Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 231 (2024) 452–457

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Workshop on Healthcare Open Data, Intelligence and Interoperability (HODII)
November 7-9, 2023, Almaty, Kazakhstan

Towards a Standardized Real-Time Data Repository based on Laboratory Test Results

Regina Sousa[a], Hugo Peixoto[a], Tiago Guimarães[a], António Abelha[a], José Machado[a,*]

[a]*LASI/Algoritmi Research Centre, University of Minho, Guimarães, Portugal*

Abstract

Healthcare facilities use huge quantities of real-time and analytical data to discover meaningful information from patient clinical lab results. Advanced analytics and machine learning algorithms help doctors identify and treat patients more accurately. Accurate models must be trained, tested, and validated with enough data. New real-time data allows healthcare practitioners to quickly and accurately analyse patient demands. Healthcare organizations can improve patient care and outcomes through knowledge discovery. The goal of this effort is to develop a real-time data repository based on patient clinical exams. This collection feeds real-time monitoring panels and machine or deep learning algorithms that forecast patient progression from clinical lab results. Integrate HL7 messages from diverse sources, preprocess them, and add them to an API-accessible data warehouse. In conclusion, the proposed method creates an international-standard data warehouse. This data warehouse can increase healthcare decision-making accuracy and efficacy when utilised with machine learning models, improving patient care and outcomes through more personalised treatment options.

Keywords: Data Warehouse; API; Real-Time Information Systems; Knowledge Discovery; Clinical Test Results; Health Data Standards

1. Introduction and Contextualization

Lack of evidence on pathogenic organisms, risk prediction, insufficient resources, and ineffective therapeutic options often violate the evidence-based medicine premise. This is considerably more apparent during pandemics. In specific, Covid-19 pandemic highlighted the urgent need to discover new solutions to developing and existing healthcare problems [19, 11]. Consequently, traditional methods have been replaced by Artificial Intelligence (AI), Machine Learning (ML), and even applied statistics. However, successfully developing and deploying AI and ML systems requires massive data sets. Access to medical data, such as Electronic Health Records (EHRs), is sometimes restricted

* Corresponding author.
E-mail address: jmac@di.uminho.pt

10.1016/j.procs.2023.12.233

by legislation designed to protect patient privacy. These restrictions make it difficult to reproduce existing results based on private health data and delay progress and studies [5]; [4]. Typically, this problem is solved by using synthetically generated health data, which protects privacy while allowing researchers to analyse and inform decisions. Indeed, this is an enormous temptation compared to the time and resources required to gather and categorise datasets containing millions of pieces retrieved from the actual world. Nevertheless, these contain simplified, generated data based on actual data [3]. The goal is for the synthetic data to accurately represent the original data. However, the problem of providing realistic data without exposing private information is recognized. Alternatively, if the synthetic data is insufficiently accurate, it will not reflect the essential patterns to the training or test data. Modeling efforts based on implausible data are unable to produce useful insights [13].

Noteworthy are data processing methods, such as anonymisation, which allow the anonymisation of real health data and provide valuable information about patient visits to health units, resulting in a time series dataset influenced by protected attributes, for instance, age, gender, and location. When entering the universe of data from the patient's medical record, the amount of data to be processed increases exponentially [8]. Millions of people produce valuable medical data records daily to feed information systems such as those used for forecasting or decision support. They require preprocessed and normalised data to sequentially improve their performance.

This is precisely the stated flaw, the absence of real-time updated data repositories containing data collected in the field. Thus, many EHR-based research papers lack follow-up data, making further study and validation a challenge. Providing a standardized and real-time data repository based on interoperability technologies, depending on the results of clinical analysis, would be highly beneficial to improve the delivery of health care, especially in supporting professionals in tracking disease progression, assessing risks and other tasks. After a health crisis comparable to that caused by Covid-19, all these benefits can be highlighted [12, 16].

For this to be possible, concepts such as Interoperability, Data Standards and Big Data must be present and well defined. Each of these notions will be defined in the following sections.

2. Background

2.1. Interoperability in Healthcare

In recent decades, the adoption of new technologies capable of digitising information on paper and assisting in decision-making has increased exponentially due to the influence of these two variables. This led to the creation of different platforms, each with its own language, data structure, etc., as they were made by different multi-functional actors. Here lies the need for interoperability that allows the secure exchange of data between the various players, considering the numerous islands of information [6]. Interoperability is needed, especially in crucial areas like healthcare where data is a critical part of the study. Clinical data interoperability refers to the ability of two or more systems to understand the data they exchange or share. There are precise syntactic, semantic, and cross-domain criteria for each level of interoperability. International eHealth communities have recommended and pushed to promote interoperability using clinical information models [15, 2].

2.2. Data Repositories

The term "Data Repository" refers to a generic infrastructure that stores segmented data for analysis or reporting. These infrastructures are currently highly appreciated because they enable businesses to make decisions supported by information that is typically more reliable than gut feeling [18]. A data repository increases the speed of data access and sharing, as well as the preservation and archiving of sensitive data. On the other hand, there are inherent weaknesses that must be addressed, such as the repository's potential behavior as the data expands or the possibility of a system failure necessitating more frequent backups [1]. Therefore, these are legitimate risks, but the data repository management team can be aware and plan for them.

3. Methodology

Even when it comes to healthcare, information systems (IS) research is primarily characterised by two paradigms: behavioral science and design science. The former seeks to push the limits of human and organisational capabilities by developing new and innovative IT artifacts, which are broadly defined as constructs, models, and/or methods. The latter seeks to push the boundaries of human and organisational capabilities [7]. Design Science Research (DSR) methodology for IS is suited for carrying out this research procedure because the work described results in an IT item.

As a result, it becomes clear that the effort began with a problem that has already been identified: the demand for more anonymized data that is updated in real-time for health research.

The main objectives that will be attained by using this strategy are the following, starting with the stated challenge:

- Analyse the data obtained through HL7 messages (Mirth Connect) in real-time;
- Sorting and processing the data gathered in the preceding point to retain just the relevant information;
- Identification of additional information of interest that should be included in the data repository but is not covered in the preceding point;
- Data availability, storage system selection and development, real-time data loading and updating using Big Data approaches;
- Validation of the data repository by health professionals.

The architecture served as the foundation for the produced solution.

The process begins at healthcare facilities that have the potential to expand. Additionally, these institutions could have several gathering locations, like emergency departments or intensive care units. These collection locations are suitable for a variety of research including blood, urine, stool, and other specimens. To ascertain the sample's analytical values, all gathered samples must be delivered to an internal or external laboratory. There is room for n values of analytical results in each sample. As a result, the lab sends an HL7, version 2 message to the integration engine for each sample that contains information about the patient's sequence number, the episode connected to the analysis, the analysis codes, the analytical values found during the test, the units for that value, and the expected limits.

The system's primary objective is to anonymise all information that can be linked to the patient or the institution for each of these messages. Adding noise to the data, namely the episode and sequence number, is one approach to achieve this. In addition, the sex at birth is altered in different ways. These modifications can take the shape of numerical values, such as 1- Male, 2- Female, or the entire words Male, Female. These options are all changed to M for Male and F for Female. All dates and character encoding are standardized as well, as one might anticipate. The diagnoses from the Diagnosis Related Groups (GDH) are given an ICD-10 code, and the analysis is given a LOINC number, to complete each record. It's necessary to keep in mind that the Extract Transform Load (ETL) process involves several other steps as well. Having said that, the fresh inputs are then instantly incorporated into the data store. The created API offers a few endpoints that other clients can access (e.g., web applications, system reports, and monitoring platforms). In this method, Mirth Connect was used for the Interoperability step and Spark written in Python (PySpark) for the ETL phase. Both artifacts will be thoroughly studied and detailed in sections 4 and 5.

4. Preliminary Implementation Results

Two artifacts were produced from the developed work. A data repository is the first, as mentioned in section 4.1, and an application programming interface is the section 4.2.

4.1. Real-Time Data Warehouse

Considering that it stores a sizeable amount of structured data from numerous data sources and transactional systems, the built data repository is categorised as a data warehouse type.

An estimated 2.5 million records are added to the repository annually for each healthcare facility. Since data from 2020 was taken into account, there are around 7 million records for the project's pilot institution. As a result, the structure of the data warehouse and a description of its tables will be explained in the parts that follow.

4.1.1. Data Warehouse Model

A single fact table, Results, and three other dimension tables, DemographicData, DiagnoseType, and ExamType were chosen as the final structure. These tables are not meant to add extra information to the fact table. As a result, the DemographicData table solely includes information on patients. The accompanying icd10 code and description can be found in the DiagnoseType table. All analytical codes, their definitions, unit breakdowns, and normal ranges are included in the ExamType table.

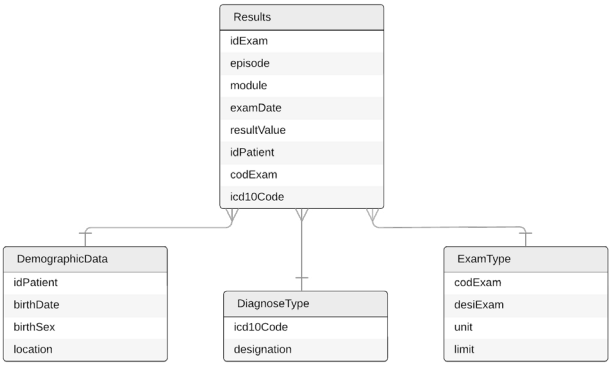

Fig. 1. Data Warehouse Architecture.

4.1.2. Description

The construction of the fact table and the three-dimension tables comes after the Star Schema.

- **Dimension Table 1:** DemographicData
  - **idPatient**: The sequential number is assigned, as the name implies, sequentially to the patient the first time he/she enters the health care institution. The same patient may be present in the database with different sequential numbers, because he/she attends n institutions. To the original sequential number and in order to guarantee the anonymisation of the dataset, a series of 4 numbers is added randomly;
  - **birthDate**: The date of birth also refers to the patient. To maintain the consistency of the data repository, all dates follow the format: dd-mm-yyyy hh:mm:ss;
  - **birthSex**: The birth sex is the attribute usually called administrative sex and refers to the identification, by a family member, midwife, nurse or doctor, of the external genital organs when the baby is born. In the developed repository this sex can contain two values: F-Female; M-Male;
  - **location**: The location refers to the place where the patient has his or her fiscal address. In Portugal this can be a city, a town or even a village.
- **Dimension Table 2:** DiagnoseType
  - **diagnoseCode**: The diagnosis used in the work developed is that of the Grupos de Diagnósticos Homogéneos(GDH), since it is the one that is most figdign. This system, GDH, is the disease classification system used in Portugal. Because it is so specific, this code was transformed into an ICD-10 code, so that in the future other countries can contribute to the repository developed.
  - **diagnoseDescription**: The description of the diagnosis is the description corresponding to the ICD-10 code.
- **Dimension Table 3:** ExamType
  - **examCode**: The examination code is the LOINC code that will pair with the designation described below;
  - **examDesignation**: The exam name is in English and is LOINC - Long Common Name;
  - **unit**: The unit corresponds to the dimensional unit associated with the value of the analysis performed;
  - **limit**: The limit refers to a standard of normality for the test code. For example, for the WBC analysis the value is 6.0 - 16.0;
- **Fact Table:** Results

- **idexam**: The exam number is, as its name indicates, the number that is assigned to the sample. In other words, there may be n results for the same test number because several parameters can be analyzed with the same sample. For example, patient ABC who has a test with the number 111111 has two results, one for the Leukocyte value and another for the Monocyte value;
- **episode**: The episode corresponds to a numeric value referring to a hospital visit. An episode is opened when the patient enters the hospital and closed when the patient is discharged. In short, each patient may contain n episodes that may occur in the scope of the emergency room, inpatient stay, consultation, among others;
- **module**: The module identifies the type of service in which the patient entered the hospital. For example, it could be CON - Appointment, INT - Internment, URG - Emergency, among others;
- **examDate**: the date of examination is only identifying the time of sample collection. Follows the pattern defined for the repository dd-mm-yyyy hh:mm:ss;
- **resultValue**: The value refers to the result obtained from the analysis of the sample. Typically it is a numeric value that will be completed by the 2 other columns that follow, the unit and the limit;
- **Foreigh Keys**: idPatient, codExam, icd10Code;

4.2. Application Programming Interface

Software programs can communicate with one another thanks to a set of guidelines, computer code, and standards known as an application programming interface (API). The created API was built on the NodeJS programming language, and swagger automatically generated interactive documentation for it.

The API provides several methods for accessing information:

- **authentication** - Generates json web token and is used for authorization on the remaining routes. This token is generated from the email and password pair.
- **getSample** - Returns the last 100 values entered in the repository;
- **getPatientResults** - According to an exam number and patient ID returns all patient results;
- **getResults** - According to patient identification returns all results for all tests;
- **getLastResult** - According to an analysis code and the patient's id returns the date of the last examination performed;
- **getResultsAge** - According to an analysis type and a certain age range, it returns the minimum, average and maximum value;
- **getModDiag** - According to the diagnosis code (ICD10), returns the most performed analyses;

Each of the aforementioned routes returns a JSON object with the most recent exam at the top. Please note that the repository is currently not open access, as the project is still ongoing.

5. Conclusion and Future Work

Customers expect businesses to be aware of their needs and expectations, according to 66% of consumers [14]. Companies can only acquire this knowledge through data collection and analysis. Through the identification of risk factors, the implementation of targeted interventions, and the optimization of resource allocation, the use of data-driven approaches in healthcare can enhance patient outcomes and lower costs, according to a systematic review published in the Journal of Medical Internet Research [9]. By making it possible to create tailored medicine, forecast disease outbreaks, and improve healthcare delivery, data analytics and machine learning techniques have the potential to revolutionise the healthcare industry [17]. Data is therefore one of the most valuable resources in the world, both in the commercial and intellectual arenas. Since knowledge that can be acquired and kept in a repository has a high monetary value for enterprises. For a number of reasons, it is crucial to create data repositories in the healthcare industry. The first benefit is that it enables the storing and organising of massive volumes of data, such as medical records, trial data, and research findings. This facilitates access to and analysis of the data by healthcare professionals, which can enhance decision-making and patient care [10]. Second, data repositories can promote data collaboration

and sharing across institutions, academics, and healthcare practitioners. This can result in the creation of novel cures and treatments as well as the discovery of patterns and trends in health data that can guide practice and policy. Finally, data repositories, which frequently have security measures to guard against unwanted access or breaches, can aid in ensuring the security and privacy of sensitive health data. In conclusion, building data repositories is essential for enhancing the effectiveness, security, and efficiency of health data management. Therefore, the created data warehouse is a very valuable artifact for both Machine Learning (ML) and Deep Learning (DL) models as well as decision support systems. It was created in accordance with the FAIR principles, making it Findable (each record has a unique key), Accessible (generic usage guidelines), Interoperable (use of internationally recognized and adopted terminologies), and Reusable (api developed with several endpoints available and dynamically documented).

The main contributions of this work are: The development of an API-accessible real-time data repository for clinical test findings; The incorporation of HL7 messages from many sources, enabling the creation of a consistent and orderly data structure that complies with global standards including HL7, ICD-10, LOINC, and FAIR; The data can be used to anticipate and enhance patient care by feeding dashboards and machine or deep learning models;The potential for this repository to be a useful tool for healthcare organizations, empowering them to take decisions based on data and progressively raise the standard of care given to patients.

Aknowledgements

This research was funded by Fundação para a Ciência e Tecnologia, within the Project Scope: UIDB/00319/2020.

References

[1] Austin, C.C., Brown, S., Fong, N., Humphrey, C., Leahey, A., Webster, P., 2016. Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements. IASSIST Quarterly 39, 24–24.
[2] Castanheira, A., Peixoto, H., Machado, J., 2020. Overcoming challenges in healthcare interoperability regulatory compliance, in: International Symposium on Ambient Intelligence, Springer. pp. 44–53.
[3] Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F., Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering 5, 493–497.
[4] Cornock, M., 2018. General data protection regulation (gdpr) and implications for research. Maturitas 111, A1–A2.
[5] Duan, Y., Edwards, J.S., Dwivedi, Y.K., 2019. Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda. International journal of information management 48, 63–71.
[6] d'Aliberti, O.G., Clark, M.A., 2022. Preserving patient privacy during computation over shared electronic health record data. Journal of Medical Systems 46, 1–8.
[7] Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design science in information systems research. MIS quarterly , 75–105.
[8] Langarizadeh, M., Orooji, A., Sheikhtaheri, A., Hayn, D., 2018. Effectiveness of anonymization methods in preserving patients' privacy: A systematic literature review. eHealth 248, 80–87.
[9] Lin, C.C., Huang, H.Y., Chen, W.C., Tsai, C.H., Chen, C.H., Cheng, T.C., Chen, H.C., Lee, H.H., Lai, M.F., Chen, T.H., 2017. Data-driven approaches in healthcare: A systematic review. Journal of Medical Internet Research 19, e269. URL: https://www.jmir.org/2017/9/e269/, doi:10.2196/jmir.7254.
[10] Oliveira, C., Sousa, R., Peixoto, H., Machado, J., 2022. Improving the effectiveness of heart disease diagnosis with machine learning, in: International Conference on Practical Applications of Agents and Multi-Agent Systems, Springer. pp. 222–231.
[11] Oliveira, D., Ferreira, D., Abreu, N., Leuschner, P., Abelha, A., Machado, J., 2021a. Prediction of covid-19 diagnosis based on openehr artefacts .
[12] Oliveira, D., Miranda, R., Leuschner, P., Abreu, N., Santos, M.F., Abelha, A., Machado, J., 2021b. Openehr modeling: improving clinical records during the covid-19 pandemic. Health and Technology 11, 1109–1118.
[13] Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M., 2018. Effective use of synthetic data for urban scene semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 84–100.
[14] Salesforce, 2022. Customer expectations. URL: https://www.salesforce.com/resources/articles/customer-expectations/?sfdc-redirect=369. accessed on [November 2022].
[15] Shull, J.G., 2019. Digital health and the state of interoperable electronic health records. JMIR medical informatics 7, e12712.
[16] Sousa, R., Miranda, R., Moreira, A., Alves, C., Lori, N., Machado, J., 2021. Software tools for conducting real-time information processing and visualization in industry: An up-to-date review. Applied Sciences 11, 4800.
[17] Topol, E., 2019. High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine 25, 44–56. URL: https://www.nature.com/articles/s41591-018-0268-4, doi:10.1038/s41591-018-0268-4.
[18] Uzwyshyn, R., 2016. Research data repositories: the what, when, why and how. Computers in Libraries 36, 8–21.
[19] Vijayvargiya, P., Garrigos, Z.E., Almeida, N.E.C., Gurram, P.R., Stevens, R.W., Razonable, R.R., 2020. Treatment considerations for covid-19: a critical review of the evidence (or lack thereof), in: Mayo Clinic Proceedings, Elsevier. pp. 1454–1466.