Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Design of a clinical database to support research purposes: Challenges and solutions

Halima Samra [1, 2, *], Alice Li [3], Ben Soh [1]

[1]Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia
[2]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[3]La Trobe Business School, La Trobe University, Melbourne, Australia

ARTICLE INFO

Article history:
Received 27 August 2020
Received in revised form
4 November 2020
Accepted 5 November 2020

Keywords:
Databases and information systems
Relational databases
Clinical research databases
Clinical research information systems
Data modeling

ABSTRACT

The aim of this paper is to propose solutions to challenges faced by database systems for clinical research purposes. Current clinical databases are primarily based on data acquisition for healthcare intentions. However, these healthcare databases lack the data analysis capability for clinical researchers. In order for clinical researchers to use the healthcare databases in an effective manner, such as in their clinical trial studies, challenges of data integration, data storage, and data retrieval in the current healthcare database settings need to be overcome. Our proposed solutions include using: 1) NoSQL to efficiently integrate clinical databases with legacy healthcare databases, (2) entity attribute value model for data retrieval, and (3) warehouse for big data storage.

1. Introduction

Advances in information and communications technology (ICT) have reformed health information systems (HISs) and influenced the way data collected, processed, stored, and retrieved (Lippeveld et al., 2000). It has resulted in the generation of an enormous amount of data that vary in type and structure according to its sources and methods of collection (Raghupathi and Raghupathi, 2014). For maximum benefit, these data should be effectively collected, organized, integrated, and stored in a shareable and accessible way. Usually, data in clinical research databases were collected either as part of the patient care process or extracted from patient medical records. The research data were then stored in databases where it can be organized and operated by the database management system (DBMS), which is specialized software used to optimize and manage the process of data storage and retrieval (Collen, 1990). In order for a database to serve multiple purposes, the content and description of the database should be comprehensively covered (Wiederhold, 2012). Numerous integration efforts have been made, such as using approaches to integrate terminologies, ontologies, and schema matching (Brazhnik, 2007). The process of data integration requires combining scientific methods and specifications needed to be stored in a database. Interoperability provides more options for integration. In order to gather further insight into the design of clinical database systems, in this paper, we review the current literature in relation to Clinical data management (Section 2), clinical research databases (Section 3), and structural designs for clinical databases (Section 4). In Section 5, we present our findings and the implications, while Section VI concludes the paper.

2. Clinical data management

Clinical data management (CDM) is defined by Madison and Plaunt (2003) as "the process of collecting and validating clinical information with the goal of converting it into an electronic format to answer research questions and to preserve it for future scientific investigation" (Chow and Liu, 2005). Therefore, obtaining data for research from HISs which lack proper linkage, such as electronic medical records (EMRs), does not support the purpose of research, as it is limited to the information gathered in relation to the health problems of a specific patient (Rocca et al., 2012). On the other hand, clinical research databases (CRDs) construction adheres to the CDM objectives and accommodates information gathered from various patients' clinical data sources, such as EMRs, sensors, implanted

* Corresponding Author.
Email Address: hsamra@kau.edu.sa (H. Samra)
https://doi.org/10.21833/ijaas.2021.03.003
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-5199-3677

devices, in-home care devices, and mobile devices. In general, CDRs tend to be specific to a disease, population, procedure, treatment, or device (Gliklich et al., 2014). The automatic transfer of patient data between patient care databases and clinical research databases can help reduce data duplication and increase consistency (Collen, 1990). The structural organization of the clinical research database is designed to support data retrieval and answer research questions using software tools for custom queries, reporting, and statistical analysis.

2.1. Data collection

Healthcare data can be obtained using primary data collection methods such as observations, surveys, and interviews. In addition, secondary sources of data, which are pre-organized healthcare data, can be obtained from the electronic patient record, research articles, the Internet, and the library. Both methods can be used for research and healthcare management. Although the primary data collection methods provide unbiased, current, and independent information, it is still a very expensive method and produces limited information. However, the secondary data collection methods can provide unlimited data but with many concerns in relation to data reliability and usefulness (Gliklich et al., 2014). Although data collected through health information systems such as electronic health records (EHR) with superior traditional data collection methods such as paper-based or by phone, their quality has been questioned. There are many reasons that affected data quality in these systems, such as incomplete, inconsistent, and noisy data. Sometimes, physicians collect their findings using free-text notes or by dictation, and these reports need to be transcribed into the computer (Dziadkowiec et al., 2016). Data collected through this method should undergo the categorization process into a group of functions such as diagnosis, treatment, and plans (Wiederhold, 2012). Otherwise, the collected data can be difficult to manage for research purposes unless it goes through the preparation process to make it useful for research. The pre-processing data phase is the best solution to improve the quality of data, which affects the analysis outcomes. Data pre-processing is concerned with the preparation and transformation of the initial dataset. Therefore, this stage contains methods of data cleaning and noise handling, data integration, and data transformation using a standard format, and finally, data reduction, i.e., summarized reports.

2.1.1. Data pre-processing

Data collection and pre-processing are the most significant and fundamental stages by which to acquire correct and appropriate data for further analysis tasks. Data preparation is essential in discovering the required knowledge, especially from a field that generates high-volume data such as healthcare. Clinical data are characterized by heterogeneity, which may come in the form of structural, unorganized, or semi-structured data. Therefore, knowledge cannot be acquired, comprehended, and automatically extracted without the application of pre-processing techniques. The secondary use of extracted data from a health information system requires the use of pre-processing measures to eliminate data quality issues that may result from missing data or incomplete medical records (Dziadkowiec et al., 2016). An efficient and robust pre-processing algorithm needs to be implemented prior to data transformation and loading into the database (Wiederhold, 2012). For example, data cleaning techniques employ methods to impute/fill incomplete data or treat noise by either polishing/correcting or filtering/removing the noisy instances (Fayyad et al., 1996).

2.2. Data storage

A data storage system contains two parts: A hardware infrastructure in the lower layer and storage methods or mechanisms on the top layer. The hardware infrastructure is a combination of both hardware equipment such as servers, routers, network links, and software components such as operating systems (Hassanien et al., 2015). In general, data storage systems must be equipped with multiple application programming interfaces (APIs), rapid query, or other software models for analyzing or interacting with data in the physical layer (Chen et al., 2014). Current storage mechanisms can be classified into three bottom-up levels: File systems, databases, and programming models (Chen et al., 2014).

2.2.1. File systems

File systems are the base for the applications at upper levels. The Google file system (GFS) is an example of a highly scalable and consistent distributed file system for large-scale data-intensive applications (Hashem et al., 2015). However, GFS has some limitations, such as poor performance for small files and a single point of failure (Chen et al., 2014).

2.2.2. Database systems

Database systems have been developed over the past decades to manage various types and scales of datasets. Database technologies, such as data warehousing, have been used for big data storage for quite some time and have contributed to the development of several storage techniques (Minelli et al., 2013). In addition to relational DB, these database technologies include object DB, XML DB, and multidimensional DB, which provide greater support for traditional datasets but are unable to meet the challenges brought by big data. Alternatively, NoSQL databases achieve greater performance with respect to traditional RDBMSs (Hashem et al., 2015). The simplest version of NoSQL

is key-value stores, where any data item can be a key to stored digital objects (Blanke, 2014). Key-value scales to unlimited data size, for example, Amazon's Dynamo, which provides incremental scalability (Srinivasa and Bhatnagar, 2012). Document stores are the dominant version in NoSQL databases (Blanke, 2014). Document stores are associated with object-oriented programming throughout the entire process, from clustering to accessing the data. Also, document stores have the same behavior as key-value stores, as a value associated with a key is the document content. They are useful for data with high complexity, such as medical records. Examples of this type are MongoDB and CouchDB (Akerkar, 2013). Column-oriented stores are databases organized into related column groups and are inspired by Google's BigTable, which is distributed, strong, and a multidimensional sorted map (Srinivasa and Bhatnagar, 2012). BigTable was developed by Google, based on the GFS to manage highly scalable structured data (Chen et al., 2014). BigTable is a sequence of nested key-value pairs where keys and values can be composed as Apache HBase, and Cassandra, an open-source database management system (Manyika et al., 2011).

2.2.3. Database programming model

Database programming models have been developed to achieve effective distribution at scale for data-intensive applications. In the context of NoSQL databases, programming languages such as MapReduce have been introduced to minimize the complex tasks for data processing and reduce the performance gap among relational databases. As a result, programming models have become a foundation for the data-processing paradigm for highly scalable, fault-tolerant, large-scale distributed applications (Kambatla et al., 2014).

MapReduce is a powerful programming model for large-scale applications that uses a simple technique that emerged from those used in the area of distributed databases (Hassanien et al., 2015). MapReduce is a parallel programming framework developed by Google based on GFS for global analysis in big data (Blanke, 2014). The fundamental role of MapReduce is based on the divide-and-conquer method. In the "map" step, the programming task is divided into sub-tasks using the mapper function, which takes the input as a key-value pair and distributes the smaller sub-tasks to be solved in a parallel and separate way. Then, in the "reduce" step, solutions from different distributed nodes for the sub-task are combined to provide a solution to the original task (Chen and Zhang, 2014). The MapReduce program can be written in a complicated low-level language such as Java, which makes writing custom jobs difficult and time-consuming and requires a highly skilled programmer. Therefore, some advanced high-level query languages have been developed within the MapReduce framework, for example, Hive, Pig, and Jaql (Srinivasa and Bhatnagar, 2012).

Dryad is a programming model that implements parallel and distributed programs that are scalable and user-friendly (Mohanty et al., 2015). Dryad's operational structure is a directed acyclic graph where a centralized job manager assigns computations to several processors, monitors the execution, and is responsible for decision making (Chen et al., 2014). Dryad is an independent system with complete functions that support job creation, monitoring, management, and visualization and also resource management, fault tolerance, and re-execution (Chen and Zhang, 2014).

2.3. Data analysis

Data analysis is the final stage of the data management process related to clarifying the meaning and understanding of the data collected and is organized for research purposes. Data analysis methods and techniques are applied for the interpretation of the results, writing reports, and evaluation (Richmond, 2006). For decades, descriptive statistics have been used merely to describe what has happened, such as in the most popular statistical package, SPSS. Also, past information predictive and prescriptive analytics are used to predict the future outcome and to direct future activities to achieve the best results, respectively (Minelli et al., 2013). The analysis of structured data reached an advanced state which now relies on a mature technology such as RDBMS, data warehouses, or OLAP (online analytical processing). The analysis is mostly based on a data mining and statistical approach in addition to statistical machine learning, which has been applied to detect anomalies within the data using mathematical models and powerful algorithms (Chen et al., 2014). On the other hand, unstructured data analysis, such as text mining, is a process of extracting useful information from unstructured text. Some text mining systems use a rule-based approach to identify patterns; however, others use machine learning techniques like natural language processing (NLP) and other algorithms to discover patterns automatically from the datasets (Franks, 2012).

Data analytics plays a significant role in making decisions in clinical practice, which can help determine the best course of action for diagnosis, treatment, and discovery to improve the quality of healthcare (Aleem et al., 2008). The identification of data patterns and the relationships among them help to develop more insight using algorithms and analytics tools (Archenaa and Anita, 2015). Big data analytics has had a pervasive impact on healthcare, which is clearly visible in different areas, such as improving the efficiency and quality of care while lowering the cost, as well as early detection and prevention of disease (Raj et al., 2015). Although healthcare systems have all the requirements for the effective application of big data analytics, such as data-intensive and critical decision support, challenges such as interoperability issues and

privacy and security concerns remain (Kudyba, 2014). As stated in the 2011 McKinsey Global Institute Report, big data analytics can effectively contribute to different areas such as clinical operations, research and development, and public health to provide a better outcome and reduce waste and inefficiency (Raghupathi and Raghupathi, 2014; Manyika et al., 2011).

Within clinical operations, outcomes-based research such as comparative effectiveness research (CER) determines the most relevant and cost-effective treatment for a patient, depending on the analysis results from a comprehensive patient and outcome data. Also, the deployment of clinical decision support systems helps lower the number of clinical care mistakes, reduce treatment errors and adverse reactions, and enhances the efficiency and quality of operations (Manyika et al., 2011). The implementation of advanced analytical methods, such as segmentation and predictive modeling on patient profiles, identifies patients at risk who may benefit from proactive care or lifestyle changes (Raghupathi and Raghupathi, 2014; Manyika et al., 2011). Furthermore, the use of evidence-based medicine for the detection and prediction of at-risk patients based on big data gathered from various healthcare sources provides sufficient evidence to identify and deliver effective clinical care (Archenaa and Anita, 2015).

In Rand D, predictive modeling has had an incredible impact in terms of disease diagnosis and treatment (Aggarwal and Reddy, 2015); it not only leads to the prediction of clinical outcomes and new drugs but also includes evaluation factors such as safety, efficacy, possible side effects, and the final trial outcomes. The clinical phase of the Rand D process can benefit from the application of statistical tools during patient recruitment to improve the design of clinical trials as well as to analyze clinical trial data and patient records. This will identify further signs and discover adverse effects as well as enable the detection of rare safety signs that appear in a typical trial and reduce drug withdrawal from the market (Manyika et al., 2011).

In public health surveillance and response, analyzing a nationwide patient and treatment database for the rapid detection of infectious diseases and outbreak provides a quick surveillance response and reduces infections. Also, these analyses can be used for the rapid development of more accurate targeted vaccines (Raghupathi and Raghupathi, 2014). The healthcare industry can exploit diverse data analytics technologies to process and analyze medical data for the improvement of healthcare services. The two widely used techniques in utilizing such data are information retrieval and data mining (Yang et al., 2015).

Information Retrieval (IR) is the most commonly used technique that deals with the process of "acquisition, organization, and searching for knowledge-based information" (Aggarwal and Reddy, 2015). IR can be used to obtain information by searching for a specific user's query within a large document collection where the retrieved subset of information is in the same format as the original with no added values (Yang et al., 2015). Traditionally, IR focuses on the retrieval of text from medical data; however, now, it covers a wide range of digital media, including the retrieval of medical images (Aggarwal and Reddy, 2015). Medical text retrieval can be considered to be a domain-specific text search with the significant challenge of dealing with the inherent complexity and ambiguity of medical terminologies that require standardization. Therefore, semantic-based text search approaches are utilized to tackle the ambiguity issue in medical text. However, for medical image retrieval, either text-based or content-based approaches can be used for this task. Text-based retrieval depends on the annotated text associated with images. But in content-based medical image retrieval, the process depends on the description of the visual features of the image, such as color, which can be automatically generated while indexing the medical image (Yang et al., 2015).

Data mining is the process of extracting patterns from massive datasets by combining methods from statistics, machine learning, and artificial intelligence with database management systems (Manyika et al., 2011). Healthcare data mining concentrates on comprehensive questions and outcomes, for example, symptoms and all the related data and clinical outcomes in combination lead to particular diagnoses and treatments (Kudyba, 2014). The application of data mining in healthcare can be classified into supervised (predictive) and unsupervised (descriptive) approaches. Supervised learning methods are used to build clinical prediction models based on predicting a function or associations from a set of training data (Manyika et al., 2011). These methods have been successfully employed in clinical prediction using statistical methods (i.e., linear regression, logistic regression, and Bayesian models), sophisticated methods in machine learning and data mining (i.e., decision trees and artificial neural networks), and survival models that try to predict the time of the occurrence of a specific event. Generally, supervised learning methods can be classified into two broad categories: Classification and regression, where both focus on discovering the underlying relationship between covariate variables and a dependent outcome variable (Aggarwal and Reddy, 2015). Unsupervised learning methods involving data clustering are a technique that finds hidden structures in unlabeled data (Manyika et al., 2011). These methods depend on grouping data into clusters according to the objects' (patients or EHRs) similarity measurements. Examples of unsupervised or descriptive data mining approaches include clustering, association rule mining, and sequence discovery (Yang et al., 2015).

3. Clinical research databases

Clinical research databases can be primary databases where data are collected specifically for

research, such as clinical trial studies (Loke, 2014). However, generally, they are secondary databases that contain a specific group of data extracted from primary databases (EMRs) with a common problem (Gliklich et al., 2014). Clinical research databases can be categorized according to their analytical purpose into (i) descriptive analyses to extract summaries of the essential features of a database, such as grouping patients with similar conditions and identifying the critical characteristics of each condition; and (ii) predictive analyses to derive classification rules, such as developing diagnostic standards which predict the course of a disease. Clinical research databases require de-identifying all patient data before including and using linking-variables to link patients who may be related to more than one source, but the patient's privacy and confidentiality always need to be maintained. The Department of Health and Human Services (HHS) enacted the Health Insurance Portability and Accountability Act (HIPAA) privacy rule that allows the use of healthcare data after removing an individual identifier (Collen, 2012).

4. Structural designs for clinical databases

Digital technology has grown rapidly in the healthcare sector, leading to a significant shift from paper to electronic records, thereby increasing the volume of healthcare data, which, as a result, requires databases to manage, manipulate, and store. Databases are fundamental to the effective use of data to serve an organization's multiple purposes (Wiederhold, 2012). The conceptual representation of an individual patient and modeling the schema for patient care during the healthcare process can be done using any structural design such as hierarchical, relational, or object-oriented, or a hybrid structural design. The design should adhere to the basic functional requirement to serve the primary goal of the medical database: (i) provide easy access to all relevant data for each patient served; and (ii) provide a resource for the scheduled retrieval of all relevant data from the records of all patients for any primary or secondary purpose (Collen, 2012).

The relational database is the most common database used in the healthcare system, which can be in the form of administrative and billing data or recording patient care, surveillance health status, and treatment advice. In addition, it can be used for research purposes to assist the researcher with studies such as drug effectiveness and diseases. (Campbell, 2004).

There are various data storage models under the relational database concept, such as the Entity Attribute Value (EAV) model, which is the most widely adopted storage model in clinical systems. The EAV, as presented in Fig. 1, has a three-column fixed schema, entity, attribute, and value, which are used to store the primary key, the attribute name, and the data value, respectively. The EAV model improves flexibility by allowing attributes to be added by simply specifying their names in the attribute column. However, the main model drawbacks are the restriction on a single value column, which hinders the ability to use multiple data types (Batra et al., 2018). The type most commonly used in healthcare is the OLTP database. The structure of the OLTP can contain applications such as electronic health records (EHR), administration, billing and payment processing, financial systems, HR, and research. The OLTP system provides real-time transactional processing (search, store, update, delete) with a fast response time. In addition, the OLAP (a database that is a data warehouse can be built on top of the existing multiple OLTP databases to combine data with analytic purposes (Cardon, 2018).

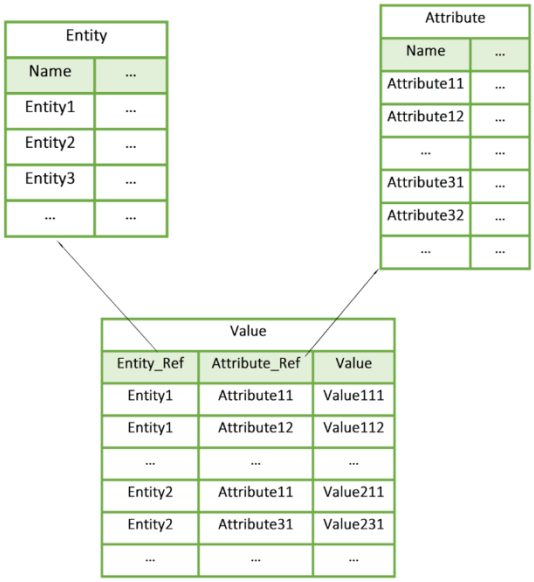

Fig. 1: The entity attribute value model (Loper et al., 2013)

4.1. Design challenges

There are many challenges that a database designer has to overcome, such as the structure and relationships of unique and complex healthcare information. For example, demographic information can relate to multiple diagnoses, which, in turn, are linked to other elements, such as the procedures performed by many doctors who can prescribe many medications (Campbell, 2004).

Although there is a constant change in information requirements due to advances in medical fields, historical data remains valuable and does not diminish like the majority of the traditional business information system. Moreover, during the design stage of the information system lifecycle, special attention is required for the medical field requirements, such as the common vocabulary of generally accepted terms for medical concepts and for administrative data used in relation to patients. A lack of medical record integration among physicians or across institutions creates great difficulties in data analysis and medical research studies (Muji et al., 2009).

4.1.1. Data integration

Data sharing through databases is common practice for clinical research as data collected at multiple sites are integrated with disease-oriented database systems since one location may not be able to collect sufficient data for analysis. Clinical institutions may also be limited in terms of research interests, so a common database can make the collected data available to researchers in a variety of locations (Wiederhold, 2012). The overall goal of data integration for the clinical research community is to be able to answer questions about aggregated data, which can be very difficult if each individual data source must be accessed separately or sequentially. The objectives of data integration in the context of health information exchange (HIE), as stated by Nadkarni and Marenco (2013), are:

- Being able to look at the "big picture": Collaboration between institutions that perform identical or highly similar operations but which are located far from each other is necessary to be able to look at consolidated summaries of structurally identical data to compare their performance.
- Identify common elements within different sources, which can then be used as a foundation for interoperability between systems that use individual sources. Such an effort was made by the Unified Medical Language System of the National Library of Medicine (UMLS), which uses controlled vocabulary to achieve standardization in certain biomedical areas.
- Eliminate repeated efforts and errors as a result of non-communication systems. This can exist in the same organization if they use multiple software packages from different vendors, which makes communication difficult and lead to duplication and inconsistency of data throughout the organization.

4.1.2. Integration approaches

Data integration from multiple types of data sources provides new knowledge using various datasets that cannot be gained from a single dataset (Gligorijević and Pržulj, 2015). Integration can be achieved using two broad strategies, physical data integration and logical data integration (Nadkarni and Marenco, 2013):

- Physical Data Integration: This approach relies on the concept of copying the original data, which is reorganized and moved from one or more repositories depending on the scope, purpose, and size of the data. The merged data is stored and managed by these new systems instead of the original source and is sorted in a single, queryable repository. The physical integration approach architecture can be represented in the form of a data warehouse for a wider scope, and great analytical capabilities, or a data mart for a small scope and special purpose focusing on one area may be used. The integration process starts with defining a global data model for the destination source (data warehouse). Then the selected data are migrated from the source to the destination using the extraction, transformation, and load (ETL) processes (Nadkarni and Marenco, 2013). Ultimately, all integrated data are transformed into the structure required by the global model, which provides quick access and excellent response time for queries (Louie et al., 2007).
- Logical Data Integration: Also called virtual integration, this approach uses conceptual schemes to bridge the representational heterogeneity of the databases and utilizing queries with the ability to collect and integrate data from distributed sources. The logical data integration architecture is based on data that are distributed in their original locations. In addition, intermediary software resides at a central location and uses a specific query protocol to communicate with the system that hosts the distributed data via the Internet. The mediator software is the point of communication between the disrupted hosts and users and mediates their request for data. Data federation is used to represent the data in the logical integration strategy. To achieve logical data integration, a global schema is defined for use as a validation model for the user query. Next, the mediator uses the mapping information to identify the location of the desired elements for the requested query. Then the proper translation of the global query to the local DBMSs query language of the distributed sources will be performed by the mediator (Nadkarni and Marenco, 2013).

4.1.3. Interoperation

Modern healthcare depends on collaboration and communication. With the growing application of health information interchange systems, there is a need for interoperability to provide information when and where necessary, facilitate decision making, reduce waste by eliminating redundant work and improve safety with fewer errors. Interoperability can be seen as four layers of "technology, data, human and institutional" with corresponding types of interoperability "technical, semantic, operational and clinical" (Benson and Grieve, 2016).

- Technical Interoperability is the technology layer, where information can be exchanged by the Health Information Technology (HIT) systems without any ability to interpret the data. This foundational layer is domain-independent, as reliable communication can be achieved over a noisy channel.
- Semantic Interoperability is the data layer, where HIT systems exchange, interpret, and use data without ambiguity. But this layer is domain and context-specific, which requires the use of standardized unambiguous codes.
- Process Interoperability is the human layer, where process interoperability is achieved when people

share a common understanding of their process artifacts across the network.
- Clinical Interoperability is a subset of process interoperability, which is the ability of two or more physicians in different care teams to transfer patients and provide smooth patient care.

As health data standards are a necessary component of interoperability in healthcare, poor implementation of interoperability results in failed large investments in digital health. The application of standards in healthcare will enhance the interoperability of healthcare systems to deliver timely services and provide better healthcare to patients (Khan et al., 2013).

5. Findings and implications

5.1. Clinical data stores (CDSs) and clinical data warehouse (CDW) for medical research

CDSs are suitable for the daily routine of clinical practices for patient care within each healthcare organization. CDSs contain disparate information across various departments and laboratories. It is difficult to access data for analysis due to the heterogeneity of data sources that require the development of a central interface for all systems and applications (Sahama and Croll, 2007). Data warehouses are always the best option for unifying scattered data in operational or transactional systems and provide a single view for useful, timely analysis for higher management and researchers. Thus, CDW provides efficient storage and powerful analysis tools to support healthcare providers' decisions and answer researchers' queries. Although the CDW can serve the purpose of data provided for research, it is difficult to build and requires lots of organizational resources for implementation and training purposes. Before applying either of these systems, whether CDSs or CDWs, the requirements of the organization must be carefully analyzed for successful implementation.

5.2. Entity attribute value (EAV) model vs. relational model for clinical research database

The relational model simplifies the representation of clinical care processes using the E-R diagram, which graphically represents the conceptual schema that can easily be transformed into a logical and physical model. Table 1 shows a simple design for modeling patient data in the database with predefined fields. The abstract (blueprint) representation of the relational model enables users to engage and facilitates user-developer communication. Normalization techniques allow for more flexibility in the relational model and provide accurate query results. On the other hand, in the entity attribute value model (EAV), there is no limitation on the number of attributes for each entity. In other words, the increased size of the logical database schema does not affect the physical schema (Batra et al., 2018; Anhøj, 2003). Although the EAV model is efficient for performing entity-centered queries, the restriction on a single value column results in less efficient performance for attribute-centered queries (Anhøj, 2003).

The database design modeling using the EAV model shown in Table 2 presents denormalized single value column attribute, which is not an efficient option to support patient-centered data with multiple attribute values, especially during the execution of a single value column attribute in large tables with numerous rows. Thus, the relational model, with its conceptual, logical, and physical modeling techniques, is easy to use and allows better communication and understanding for users during the development processes. Furthermore, features such as normalization provide flexibility and more accurate query results. This will reduce the time spent in the development of the system and allow more time for the users to test and evaluate the system.

Table 1: Relational database design example for simple Patients data

PatientID	Name	DOB	Gender
1	Patent1	01-01-1970	Male
2	Patent2	01-01-1980	Female

Table 2: EAV (Entity-Attribute-Value) database design example for simple Patients data

PatientID	Attribute	Value
1	Name	Patent1
1	DOB	01-01-1970
1	Gender	Male
2	Name	Patent2
2	DOB	01-01-1980
2	Gender	Female

5.3. Supportive technologies for clinical research database integration

5.3.1. Cloud computing

Cloud computing is one of the powerful technological advances that has emerged in modern ICT. In cloud computing, data, scalable computing resources, and other services are provided over the Internet at a lower cost (Manyika et al., 2011). Cloud computing provides services such as virtual resources, parallel processing, data integration, and scalable data storage (Erturk and Jyoti, 2015).

5.3.2. NoSQL databases

Although relational database architecture provides numerous advantages such as high consistency and availability, its performance decreases as the data grows and faces scalability constraints as it is impossible to scale horizontally, and its vertical growth is limited. NoSQL databases provide solutions for data aggregation and handling unstructured data, and its schema structure is flexible. In addition, it provides scalability for a quickly growing data repository, where horizontal

scalability is one of the NoSQL databases features (Kaur and Rani, 2013).

6. Conclusion

The complexity, rapid development, and expansion of the clinical information field make it difficult to develop and maintain clinical databases (Anhøj, 2003). The design and implementation of a clinical research database need more attention paid to the requirements to understand the conversion that takes place from the abstract model to a functioning database. As discussed, when it comes to using the database in general research, the relational database is the best option to allow more search options that supports complex queries for research questions and allows easy reporting options for novice users. However, for larger healthcare research organizations, CDW provides a comprehensive view of integrated data to support large studies as well as clinical decisions. Also, immersive technologies such as cloud computing and NoSQL databases can be used to integrate data from disparate sources with different formats and structures into a unified repository.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Aggarwal CC and Reddy CK (2015). Healthcare data analytics. Volume 36, CRC Press, Boca Raton, USA.

Akerkar R (2013). Big data computing. CRC Press, Boca Raton, USA. https://doi.org/10.1201/b16014

Aleem IS, Schemitsch EH, and Hanson BP (2008). What is a clinical decision analysis study? Indian Journal of Orthopaedics, 42(2): 137-139. https://doi.org/10.4103/0019-5413.40248 **PMid:19826517 PMCid:PMC2759613**

Anhøj J (2003). Generic design of web-based clinical databases. Journal of Medical Internet Research, 5(4): e27. https://doi.org/10.2196/jmir.5.4.e27 **PMid:14713655 PMCid:PMC1550574**

Archenaa J and Anita EM (2015). A survey of big data analytics in healthcare and government. Procedia Computer Science, 50: 408-413. https://doi.org/10.1016/j.procs.2015.04.021

Batra S, Sachdeva S, and Bhalla S (2018). Entity attribute value style modeling approach for archetype based data. Information, 9(1): 2. https://doi.org/10.3390/info9010002

Benson T and Grieve G (2016). Principles of health interoperability: SNOMED CT, HL7 and FHIR. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-319-30370-3

Blanke T (2014). Digital asset ecosystems: Rethinking crowds and clouds. Elsevier, Amsterdam, Netherlands.

Brazhnik O (2007). Databases and the geometry of knowledge. Data and Knowledge Engineering, 61(2): 207-227. https://doi.org/10.1016/j.datak.2006.05.005

Campbell RJ (2004). Database design: What HIM professionals need to know. Perspectives in Health Information Management, 1: 6.

Cardon D (2018). Healthcare databases: Purpose, strengths, weaknesses. Health Catalyst, Salt Lake City, USA.

Chen CP and Zhang CY (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information Sciences, 275: 314-347. https://doi.org/10.1016/j.ins.2014.01.015

Chen M, Mao S, Zhang Y, and Leung VC (2014). Big data: Related technologies, challenges and future prospects. Volume 96, Springer, Berlin, Germany. https://doi.org/10.1007/978-3-319-06245-7

Chow SC and Liu JP (2005). Design and analysis of clinical trials: Concepts and methodologies. 2nd Edition, Wiley and Sons, Hoboken, USA.

Collen MF (1990). Clinical research databases: A historical review. Journal of Medical Systems, 14(6): 323-344. https://doi.org/10.1007/BF00996713 **PMid:2132040**

Collen MF (2012). Secondary medical research databases. In: Collen MF (Ed.), Computer medical databases: 183-193. Springer, London, UK. https://doi.org/10.1007/978-0-85729-962-8_6

Dziadkowiec O, Callahan T, Ozkaynak M, Reeder B, and Welton J (2016). Using a data quality framework to clean data extracted from the electronic health record: A case study. eGEMs: Generating Evidence and Methods to improve patient outcomes, 4(1): 1201. https://doi.org/10.13063/2327-9214.1201 **PMid:27429992 PMCid:PMC4933574**

Erturk E and Jyoti K (2015). Perspectives on a big data Application: What database engineers and IT students need to know. Engineering, Technology and Applied Science Research, 5(5): 850-853. https://doi.org/10.48084/etasr.592

Fayyad U, Piatetsky-Shapiro G, and Smyth P (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3): 37-37.

Franks B (2012). Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics. Volume 49, John Wiley and Sons, Hoboken, USA. https://doi.org/10.1002/9781119204275

Gligorijević V and Pržulj N (2015). Methods for biological data integration: Perspectives and challenges. Journal of the Royal Society Interface, 12(112): 20150571. https://doi.org/10.1098/rsif.2015.0571 **PMid:26490630 PMCid:PMC4685837**

Gliklich RE, Dreyer NA, and Leavy MB (2014). Registries for evaluating patient outcomes: A user's guide. Government Printing Office, Washington, USA.

Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, and Khan SU (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47: 98-115. https://doi.org/10.1016/j.is.2014.07.006

Hassanien AE, Azar AT, Snasael V, Kacprzyk J, and Abawajy JH (2015). Big data in complex systems. Volume 9, Springer, Berlin, Germany. https://doi.org/10.1007/978-3-319-11056-1

Kambatla K, Kollias G, Kumar V, and Grama A (2014). Trends in big data analytics. Journal of Parallel and Distributed Computing, 74(7): 2561-2573. https://doi.org/10.1016/j.jpdc.2014.01.003

Kaur K and Rani R (2013). Modeling and querying data in NoSQL databases. In the IEEE International Conference on Big Data, IEEE, Silicon Valley, USA: 1-7. https://doi.org/10.1109/BigData.2013.6691765

Khan WA, Hussain M, Latif K, Afzal M, Ahmad F, and Lee S (2013). Process interoperability in healthcare systems with dynamic

semantic web services. Computing, 95(9): 837-862. https://doi.org/10.1007/s00607-012-0239-3

Kudyba S (2014). Big data, mining, and analytics: Components of strategic decision making. CRC Press, Boca Raton, USA. https://doi.org/10.1201/b16666

Lippeveld T, Sauerborn R, Bodart C, and WHO (2000). Design and implementation of health information systems. World Health Organization, Geneva, Switzerland.

Loke YK (2014). Use of databases for clinical research. Archives of Disease in Childhood, 99(6): 587-589. https://doi.org/10.1136/archdischild-2013-304466 **PMid:24489362**

Loper D, Klettke M, Bruder I, and Heuer A (2013). Enabling flexible integration of healthcare information using the entity-attribute-value storage model. Health Information Science and Systems, 1: 9. https://doi.org/10.1186/2047-2501-1-9 **PMid:25825661 PMCid:PMC4340778**

Louie B, Mork P, Martin-Sanchez F, Halevy A, and Tarczy-Hornoch P (2007). Data integration and genomic medicine. Journal of Biomedical Informatics, 40(1): 5-16. https://doi.org/10.1016/j.jbi.2006.02.007 **PMid:16574494**

Madison T and Plaunt M (2003). Clinical data management. 2nd Edition, Encyclopedia of Biopharmaceutical Statistics, New York, USA. https://doi.org/10.1201/b14760-26

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, and Hung Byers A (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Minelli M, Chambers M, and Dhiraj A (2013). Big data, big analytics: Emerging business intelligence and analytic trends for today's businesses. Volume 578, John Wiley and Sons, Hoboken, USA. https://doi.org/10.1002/9781118562260

Mohanty H, Bhuyan P, and Chenthati D (2015). Big data: A primer. Volume 11, Springer, Berlin, Germany. https://doi.org/10.1007/978-81-322-2494-5

Muji M, Ciupa RV, Dobru D, Bica C, Olah P, Bacarea V, and Marusteri M (2009). Database design patterns for healthcare information systems. In the International Conference on Advancements of Medicine and Health Care Through Technology, Springer, Cluj-Napoca, Romania: 63-66. https://doi.org/10.1007/978-3-642-04292-8_14

Nadkarni P and Marenco L (2013). Data integration: An overview. In: Sarkar IN (ed.), Methods in biomedical informatics: A pragmatic approach: 15-47. Elsevier Inc., Waltham, USA. https://doi.org/10.1016/B978-0-12-401678-1.00002-6 **PMCid:PMC4325484**

Raghupathi W and Raghupathi V (2014). Big data analytics in healthcare: Promise and potential. Health Information Science and Systems, 2: 3. https://doi.org/10.1186/2047-2501-2-3 **PMid:25825667 PMCid:PMC4341817**

Raj P, Raman A, Nagaraj D, and Duggirala S (2015). Big data analytics for healthcare. In: Raj P, Raman A, Nagaraj D, and Duggirala S (Eds.), High-performance big-data analytics: 391-424. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-20744-5_14

Richmond B (2006). Introduction to data analysis handbook. Academy for Educational Development, Durham, UK.

Rocca WA, Yawn BP, Sauver JLS, Grossardt BR, and Melton LJ (2012). History of the Rochester epidemiology project: Half a century of medical records linkage in a US population. Mayo Clinic Proceedings, 87(12): 1202-1213. https://doi.org/10.1016/j.mayocp.2012.08.012 **PMid:23199802 PMCid:PMC3541925**

Sahama TR and Croll PR (2007). A data warehouse architecture for clinical data warehousing. In: Warren J, Roddick J, Steketee C, Brankovic L, Coddington P, and Wendelborn A (Eds.), ACSW frontiers 2007: Proceedings of 5th Australasian symposium on grid computing and e-research: 227-232. Australian Computer Society, Darlinghurst, Australia.

Srinivasa S and Bhatnagar V (2012). Big data analytics. In the 1st International Conference on Big Data Analytics BDA, Springer, New Delhi, India: 24-26. https://doi.org/10.1007/978-3-642-35542-4

Wiederhold G (2012). Databases for health care. 1st Edition, Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-93174-1_1

Yang JJ, Li J, Mulder J, Wang Y, Chen S, Wu H, and Pan H (2015). Emerging information technologies for enhanced healthcare. Computers in Industry, 69: 3-11. https://doi.org/10.1016/j.compind.2015.01.012