and targeted interventions.

5. References

[1] W.O. Hackl, and A. Hoerbst, Clinical Information Systems Research in the Pandemic Year 2020, *Yearb Med Inform*. **30** (2021). doi:10.1055/s-0041-1726516.

[2] W.O. Hackl, and A. Hoerbst, On the Way to Close the Loop in Information Logistics: Data from the Patient - Value for the Patient, *Yearb Med Inform*. **27** (2018) 91–97. doi:10.1055/s-0038-1667076.

[3] R. Dendere, C. Slade, A. Burton-Jones, C. Sullivan, A. Staib, and M. Janda, Patient Portals Facilitating Engagement With Inpatient Electronic Medical Records: A Systematic Review, *J Med Internet Res*. **21** (2019). doi:10.2196/12779.

[4] R. Kimball, and M. Ross, The Data Warehouse toolkit, Wiley Publishing, 2013.

[5] W.O. Hackl, and E. Ammenwerth, SPIRIT: Systematic planning of intelligent reuse of integrated clinical routine data: A conceptual best-practice framework and procedure model, *Methods Inf Med*. **55** (2016) 114–124. doi:10.3414/ME15-01-0045.

[6] F.J. Martin-Sanchez, V. Aguiar-Pulido, G.H. Lopez-Campos, N. Peek, and L. Sacchi, Secondary Use and Analysis of Big Data Collected for Patient Care, *Yearb Med Inform*. **26** (2017) 28–37. doi:10.15265/IY-2017-008/ID/JR008-64.

[7] B. Pfeifer, J. Aschaber, C. Baumgartner, S. Dreiseitl, R. Modre, G. Schreier, and B. Tilg, A Life Science Data Warehouse System to enable Systems Biology in Prostate Cancer, in: S. Cohen-Boulakia, and V. Tannen (Eds.), Data Integration in the Life Sciences, 4th International Workshop, Philadelphia, 2007: pp. 9-ff.

[8] B. Pfeifer, J. Aschaber, C. Baumgartner, S. Dreiseitl, R. Modre, G. Schreier, and B. Tilg, A data warehouse for prostate cancer biomarker discovery, in: H.R. Arabnia, M.Q. Yang, and J.Y. Yang (Eds.), International Conference on Bioinformatics & Computational Biology, BIOCOMP 2007, Volume II, June 25-28, 2007, Las Vegas Nevada, USA, CSREA Press, 2007: pp. 323–327.

[9] G. Lechleitner, K.P. Pfeiffer, I. Wilhelmy, and M. Ball, Cerner Millennium: the Innsbruck experience, *Methods Inf Med*. **42** (2003) 8–15. doi:10.1055/s-0038-1634204.

[10] A. Bauer, and H. Günzel, Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung, 4th ed., dpunkt.verlag, Heidelberg, 2013.