الجمهوريـــــــــــــــة الجزائريـــــــــة الديمقراطيـــــــة الشعبيـــــة
وزارة التعليـــــــم العـــــــالي و البحـــــــث العلـــــمي
جامعة وهران للعلوم والتكنولوجيا محمد بوضياف
كلية الرياضيات و الاعلام الالي
République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur Et de la Recherche Scientifique
Université des Sciences et de la Technologie d'Oran Mohamed BOUDIAF
Faculté des Mathématiques et Informatique

***Département** : Informatique*

Mémoire de fin d'études

Visualization of Health Data In A Medical Sector

Pour l'obtention du diplôme
de **Master**

Domaine : **Mathématiques – Informatique**

Filière : **Informatique**

Spécialité : **Systèmes d'Information et Données (SID)**

Présenté le : 09/06/2022
Par :

-BENZERGA Malek Kheira

Jury	Nom et Prénom	Grade	Université
Président	NAIT Bahloul Sara	MCA	USTOMB
Encadrant	MEBARKI Abdelkrim	MCA	USTOMB
Examinateur	MENAD Nadia	MCB	USTOMB

2021/2022

University of Science and Technology of Oran - Mohamed Boudiaf
MASTER DEGREE OF DATA & INFORMATION SYSTEMS
Department of Computer Science
Oran, Algeria

Visualisation of health data in a medical sector

Authored by

Malek kheira Benzerga

Supervised by

Abdelkarim Mebarki

2022

Acknowledgements & Dedication

I want to present all my gratitude to my supervisor Mr MEBARKI Abdelkarim who was always there to answer my questions, provide his insights, and to guide me in the right direction when it was needed and for accepting to supervise and helping me.

I present a special thanks to all the examiners Miss NAIT Bahloul and Miss MENAD Nadia for accepting to review and judge my thesis.

I would also like to take this opportunity to thank all the friends who have always been there for me : Oussama, Meriem, Houari, Soumia, Sirine, and to my colleagues from the UDev scientific club.

Finally, I wish to express my very profound gratitude to my family and special thanks to my parents for providing me with unconditional support and continuous encouragement throughout my years of study. This accomplishment would not have been possible without you. Thank you!

Malek.

Abstract

The care of patients in a medical structure is provided by medical and paramedical staff. Coordination between staff involved in care requires efficient and synchronized communication. The presence of visual representations in this communication adds a dimension to the data which can improve the reading of these data, their interpretations, and thus accelerate the response time.

In this project, four major points will be addressed: healthcare and data, information visualization in this field, the different data management systems for multiple medical actors and the visualization system proposed to display personalized dashboards for different actors.

***Keywords*—** Visualization; Medical data; Graphical Representation; Modeling; XML; Personalized Dashboard.

Résumé

La prise en charge des patients dans une structure médicale est assurée par du personnel médical et paramédical. La coordination entre le personnel impliqué dans les soins nécessite une communication efficace et synchronisée. La présence de représentations visuelles dans cette communication ajoute une dimension aux données qui peut améliorer la lecture de ces données, leurs interprétations, et ainsi accélérer le temps de réponse.

Dans ce projet, quatre points majeurs seront abordés : la santé et les données, la visualisation de l'information dans ce domaine, les différents systèmes de gestion de données pour de multiples acteurs médicaux et le système de visualisation proposé pour afficher des tableaux de bord personnalisés pour différents acteurs.

Mots clés— Visualisation ; Données médicales ; Représentation graphique ; La modélisation ; XML ; Tableau de bord personnalisé.

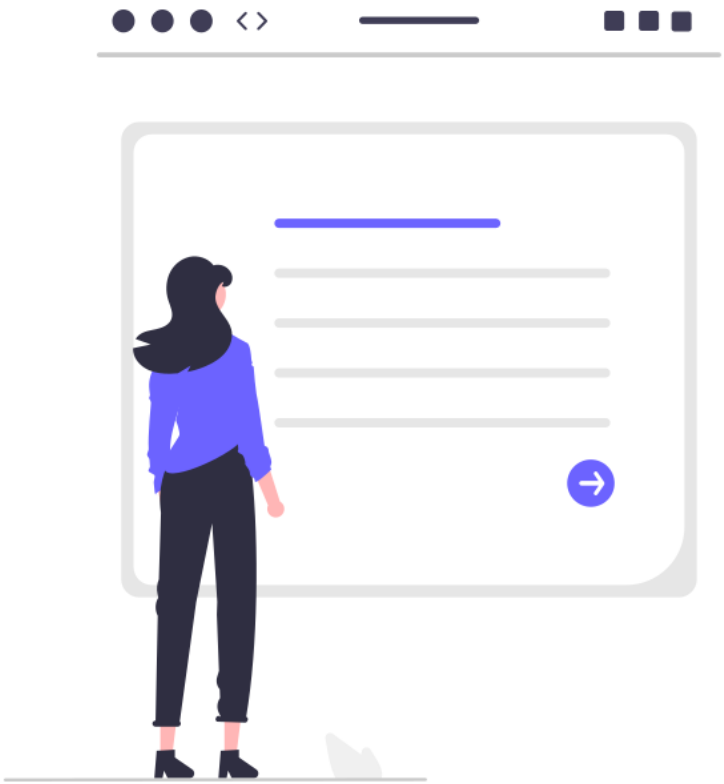
Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction



Health has played an important role in human history, helping civilization, behind the curtains, to evolve into the society of today[65]. Recently, the healthcare sector has witnessed the development of a wide range of IoT devices and applications[52]. And a new field has been unlocked : Healthcare information technology (HIT).

1 Background

Healthcare information technology (HIT) has been defined as « the application of information processing involving both computer hardware and software that deals with the storage, retrieval, sharing, and use of healthcare information, data, and knowledge for communication and decision making[38] » where the oil of it is the Medical Informatics as Morris F Collen defines it : « Medical informatics is the application of computer technology to all fields of medicine-medical care, medical teaching and medical research » ; in other words The medical informatics is the foundation for understanding and practice of the up-to-day medicine. Its basic tool is the computer, the subject of studying and the means by which the aspects and achievement in the new knowledge in studying of a man, his health and disease and functioning of the total health activities is performed[49].

Medical informatics as a discipline is still young, in particular when you compare it with other medical disciplines. However, within the past decades, societies in general, and medicine and healthcare in particular, have tremendously changed by the adoption of health information technology. This change has significantly impacted the healthcare field as well[43]. As a result, health information technology improves patient's safety by reducing medication errors, reducing adverse drug reactions, and improving compliance to practice guidelines. There should be no doubt that health information technology is an important tool for improving healthcare quality and safety[31].

2 Problem & motivation

With the progress of health information technology, the healthcare data is increasingly digitized and, like in most other industries, data is growing in Velocity, Volume and Value. According to Statista[8], the amount of global healthcare data is expected to increase dramatically by the year 2020. Early estimates from 2013 suggest that there were about 153 exabytes of healthcare data generated in that year. However, projections indicate that there could be as much as 2,314 exabytes of new data generated in 2020 (Figure 1.1).

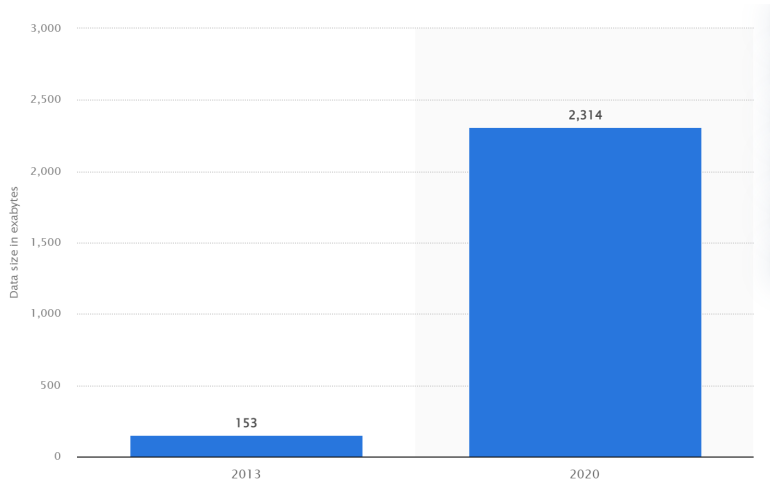


FIGURE 1.1 – Total amount of global healthcare data generated in 2013 and a projection for 2020* (in exabytes).

Health Data Management is the practice of making sense of this data and managing it to the benefit of healthcare organizations, practitioners, and ultimately patient well being and health, It enables the integration and analysis of medical data to make patient care more efficient, and extract insights that can improve medical outcomes, while protecting the security and privacy of the data. In the past forty years, medical data began a transition from purely paper-based tracking to digitized information. Even today, many types of medical data have yet to be digitized, or have not yet been integrated into Health Data Management systems. Some of the important challenges facing health data professionals today are[7] :

- ***Fragmented data :*** medical data can be structured data in spreadsheets or databases, images or video files, digital documents, scanned paper documents, or may be stored in specialized formats such as the DICOM format used for MRI scans. Data is widely duplicated, collected multiple times and stored in different versions by healthcare providers, public health organizations, insurance bodies, pharmacies, and patients themselves. There is no one source of truth for information on patient well being.
- ***Changes to data :*** medical data constantly changes as do the names, professions, locations and conditions of patients and physicians. Patients undergo numerous tests and are administered many types of treatment over the years, and the treatments and medications themselves evolve over time. New types of medical treatment, such as telehealth models, create new types of data.

- ***Regulations and compliance :*** medical data is sensitive and must adhere to government regulations, such as the USA Health Insurance Portability and Accountability Act (HIPAA). Data discovery challenges and poor data quality make it much more difficult to perform the required audits and meet regulatory requirements and limits the diversity of data healthcare providers can use for the benefit of patients.

In short, these challenges together with the lack of data management systems that provide the right insights to the right actors constitute an obstacle to the development of medical informatics.

3 Purpose & delimitations

The goal of this work is to design a visualization system in medical healthcare that provides personalized dashboards and insights for medical actors.

We intend to achieve our goal by designing a visualization system that manages data from different sources, and structures it, then provides a personalized dashboard for each actor.
First, we will create a data integration system that takes care of the data management.
Second, we will transform the focus data to a structured format.
Third, we will design a visual presentation after the data processing and formatting (a dashboard application).

4 Document structure

This document is presented in 6 chapters, starting with the chapter1 : Introduction, in which we present a bit of background of the topic and then delve into formally defining the problem we intend to tackle, followed by a brief description of what lies within and beyond the scope of this work.

In Chapter 2 : Health sector & Data, we present the healthcare environment including the principal actors, activities, and the type of data generated from each one. Then we presented the Medical records management and its various electronic types and how important security is to them.

In chapter 3 : Information Visualization, we present its definition, then we explain the visualization pipeline : how it deals with data, then we move to the data warehouse and its data integration approaches.

Next comes chapter 4 : Contribution & Discussion, we talk about the work objectives then we go through several related work reviews and then present the proposed work.

Chapter 5 : Implementation, a chapter about the implementation of the system, which tools are introduced and results are displayed using screenshots and diagrams.

Finally, in chapter 6 : Conclusion & Future work, the results and insights gained through the journey of making the proposed solution, few conclusions drawn and perspectives on what could be enhanced moving forward with this project.

Chapitre 2

Health sector & Data

In this chapter, we provide an overview of data from the health sector and the actors involved in it, by presenting the components of a health system and the types of data generated by each activity. In the second part, we will move on to the multiple electronic records used in the healthcare.

1 Healthcare Environment

Healthcare is a multi-dimensional system established with the sole aim for the prevention, diagnosis, and treatment of health-related issues or impairments in humans. There is three components of a healthcare system[39] :

- The health professionals (physicians or nurses) : belong to various health sectors like dentistry, medicine, midwifery, nursing, psychology, physiotherapy, and many others.
- Health facilities (clinics, hospitals for delivering medicines and other diagnosis or treatment technologies).
- Financing institution supporting the former two.

Healthcare is required at several levels depending on the urgency of situation :

1. ***Primary care :*** Professionals serve it as the first point of consultation.
2. ***Secondary care :*** acute care requiring skilled professionals.
3. ***Tertiary care :*** advanced medical investigation and treatment.
4. ***Quaternary care :*** highly uncommon diagnostic or surgical procedures.

At all these levels, the health professionals are responsible for different kind of information such as a patient's medical history (diagnosis and prescriptions related data), medical and clinical data (like data from imaging and laboratory examinations), and other private or personal medical data

Regardless of what form it takes, data has the potential to tell stories, identify cost savings and efficiencies, new connections and opportunities, and enable improved understanding of the past to shape a better future[66].

The term "big data" has become a buzzword in recent years, with its usage frequency having doubled each year in the last few years according to common search engines (Figure 2.1).

Big data is a vague term with a definition that is not universally agreed upon. A definition by Demchenko et al[40] who define Big Data by five V's : Volume, Velocity, Variety, Veracity, and Value. Volume pertains to vast amounts of data, Velocity applies to the high pace at which new data is generated, Variety pertains to the level of complexity of the data, Veracity measures the genuineness of data, and Value evaluates how good the quality of the data is in reference to the intended results.

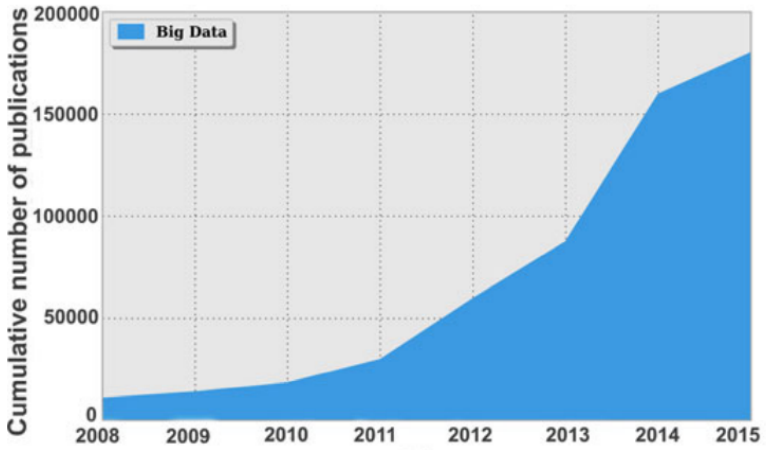


FIGURE 2.1 – Cumulative number of publications referring to "big data" indexed by Google Scholar.

If we trace the relationship between the use of the term big data per health research , we can easily infer the growth of medical informatics (Figure 2.2). Big data in health is concerned with meaningful datasets that are too big, too fast, and too complex for healthcare providers to process and interpret with existing tools[34].

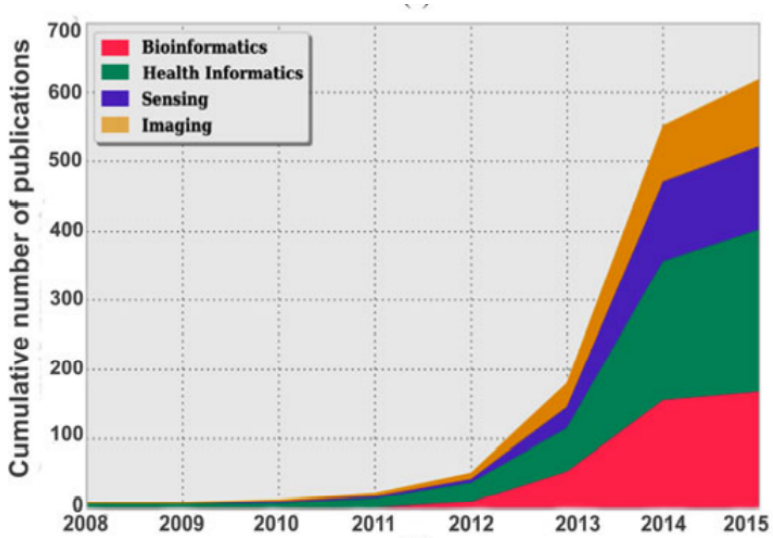


FIGURE 2.2 – Cumulative number of publications per health research area referring to "big data," as indexed in IEEE Xplore, ACM Digital library, PubMed (National Library of Medicine, Bethesda, MD), Web of Science, and Scopus.

There are numerous current areas of research within the field of Health Informatics, including Bioinformatics, Image Informatics (e.g. Neuroinformatic), Clinical Informatics, Public Health Informatics, and also Translational BioInformatics (TBI). Research done in Health Informatics (as in all its subfields) can range from data acquisition, retrieval, storage, analytics employing data mining techniques, and so on.

Data gathered for Health Informatics research does exhibit many of these qualities. Big Volume comes from large amounts of records stored for patients : for example, in some datasets each instance is quite large (e.g. datasets using MRI images or gene microarrays for each patient), while others have a large pool with which to gather data (such as social media data gathered from a population). Big Velocity occurs when new data is coming in at high speeds, which can be seen when trying to monitor real-time events whether that be monitoring a patient's current condition through medical sensors or attempting to track an epidemic through multitudes of incoming web posts (such as from Twitter). Big Variety pertains to datasets with a large amount of varying types of independent attributes, datasets that are gathered from many sources (e.g. search query data comes from many different age groups that use a search engine), or any dataset that is complex and thus needs to be seen at many levels of data throughout Health Informatics.

Schematically, several health-related activities can be distinguished[48] :

- Pre-admission, admission and administrative discharge activities.
- T2A invoicing or valuation activities.
- Care activities in the accommodation service.
- Activities in operating theaters and technical platforms.
- Laboratory activities.
- Imaging activities.

For each activity, different types of data are generated. In France and most developed countries, the following data is collected and digitally available (in chronological order)[1, 37] :

1. **Administrative data** related to patient movements (identity, dates, places, etc.), demographic (age, sex, place of residence, etc.) and insurance (health coverage, etc.).
2. **Results of biological analyses**, generally taken by nurses and analyzed by professionals or by robots.
3. **Medical data** produced automatically by autonomous medical devices. These devices can be implantable or external.

4. **Data relating to the drugs administered to the patient**, generally by nurses or doctors, possibly as part of a diagnostic or therapeutic procedure.
5. **Data relating to medical devices** implanted in the patient during surgery.
6. **Data relating to medical procedures**, whether diagnostic or therapeutic. These data are generally coded by the producer, sometimes by the machine which produces them.
7. **Comments** in free text, possibly formalized in letters or reports.
8. **Medical diagnoses**, coded a posteriori by the doctors who treated the patient, or by specialized technicians reading the letters[42].

This data can be structured (which can be used directly by an algorithm) or unstructured (they are stored without a predefined format, such as the text of reports or medical letters, and are interpreted by humans). Machines generally produce raw structured information (eg medical biology measurements), while healthcare professionals exchange unstructured information with high interpretative value (eg a diagnosis). Medical records are created by aggregating information from different sources : (Table 2.1) gives an overview of this data[48].

Category	Nature	Structure
Administrative data	Values, Text	Structured
Communications between caregivers (Transmissions and medical observations, medical letters, reports)	Text	Unstructured
Data managed by centralized pharmacies	Values	Structured
Medical biology results	Values, Text	Structured
Data from monitoring devices	Values	Structured
Image Data Images	Images, Text	Unstructured
PMSI data and codes	Codes, Values	Structured

TABLE 2.1 – Origin, nature and structure of medical record data.

2 Medical records management

Self-tracking and documenting information about aspects of one's personal and daily life has a long history. It is an effective method which helps us to learn more about ourselves, rather than depending on our limited memory[32].

The medical record is a multifunctional document that is used to communicate and document critical information about patients' medical care among health care professionals. Comprehensive medical records are a cornerstone in the quality and efficiency of patient care during the hospitalization and in subsequent follow-up visits, as they can provide a complete and accurate chronology of treatments, patient results and future plans for care[64], it involves many kinds of records, including patient charts, x-rays, images, scans, and even emails. Additionally, it involves making sure all of these items are accessible, safe, and secure. There are multiple electronic records used in the healthcare.

2.1 Electronic medical record

Electronic medical record (EMR) systems, defined as "an electronic record of health-related information on an individual that can be created, gathered, managed, and consulted by authorized clinicians and staff within one health care organization," have the potential to provide substantial benefits to physicians, clinic practices, and health care organizations.

It is a digital version of the paper medical record that has been used for years and it will contain the patient's medical and surgical history, allergy information, treatment history, current, and past prescriptions, and other pertinent information that can be used in making future medical decisions[24]. These systems can facilitate workflow and improve the quality of patient care and patient safety[6].

2.2 Electronic health record

An Electronic Health Record (EHR) is an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the key administrative clinical data relevant to that persons care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports. The EHR automates access to information and has the potential to streamline the clinician's workflow. It also has the ability to support other care-related activities directly or indirectly through various interfaces, including evidence-based decision support, quality management, and outcomes reporting.

EHRs are the next step in the continued progress of healthcare that can strengthen the relationship between patients and clinicians. The data, and the timeliness and availability of it, will enable providers to make better decisions and provide better care[5].

2.3 The Difference Between EMR & EHR

Both the EMR and EHR contain electronic versions of a patient's medical history. Most of the information in an EMR goes into an EHR.

The EMR can contain medical history, diagnoses, medications, immunizations and dates, allergies, etc. Often, a patient needs to ask for a printed copy of an EMR to share with another medical provider.

The EHR contains similar details as the EMR, but also other relevant data like information from wearable devices, demographics, and insurance information. It can also contain lab data and imaging reports that come from other offices or practices. Assuming the software is compatible, other offices and practices can access the information within an EHR to help coordinate care and make clinical decisions[13] (Table 2.2) gives an overview of this difference :

EMR (electronic medical record)	**EHR (electronic health record)**
Medical and clinical data gathered in one provider's office	Medical and clinical data gathered from many providers' offices and hospitals
Narrower view	Broader view
Digital version of a paper chart in one office	Digital version of varied health information
Not designed for sharing	Designed for sharing outside of an individual medical practice
Providers use mainly for diagnosis and treatment	Providers have access to many diagnostic tools to make decisions

TABLE 2.2 – EMR vs. EHR : Similarities and Differences.

2.4 Personal health record

Electronic Personal Health Records (ePHRs) are a representation of health records connected to the care of a patient and are managed by the patient[41], unlike EHRs, which are managed by health care providers. ePHRs allow healthcare consumers the luxury of deciding which health information to share with healthcare providers[59]. Ozok et al.[15] defined ePHR systems as patient

centric, multi-functional, health management systems developed for managing and storing lifelong personal health information for various purposes from chronic to critical, medical and preventive care[33].

The information in an EHR is keyed in by healthcare providers and is only accessible to healthcare providers. In addition, an EHR might only contain information from a single healthcare provider. On the other hand, an individual will retain control of their own ePHR, which might encompass health information from different sources, such as various healthcare providers, as well as from the patient, as integrated ePHRs have the capability to incorporate data from different sources. Thus, at any one time, there may be various EHRs for one person but only one ePHR[33].

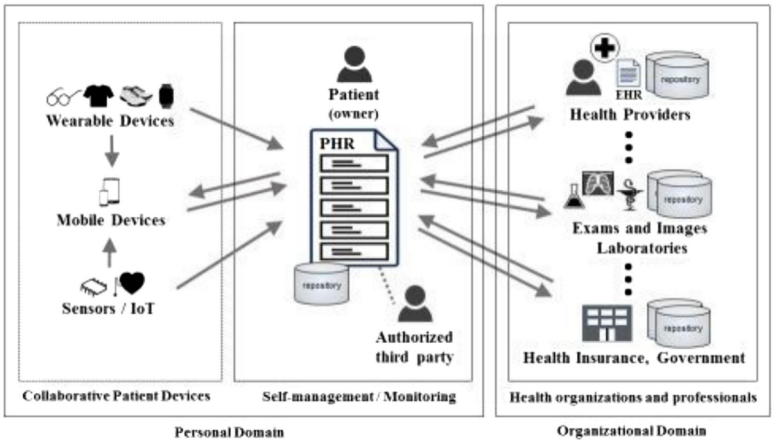


FIGURE 2.3 – Difference between an EHR and ePHR (Taken from [33]).

3 Security in medical records management

The actual technology of an electronic medical record seems to be falling into place. As the world moves more toward the use of telemedicine to preclude the movement of patients to more advanced facilities, the need for a fully functioning medical record is paramount. One of the important ethical issues in electronic records management involves privacy. Privacy is the “claim of individuals to be left alone, free from surveillance or interference from other individuals, organizations, including state” (Laudon and Laudon 2005 :159), privacy deals with the collection and use or misuse of data. Data is constantly being collected and stored on each of us. This data is often distributed over easily accessed networks and without our knowledge or consent[53].

There are a number of security dilemmas in electronic records management. There can be illegal access and use of records, data alteration and destruction (Stair and Reynolds 2006 : 583).

Typical cross-organizational e-health applications are[55] :

- sharing of patient records among different healthcare professionals.
- Access to distributed EHRs any place and any time.
- On-line teleconsultation, telemonitoring and assistance.
- Patient-doctor consultation services.
- Patients' access to their own EHRs.

That is why Electronic health records management attracts significant international interest and sets the scenery for the establishment of a distributed, coalition-based, security policy enhanced records exchange framework among different medical domains. Several European projects have proposed candidate solutions for secure inter-operations between medical domains. In the HARP project, security profiles related to access rights are dynamically downloaded to the client side. The MEDITRAV EUproject attempts to overcome national or linguistic barriers by adopting the solution of a multilingual portable personal record. These approaches, pose mainly their research effort on the security requirements for effective electronic health record management, still they confront mainly to stable infrastructures[36].

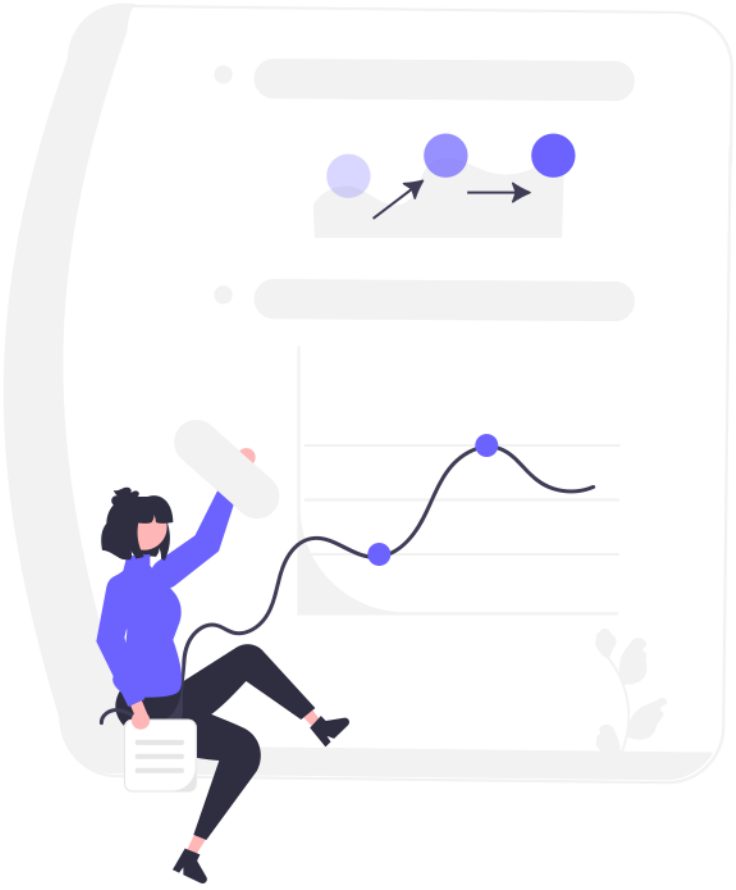
4 Conclusion

We have seen in this chapter the principal components of the healthcare sector from its principal actors through its activities to the data generated from each activity. Then we moved to the different medical electronic records management types, and we presented the differences between them. Finally, we moved on to the importance of the security field in the treatment of the patient data and medical data, then we did a quick overview of some of the projects implemented in this regard.

In the next chapter, we will present the Information visualization field and its applications.

Chapitre 3

Information Visualization -infoVis-



Human mind is very visual, following Williams et al., visualization is “a cognitive process performed by humans in forming a mental image of a domain space. In computer and information science it is, more specifically, the visual representation of a domain space using graphics, images, animated sequences, and sound augmentation to present the data, structure, and dynamic behavior of large, complex data sets that represent systems, events, processes, objects, and concepts”[63]. The (Figure 3.1) below presents the Florence Nightingale’s ’Rose diagram’ published in 1858 showing the reduction in the number of deaths in military hospitals in Scutari arising from the changes she instituted[58]
.

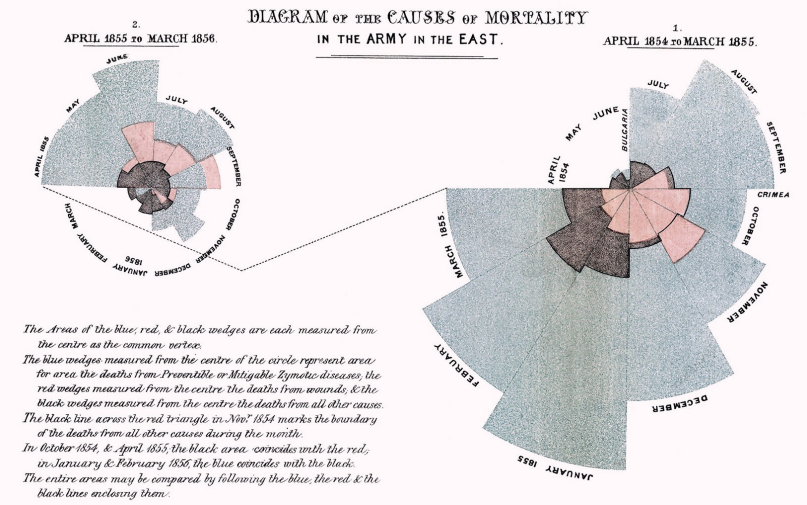


FIGURE 3.1 – Rose Diagram

Data visualization involves presenting data in graphical or pictorial form, which makes the information easy to understand. It helps to explain facts and determine courses of action. It will benefit any field of study that requires innovative ways of presenting large, complex information[58].

Traditionally, data visualization has been structured along two main fields : scientific visualization and information visualization. A third, newer field, called visual analytics has emerged in the past several years, as a bridge between and also an extension of the former two fields[60]. In this paper, we will focus mainly on the information visualization field.

1 Definition

Information visualization (InfoVis) is the practice of representing data in a meaningful, visual way that users can interpret and easily comprehend, it is a research area that aims to aid users in exploring, understanding, and analyzing data through progressive, iterative visual exploration. With the boom in big data analytics, InfoVis is being widely used in a variety of data analysis applications in different domains, ranging from finance to sports to politics[47].

Information visualizations are often created with an audience in mind and designed to display certain important information that they need to understand. With an idea of how the visualization will be used, using multiple tools (Column chart, Bar graph, Network graph, Stacked bar graph, Histogram, Line chart, Pie chart, Box plot, Bubble chart, Dual-axis chart,...) that can help users compare different values, show the bigger picture, track trends in the data, and understand different relationships between variables[21]. These tools follow the model of the visualization pipeline.

FIGURE 3.2 – Information visualization tools

2 Visualization pipeline

A visualization pipeline embodies a dataflow network in which computation is described as a collection of executable modules that are connected in a directed graph representing how data moves between modules. In a basic pipeline (Figure 3.3), there are three types of modules : sources, filters, and sinks. A source module produces data that it makes available through an output. File readers and synthetic data generators are typical source modules. A sink module accepts data through an input and performs an operation with no further

result (as far as the pipeline is concerned). Typical sinks are file writers and rendering modules that provide images to a user interface. A filter module has at least one input from which it transforms data and provides results through at least one output[51].

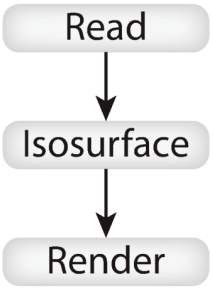


FIGURE 3.3 – A simple visualization pipeline.

As science progresses, this model has been detailed, Figure 3.4 provides an overview of the infoVis pipeline. It has five main modules : Data Analysis, Filtering, Mapping, Rendering, Image data, explained as follows :

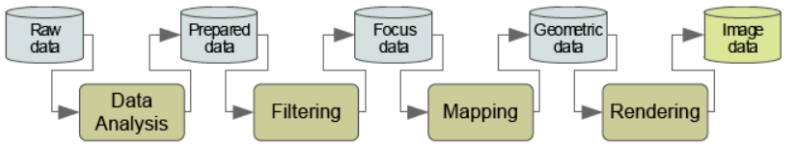


FIGURE 3.4 – A visualization pipeline describes the process of creating visual representations of data.

1. ***Raw data :*** First, we have to import the data. This implies finding a representation of the original information we want to investigate in terms of a data set. Practically, importing data means choosing a specific dataset implementation and converting the original information to the representation implied by the chosen dataset in order to turn this data into information using Data analysis.
2. ***Data Analysis :*** Is the process of bringing order and structure to collected data. Mostly using data warehouse systems (3), it turns data into information that teams can use. Analysis is done using systematic methods to look for trends, groupings, or other relationships between different types of data[3], following this process :
   - ***Data Requirements Specification :*** The data required for analysis is based on a question or an experiment. Based on the requirements

of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data may be numerical or categorical[2].

- ***Data Collection :*** Guided by the requirements identified, Data can be collected through several sources, including online sources, computers, personnel, and sources from the community.
- ***Data processing :*** The data that is collected must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (known as structured data) for further analysis, often through the use of spreadsheet or statistical software[27].
- ***Data Cleaning :*** The processed and organized data may be incomplete, contain duplicates, or contain errors. Data Cleaning is the process of preventing and correcting these errors[2].
- ***Perform data analysis :*** One of the last steps in the data analysis process is analyzing and manipulating the data. This can be done in a variety of ways depending on the cleaned data nature[23].

The data analysis step produces the prepared data.

3. ***Filtering :*** Data filtering is the process of choosing a smaller part of your data set and using that subset for viewing or analysis[20], this portion of data called focus data.
4. ***Mapping :*** Focus data are mapped to geometric primitives (e.g., points, lines) and their attributes (e.g., color, position, size) ; most critical step for achieving expressiveness and effectiveness.
5. ***Rendering :*** The rendering operation is the final step of the visualization process, rendering takes the geometric data created by the mapping operation and transforms it to an image data.

3 Data Warehouse Systems

The concept of "data warehousing" arose in the mid-1980s with the intention to support huge information analysis and management reporting[62]. Data warehouse was defined According to Bill Inmon a “subject-oriented, integrated, time variant and non-volatile collection of data in support of management’s decision making process”[45].

According to Ralph Kimball "a data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making"[46].

There are three major areas in the data warehouse architecture as following :

- ***Data Acquisition :*** This step covers the process of extracting data from the multi sources, moving all extracted data to the stage and preparing the data for loading into the repository. The two main architectural components of this area are the source data and the data store, which is where all the extracted data is gathered and prepared for loading into the data warehouse.
- ***Data processing and storage :*** This stage covers all preparations and analysis that take place on our data from data cleaning to data processing to filtering till the rendering step (the steps mentioned in section2). At this stage, the data goes through a series of transformations to extract the focus data into clear and clean formats to build a sort of common language.
- ***Information visualization :*** This step focuses mainly on the visualization part, it makes it easy for the users to access the information directly from the data warehouse.

4 Data Integration Approaches

Data integration is the most tedious and time-consuming step in setting up a decision-making information system. During this step, the data is transformed and filtered to represent a homogeneous, common and stable source of information. The performance of the SID is closely linked to the quality of data integration. It should be noted that the data integration step is not limited to the decision-making domain. It is more general and can be applied for different needs : bringing together and requesting several operational information systems, communicating applications that have been made in silos (independently of each other), etc.

Several approaches have been developed depending on the integration needs. We present the most used approaches[35] :

- ***Extract Transform and Load (ETL) :***
  This is the most used approach in setting up a data warehouse. In this approach, the integration is done in three steps :
  — Extracting data from sources.
  — Data transformation, which consists of cleaning and aggregating data to integrate them into a predefined schema.
  — Loading data into the target (the data warehouse).
- ***Enterprise Information Integration(EII) :***
  In the EII approach, no physical integration is performed. Heterogeneous data sources are consolidated using a virtual database, transparent to

applications using the data. The virtual database provides a unified view of data. Users send their request directly to the database. The query is then broken down into sub-queries that will be sent to the respective sources. The answers are assembled into a final result.

- ***Enterprise Application Integration(EAI) :***
  In order to connect applications built in different environments and with different technologies, the EAI approach is based on application integration and sharing of their data using web services (SOA architecture). This approach allows for real-time communication. It is also used to feed data warehouses. This approach does not replace ETL.

Criteria	**ETL**	**EII**	**EAI**
Data flow	Unidirectional	Bidirectional	Bidirectional
Latency	Daily to monthly	Real time	Real time
Data transformation	Big capacity	Medium capacity	Low capacity
Context of use	-Consolidation of a large amount of data. -Complex transformations.	-Link an existing warehouse with specific data sources. -Source data volatile and accessible using simple queries.	-Sources not directly accessible. -Simple queries.

TABLE 3.1 – Comparison between different integration approaches.

5 Conclusion

In this chapter, we went through the definition of Information visualization, presenting the visualization pipeline process, and then moving to the data warehouse and its different data integration approaches.
In the next chapter, we present our contribution based on the different concepts we have already seen in the previous chapters.

Chapitre 4

Contribution & Discussion

In this chapter, we propose our data management and visualization system for medical data that aims to manage the uncommon data from the various sources then provides a personalized dashboard for each actor based on their predefined needs.

1 Work Objectives

The aim of this work is to propose a system that organizes data coming from different sources then structures it in a way that enables easily to visualize it in most convenient ways (both useful and easy to read) for each actor. We are interested in the step between Rendering and Data image in the pipeline process(3.4). After reviewing and analyzing similar proposed works, we created our own visualization system that will be detailed in the coming sections.

2 Literature & Related works review

We present in this section architecture for healthcare data management systems, data warehouses and solutions that attempt to integrate infoVis into medical data and medical structures, which could be used by executive managers, doctors, physicians and other health professionals to support the healthcare process. Medical data existing today in multiple sources with different formats makes it necessary to have certain data integration techniques. A healthcare data warehouse is therefore needed to integrate the different data sources into a central data repository and analyze this data.

- ***A Healthcare Data Warehouse for Cancer Diseases :*** Dr.Osama E.Sheta and Ahmed Nour Eldeen discussed in their paper[57] the implementation of a healthcare data warehouse for cancer diseases, they proposed two stages approach for the building cancer data warehouse :

  1. *Business Analysis :* Consist of business process analysis and business requirement analysis (Figure4.1).

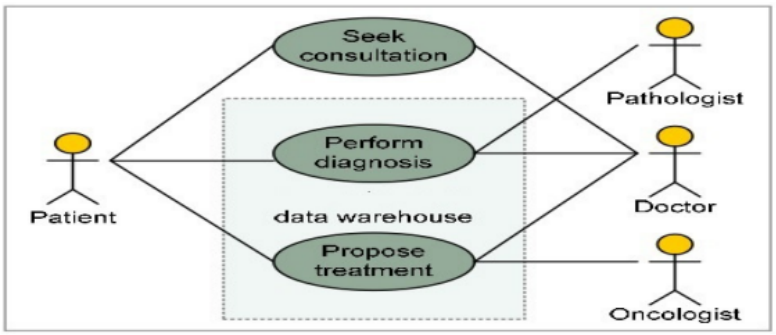


FIGURE 4.1 – Cancer data warehouse use case diagram.

2. *Architecture Design :* Data is imported from several sources and transformed within a staging area before it is integrated and stored in the production data warehouse for further analysis (Figure4.2).

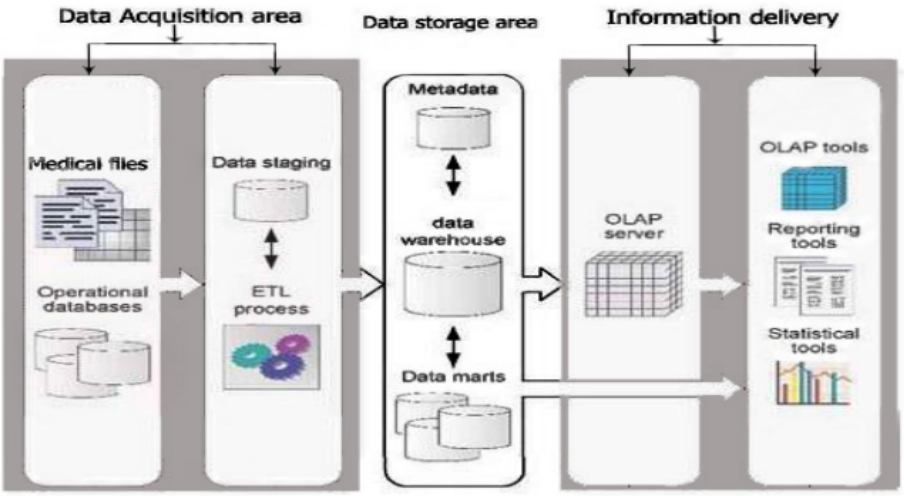


FIGURE 4.2 – Cancer data warehouse Architecture Taken from the source.

- ***Data Warehouse Framework in Pharmaceutical Sector :*** In this paper[29] authors proposed a data warehouse framework to enhance decisions of distribution systems in pharmaceutical companies to decrease the medicine industry cost and increase productivity. The framework can be described in four phases shown in (Figure 4.3). Phase one consists of a data preparation, a phase which has four steps (data collection, building DBs, DWH and data cleaning). Phase two consists of training the data which is applying time series to three types of Neural Networks techniques (levenberg marquardt, Bayesian regularized, and Scaled conjugate gradient). Phase three is testing the performance based on mean square error (MSE). Phase four consists of evaluating the performance of the best prediction model.

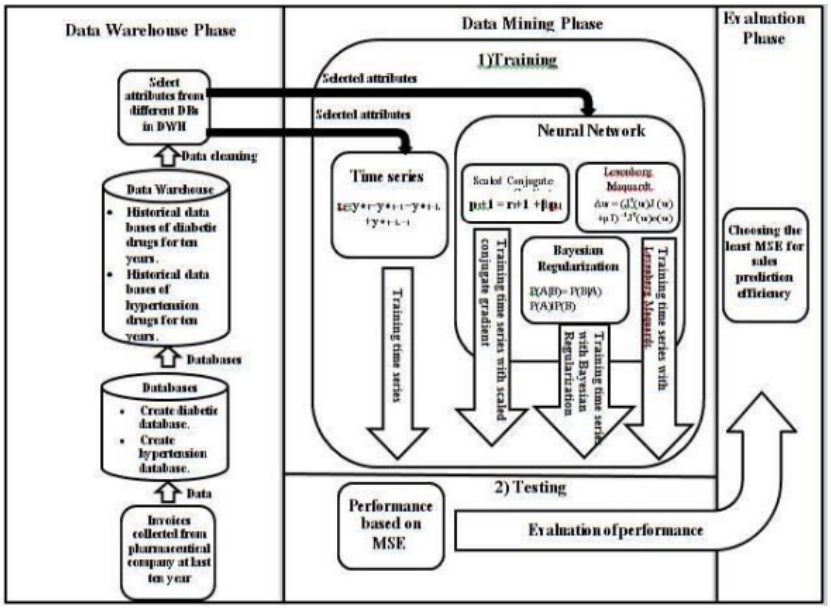


FIGURE 4.3 – The Proposed Framework of Sales Prediction.

- ***Big Bata Warehouse Based On Hadoop Architecture :*** In this paper[56] entitled “Medical Big Data Warehouse : Architecture and System Design, a Case Study : Improving Healthcare Resources Distribution” authors proposed a system architecture and a conceptual data model for a MBDW (Medical Big Data Warehouse), and then offer a solution to overcome both the growing of fact table size and the lack of primary and foreign keys in the framework Apache Hive required in the conceptual data model. This solution is based on nested partitioning according to the dimension tables keys, then applying their solution to implement a MBDW to improve medical resources distribution for the health sector in the Bejaia region (in Algeria).

  The overall architecture is depicted in (Figure 4.4). It is a scalable, reliable, and distributed architecture to extract, store, analyze, and visualize healthcare data extracted from various resources HIS (Hospitals Information systems).

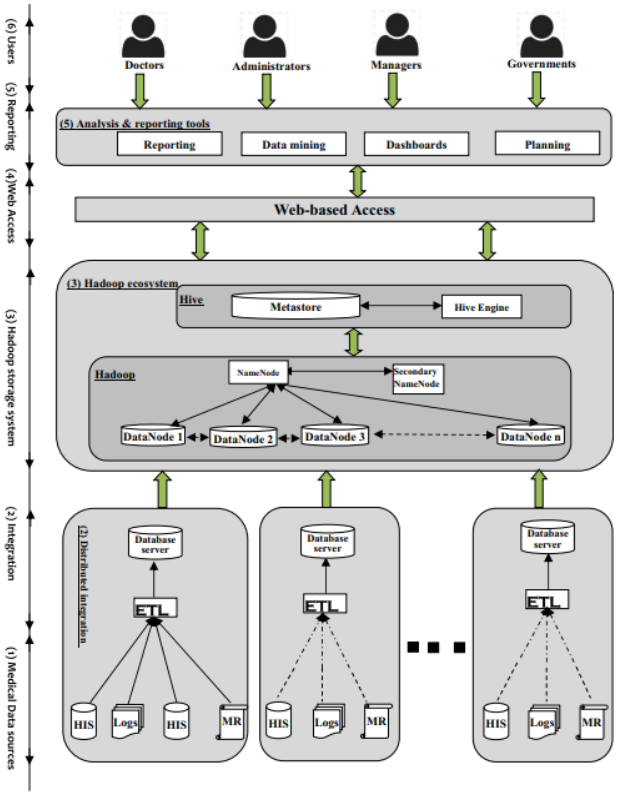


FIGURE 4.4 – Hadoop-based system architecture of medical big data warehousing.

3 Proposed Solution

As mentioned in the project description, the presence of a visualization presentation in the medical sector is necessary. The proposed solution was to create a data management and visualization system that organizes and structures the data coming from various sources, then produces focus data in the eXtensible Markup Language (XML) format, this focus data will be visualized based on each actor and their preferend needs.
We mainly focused on the "Rendering" step in the visualization pipeline(3.4), and we worked on the medical data type that are mentioned in the Table2.1.

4 Why XML?

The rise of XML (eXtensible Markup Language) in patient care has been driven by the needs for communication among health professionals and between healthcare organizations such as hospitals and health insurance companies.The main advantage of XML is its flexibility, as it allows creators to describe any content easily by generating their own tags[61]. Some of XML's features are[30] :

- The XML and DTD files are human readable and thus can be easily edited by people with only a few computer skills. Updating a data model is, therefore, straightforward (at least from a technical point of view).
- XML is Internet-oriented and has very rich capabilities for linking data; this can be used for interconnecting databases.
- XML provides an open framework for defining standard specifications. This is an important point because medical informatics clearly lacks standardization. For example, querying on multiple molecular biology databases could be greatly facilitated if each database would offer an XML view of their content.

On the other hand, XML has some weaknesses :

- The overhead of a text based format in data parsing, storage and transmission needs to be evaluated before adopting XML as a general solution. However, a text format means that the source code can be read and edited with any text editor.
- It is not clear whether XML satisfactorily addresses the problems of technological scalability. Indeed if XML data are stored in flat files, queries on XML files will not scale because XML in itself does not provide scalable facilities such as indexing or data clustering. This means that parsing should be done on the fly which leads to poor performances. One solution could be to have query optimizations done externally for example

using a Database Management System (DBMS).

At this point the question to be answered is whether the pros prevail over the cons, for this reason Frederic Achard and al[30] have provided a comparison between XML and some of the most popular solutions that are used for the management and exchange of bioinformatics data summarized in the (Table4.1), each one is rated with one to four stars for different criteria : the higher the number of stars, the better the solution with regard to the criteria.

Criteria	**XML**	**Field/ value**	**ASN.1**	**CORBA**	**Java RMI**	**OODBMS**
Model expressiveness	**	*	***	***	***	****
Constraints	**	*	*	**	***	****
Self-descriptive	yes	no	yes	yes	yes	yes
Query language	soon[a]	no	no	soon[b]	no	yes
Flexibility	****	*	***	***	***	****
Simplicity	****	****	***	*	**	**
Scalability	**	*	**	***	***	****
Interoperability	****	*	**	****	****	***

TABLE 4.1 – Summary of comparison of different alternatives to XML.

They conclude that the use of XML as an intermediate medium would be really efficient only if all databases share common or very similar DTDs. Whatever language is used, it is always difficult to find an agreement on a common semantics, and when one is found, it is often revised. However, XML would be an excellent candidate for this role because of its flexibility.

5 The Proposed System Process

Our system process is divided into three main phases as illustrated in the following figure. In this part, we will explain in detail the steps of each phase.

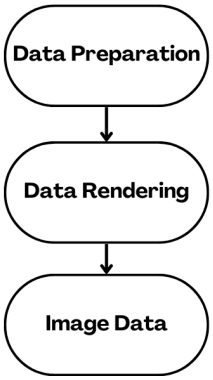


FIGURE 4.5 – The proposed system architecture.

5.1 The Data Preparation

It consists of 4 stages : The Data Collection, pretreatment, Data Integration & Treatment, and output data as illustrated in the following figure.

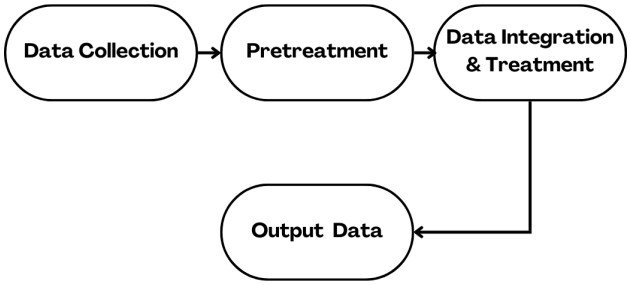


FIGURE 4.6 – The Data preparation flowchart.

The Data Collection

The datasets chosen for this work come from kaggle[12]. Kaggle is an online community of data scientists and machine learning practitioners[28], it enables users to find and publish datasets, explore and build models in a data science environment based on the web, etc.
For the purposes of the creation process, we collected the following data (each is between 100K and 240K line) in csv format :

- **Drugs Prescriptions with Providers Profile Dataset :** It contains recommended drugs based on medical practitioners practicing specialty, years of practicing and previous suggested drugs in prescriptions by other providers or practitioners (Figure4.7).

#	specialty	years_pra...	cms_presc...
0	Nephrology	7	DOXAZOSIN MESYLATE, MIDODRINE HCL, MEGESTROL ACETATE, BENAZEPRIL HCL, METOLAZONE, NOVOLOG, DIAZEPAM, ...
1	General Practice	7	CEPHALEXIN, AMOXICILLIN, HYDROCODONE-ACETAMINOPHEN
2	General Practice	7	CEPHALEXIN, AMOXICILLIN, CLINDAMYCIN HCL
3	General Practice	7	AMOXICILLIN

FIGURE 4.7 – A sample of Drugs prescriptions dataset.

- **Health Prescription Dataset :** It contains the patient's diagnosis, the admission type and the dates related (Figure4.8).

SUBJECT_ID	ROW_ID	HADM_ID	CATEGORY	ADMISSION_TYPE	DIAGNOSIS	TEXT
273 – 99.8k	1 – 13.2k	100k – 200k	1 unique value	EMERGENCY 85% ELECTIVE 14% Other (7) 1%	PNEUMONIA 4% SEPSIS 3% Other (685) 92%	744 unique values
26880	178	135453	Discharge summary	EMERGENCY	S/P FALL;TELEMETRY	Admission Date: [**2162-3-3**] Discharge Date: [**2162-3-25**] Date of Birth: [**2...
42130	181	114236	Discharge summary	ELECTIVE	LEFT SPHENOID MENENGIOMA/SDA	Admission Date: [**2150-2-25**] Discharge Date: [**2150-3-1**] Date of Birth: [**2...
76874	212	113329	Discharge summary	EMERGENCY	TYLENOL BENZO OVERDOSE	Admission Date: [**2101-10-25**] Discharge Date: [**2101-10-28**] Date of Birth: [...
66479	220	134640	Discharge summary	EMERGENCY	PEDISTRIAN STRUCK	Admission Date: [**2148-2-3**] Discharge Date: [**2148-2-7**] Date of Birth: [**20...
31502	8	125483	Discharge summary	EMERGENCY	RESPIRATORY FAILURE;CONGESTIVE HEART FAILURE	Admission Date: [**2174-5-29**] Discharge Date: [**2174-6-9**] Date of Birth: [**2...

FIGURE 4.8 – A sample of Health Prescription dataset.

- We have collected a collection of x-ray images and other images type.

- **Liver Patient Records Dataset :** This data set contains liver and non-liver patient records collected from North East of Andhra Pradesh, India. The dataset contains the age and the gender of the patient, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase and others (Figure4.9).

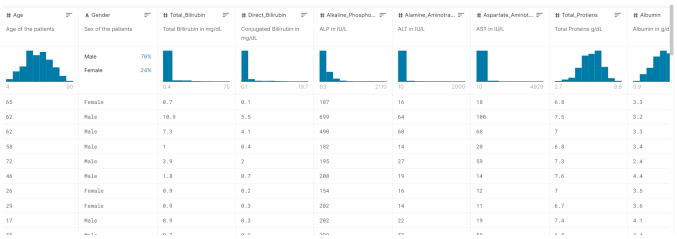


FIGURE 4.9 – A sample of Liver Patient Records dataset.

- **Disease Symptom Prediction Dataset :** There are columns containing diseases, their symptoms , precautions to be taken, and their weights (Figure4.10).

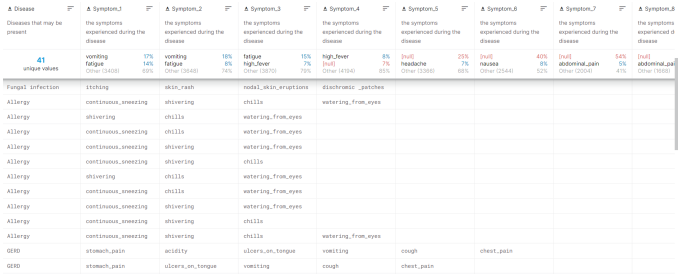


FIGURE 4.10 – A sample of Disease Symptom Prediction dataset.

- **Patient Treatment Classification :** This dataset is Electronic Health Record Predicting collected from a private Hospital in Indonesia. It contains the patient's laboratory test results : name, data type, value sample, description (Figure4.11).

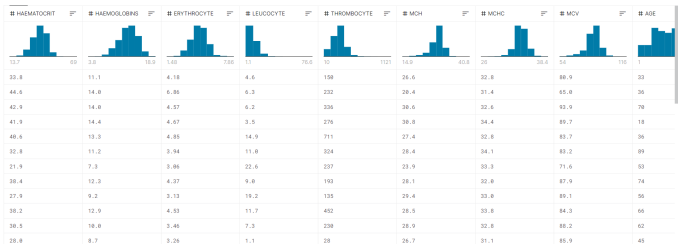
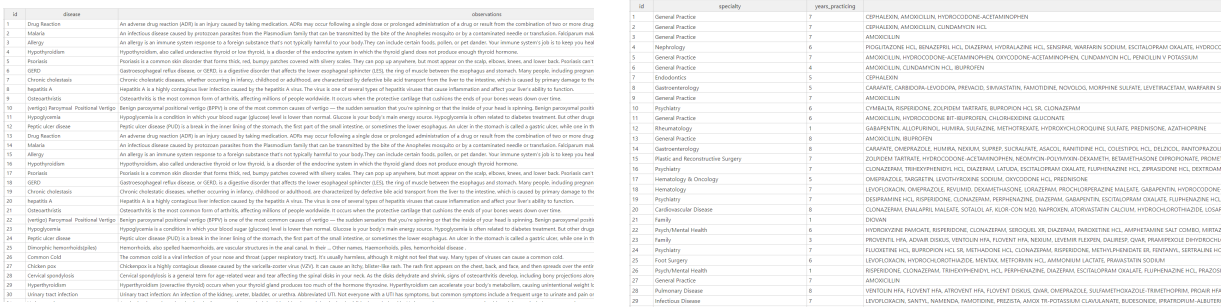


FIGURE 4.11 – A sample of Patient Treatment Classification dataset.

Pretreatment

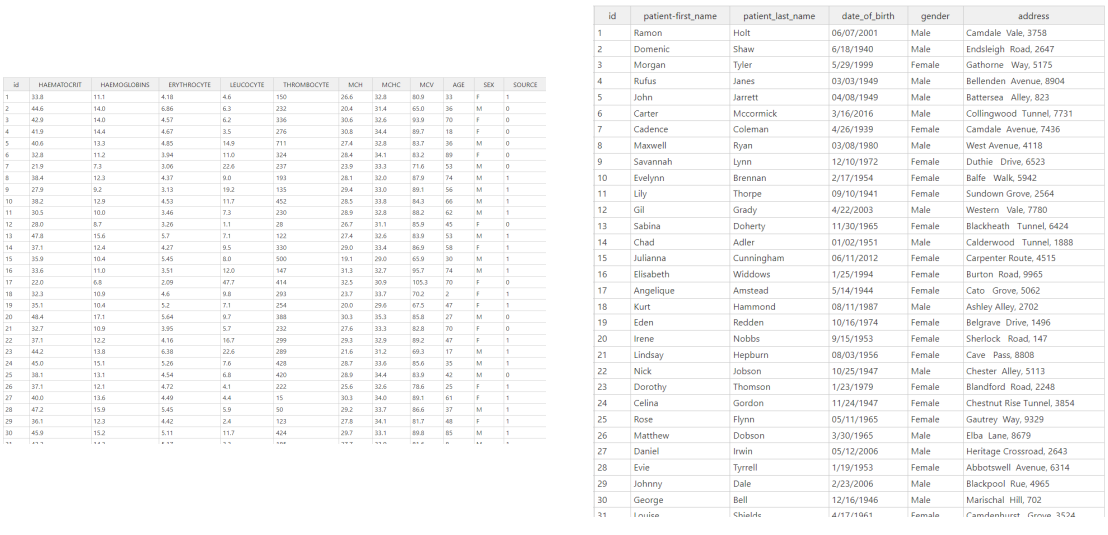
We wanted to present the entered data as if it comes from several medical actors, for this we created a fake data based on the dataset collected before, using python libraries (1.1) to present the patients information, their lab test results, their diagnosis etc.



id	disease	observations
1	Drug Reaction	An adverse drug reaction (ADR) is an injury caused by taking medication. ADRs may occur following a single dose or prolonged administration of a drug or result from the combination of two or more drug
2	Malaria	An infectious disease caused by protozoan parasites from the Plasmodium family that can be transmitted by the bite of the Anopheles mosquito or by a contaminated needle or transfusion. Falciparum mal
3	Allergy	An allergy is an immune system response to a foreign substance that's not typically harmful to your body.They can include certain foods, pollen, or pet dander. Your immune system's job is to keep you heal
4	Hypothyroidism	Hypothyroidism, also called underactive thyroid or low thyroid, is a disorder of the endocrine system in which the thyroid gland does not produce enough thyroid hormone.
5	Psoriasis	Psoriasis is a common skin disorder that forms thick, red, bumpy patches covered with silvery scales. They can pop up anywhere, but most appear on the scalp, elbows, knees, and lower back. Psoriasis can't
6	GERD	Gastroesophageal reflux disease, or GERD, is a digestive disorder that affects the lower esophageal sphincter (LES), the ring of muscle between the esophagus and stomach. Many people, including pregnan
7	Chronic cholestasis	Chronic cholestatic diseases, whether occurring in infancy, childhood or adulthood, are characterized by defective bile acid transport from the liver to the intestine, which is caused by primary damage to th
8	hepatitis A	Hepatitis A is a highly contagious liver infection caused by the hepatitis A virus. The virus is one of several types of hepatitis viruses that cause inflammation and affect your liver's ability to function.
9	Osteoarthritis	Osteoarthritis is the most common form of arthritis, affecting millions of people worldwide. It occurs when the protective cartilage that cushions the ends of your bones wears down over time.
10	(vertigo) Paroxysmal Positional Vertigo	Benign paroxysmal positional vertigo (BPPV) is one of the most common causes of vertigo — the sudden sensation that you're spinning or that the inside of your head is spinning. Benign paroxysmal positi
11	Hypoglycemia	Hypoglycemia is a condition in which your blood sugar (glucose) level is lower than normal. Glucose is your body's main energy source. Hypoglycemia is often related to diabetes treatment. But other drugs
12	Peptic ulcer disease	Peptic ulcer disease (PUD) is a break in the inner lining of the stomach, the first part of the small intestine, or sometimes the lower esophagus. An ulcer in the stomach is called a gastric ulcer, while one in th
13	Drug Reaction	An adverse drug reaction (ADR) is an injury caused by taking medication. ADRs may occur following a single dose or prolonged administration of a drug or result from the combination of two or more drug
14	Malaria	An infectious disease caused by protozoan parasites from the Plasmodium family that can be transmitted by the bite of the Anopheles mosquito or by a contaminated needle or transfusion. Falciparum mal
15	Allergy	An allergy is an immune system response to a foreign substance that's not typically harmful to your body.They can include certain foods, pollen, or pet dander. Your immune system's job is to keep you heal
16	Hypothyroidism	Hypothyroidism, also called underactive thyroid or low thyroid, is a disorder of the endocrine system in which the thyroid gland does not produce enough thyroid hormone.
17	Psoriasis	Psoriasis is a common skin disorder that forms thick, red, bumpy patches covered with silvery scales. They can pop up anywhere, but most appear on the scalp, elbows, knees, and lower back. Psoriasis can't
18	GERD	Gastroesophageal reflux disease, or GERD, is a digestive disorder that affects the lower esophageal sphincter (LES), the ring of muscle between the esophagus and stomach. Many people, including pregnan
19	Chronic cholestasis	Chronic cholestatic diseases, whether occurring in infancy, childhood or adulthood, are characterized by defective bile acid transport from the liver to the intestine, which is caused by primary damage to th
20	hepatitis A	Hepatitis A is a highly contagious liver infection caused by the hepatitis A virus. The virus is one of several types of hepatitis viruses that cause inflammation and affect your liver's ability to function.
21	Osteoarthritis	Osteoarthritis is the most common form of arthritis, affecting millions of people worldwide. It occurs when the protective cartilage that cushions the ends of your bones wears down over time.
22	(vertigo) Paroxysmal Positional Vertigo	Benign paroxysmal positional vertigo (BPPV) is one of the most common causes of vertigo — the sudden sensation that you're spinning or that the inside of your head is spinning. Benign paroxysmal positi
23	Hypoglycemia	Hypoglycemia is a condition in which your blood sugar (glucose) level is lower than normal. Glucose is your body's main energy source. Hypoglycemia is often related to diabetes treatment. But other drugs
24	Peptic ulcer disease	Peptic ulcer disease (PUD) is a break in the inner lining of the stomach, the first part of the small intestine, or sometimes the lower esophagus. An ulcer in the stomach is called a gastric ulcer, while one in th
25	Dimorphic hemorrhoids(piles)	Hemorrhoids, also spelled haemorrhoids, are vascular structures in the anal canal. In their ... Other names. Haemorrhoids, piles, hemorrhoidal disease ...
26	Common Cold	The common cold is a viral infection of your nose and throat (upper respiratory tract). It's usually harmless, although it might not feel that way. Many types of viruses can cause a common cold.
27	Chicken pox	Chickenpox is a highly contagious disease caused by the varicella-zoster virus (VZV). It can cause an itchy, blister-like rash. The rash first appears on the chest, back, and face, and then spreads over the entir
28	Cervical spondylosis	Cervical spondylosis is a general term for age-related wear and tear affecting the spinal disks in your neck. As the disks dehydrate and shrink, signs of osteoarthritis develop, including bony projections alon
29	Hyperthyroidism	Hyperthyroidism (overactive thyroid) occurs when your thyroid gland produces too much of the hormone thyroxine. Hyperthyroidism can accelerate your body's metabolism, causing unintentional weight lo
30	Urinary tract infection	Urinary tract infection: An infection of the kidney, ureter, bladder, or urethra. Abbreviated UTI. Not everyone with a UTI has symptoms, but common symptoms include a frequent urge to urinate and pain or

id	specialty	years_practicing	
1	General Practice	7	CEPHALEXIN, AMOXICILLIN, HYDROCODONE-ACETAMINOPHEN
2	General Practice	7	CEPHALEXIN, AMOXICILLIN, CLINDAMYCIN HCL
3	General Practice	7	AMOXICILLIN
4	Nephrology	6	PIOGLITAZONE HCL, BENAZEPRIL HCL, DIAZEPAM, HYDRALAZINE HCL, SENSIPAR, WARFARIN SODIUM, ESCITALOPRAM OXALATE, HYDROCC
5	General Practice	7	AMOXICILLIN, HYDROCODONE-ACETAMINOPHEN, OXYCODONE-ACETAMINOPHEN, CLINDAMYCIN HCL, PENICILLIN V POTASSIUM
6	General Practice	4	AMOXICILLIN, CLINDAMYCIN HCL, IBUPROFEN
7	Endodontics	5	CEPHALEXIN
8	Gastroenterology	5	CARVATE, CARBIDOPA-LEVODOPA, PREVACID, SIMVASTATIN, FAMOTIDINE, NOVOLOG, MORPHINE SULFATE, LEVETIRACETAM, WARFARIN SI
9	General Practice	7	AMOXICILLIN
10	Psychiatry	6	CYMBALTA, RISPERIDONE, ZOLPIDEM TARTRATE, BUPROPION HCL SR, CLONAZEPAM
11	General Practice	6	AMOXICILLIN, HYDROCODONE BIT-IBUPROFEN, CHLORHEXIDINE GLUCONATE
12	Rheumatology	1	GABAPENTIN, ALLOPURINOL, HUMIRA, SULFAZINE, METHOTREXATE, HYDROXYCHLOROQUINE SULFATE, PREDNISONE, AZATHOPRINE
13	General Practice	8	AMOXICILLIN, IBUPROFEN
14	Gastroenterology	8	CARVATE, OMEPRAZOLE, HUMIRA, NEXIUM, SUPREP, SUCRALFATE, ASACOL, RANITIDINE HCL, COLESTIPOL HCL, DELTICOL, PANTOPRAZOLE
15	Plastic and Reconstructive Surgery	7	ZOLPIDEM TARTRATE, HYDROCODONE-ACETAMINOPHEN, NEOMYCIN-POLYMYXIN-DEXAMETH, BETAMETHASONE DIPROPIONATE, PROMI
16	Psychiatry	7	CLONAZEPAM, TRIHEXYPHENIDYL HCL, DIAZEPAM, LATUDA, ESCITALOPRAM OXALATE, FLUPHENAZINE HCL, ZIPRASIDONE HCL, DEXTROAM
17	Hematology & Oncology	5	OMEPRAZOLE, TARGRETIN, LEVOTHYROXINE SODIUM, OXYCODONE HCL, PREDNISONE
18	Hematology	7	LEVOFLOXACIN, OMEPRAZOLE, REVLIMID, DEXAMETHASONE, LORAZEPAM, PROCHLORPERAZINE MALEATE, GABAPENTIN, HYDROCODONE
19	Psychiatry	7	DESIPRAMINE HCL, RISPERIDONE, CLONAZEPAM, PERPHENAZINE, DIAZEPAM, GABAPENTIN, ESCITALOPRAM OXALATE, FLUPHENAZINE HCL
20	Cardiovascular Disease	8	CLONAZEPAM, ENALAPRIL MALEATE, SOTALOL AF, KLOR-CON M20, NAPROXEN, ATORVASTATIN CALCIUM, HYDROCHLOROTHIAZIDE, LOSAF
21	Family	1	DIOVAN
22	Psych/Mental Health	6	HYDROXYZINE PAMOATE, RISPERIDONE, CLONAZEPAM, SEROQUEL XR, DIAZEPAM, PAROXETINE HCL, AMPHETAMINE SALT COMBO, IRBESA
23	Family	3	PROVENTIL HFA, ADVAIR DISKUS, VENTOLIN HFA, FLOVENT HFA, NEXIUM, LEVEMIR FLEXPEN, DALIRESP, QVAR, PRAMIPEXOLE DIHYDROCHL
24	Psychiatry	7	FLUOXETINE HCL, BUPROPION HCL SR, METHADONE HCL, CLONAZEPAM, RISPERIDONE, METHYLPHENIDATE ER, FENTANYL, SERTRALINE HC
25	Foot Surgery	6	LEVOFLOXACIN, HYDROCHLOROTHIAZIDE, MENTAX, METFORMIN HCL, AMMONIUM LACTATE, PRAVASTATIN SODIUM
26	Psych/Mental Health	1	RISPERIDONE, CLONAZEPAM, TRIHEXYPHENIDYL HCL, PERPHENAZINE, DIAZEPAM, ESCITALOPRAM OXALATE, FLUPHENAZINE HCL, PRAZOSI
27	General Practice	7	AMOXICILLIN
28	Pulmonary Disease	8	VENTOLIN HFA, FLOVENT HFA, ATROVENT HFA, FLOVENT DISKUS, QVAR, OMEPRAZOLE, SULFAMETHOXAZOLE-TRIMETHOPRIM, PROAIR HFA
29	Infectious Disease	7	LEVOFLOXACIN, SANTYL, NAMENDA, FAMOTIDINE, PREZISTA, AMOX TR-POTASSIUM CLAVULANATE, BUDESONIDE, IPRATROPIUM-ALBUTERI

FIGURE 4.12 – Samples of our data after the pretreatment.



id	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	AGE	SEX	SOURCE
1	33.8	11.1	4.18	4.6	150	26.6	32.8	80.9	33	F	1
2	44.6	14.0	6.86	6.3	232	20.4	31.4	65.0	36	M	0
3	42.9	14.0	4.57	6.2	336	30.6	32.6	93.9	70	F	0
4	41.9	14.4	4.67	3.5	276	30.8	34.4	89.7	18	F	0
5	40.6	13.3	4.85	14.9	711	27.4	32.8	83.7	36	M	0
6	32.8	11.2	3.94	11.0	324	28.4	34.1	83.2	89	F	0
7	21.9	7.3	3.06	22.6	237	23.9	33.3	71.6	53	M	0
8	38.4	12.3	4.37	9.0	193	28.1	32.0	87.9	74	M	1
9	27.9	9.2	3.13	19.2	135	29.4	33.0	89.1	56	M	1
10	38.2	12.9	4.53	11.7	452	28.5	33.8	84.3	66	M	1
11	30.5	10.0	3.46	7.3	230	28.9	32.8	88.2	62	M	1
12	28.0	8.7	3.26	1.1	28	26.7	31.1	85.9	45	F	0
13	47.8	15.6	5.7	7.1	122	27.4	32.6	83.9	53	M	1
14	37.1	12.4	4.27	9.5	330	29.0	33.4	86.9	58	F	1
15	35.9	10.4	5.45	8.0	500	19.1	29.0	65.9	30	M	1
16	33.6	11.0	3.51	12.0	147	31.3	32.7	95.7	74	M	1
17	22.0	6.8	2.09	47.7	414	32.5	30.9	105.3	70	F	0
18	32.3	10.9	4.6	9.8	293	23.7	33.7	70.2	2	F	1
19	35.1	10.4	5.2	7.1	254	20.0	29.6	67.5	47	F	1
20	48.4	17.1	5.64	9.7	388	30.3	35.3	85.8	27	M	0
21	32.7	10.9	3.95	5.7	232	27.6	33.3	82.8	70	F	0
22	37.1	12.2	4.16	16.7	299	29.3	32.9	89.2	47	F	1
23	44.2	13.8	6.38	22.6	289	21.6	31.2	69.3	17	M	1
24	45.0	15.1	5.26	7.6	428	28.7	33.6	85.6	35	M	1
25	38.1	13.1	4.54	6.8	420	28.9	34.4	83.9	42	M	0
26	37.1	12.1	4.72	4.1	222	25.6	32.6	78.6	25	F	1
27	40.0	13.6	4.49	4.4	15	30.3	34.0	89.1	61	F	1
28	47.2	15.9	5.45	5.9	50	29.2	33.7	86.6	37	M	1
29	36.1	12.3	4.42	2.4	123	27.8	34.1	81.7	48	F	1
30	45.9	15.2	5.11	11.7	424	29.7	33.1	89.8	85	M	1

id	patient_first_name	patient_last_name	date_of_birth	gender	address
1	Ramon	Holt	06/07/2001	Male	Camdale Vale, 3758
2	Domenic	Shaw	6/18/1940	Male	Endsleigh Road, 2647
3	Morgan	Tyler	5/29/1999	Female	Gathorne Way, 5175
4	Rufus	Janes	03/03/1949	Male	Bellenden Avenue, 8904
5	John	Jarrett	04/08/1949	Male	Battersea Alley, 823
6	Carter	Mccormick	3/16/2016	Male	Collingwood Tunnel, 7731
7	Cadence	Coleman	4/26/1939	Female	Camdale Avenue, 7436
8	Maxwell	Ryan	03/08/1980	Male	West Avenue, 4118
9	Savannah	Lynn	12/10/1972	Female	Duthie Drive, 6523
10	Evelynn	Brennan	2/17/1954	Female	Balfe Walk, 5942
11	Lily	Thorpe	09/10/1941	Female	Sundown Grove, 2564
12	Gil	Grady	4/22/2003	Male	Western Vale, 7780
13	Sabina	Doherty	11/30/1965	Female	Blackheath Tunnel, 6424
14	Chad	Adler	01/02/1951	Male	Calderwood Tunnel, 1888
15	Julianna	Cunningham	06/11/2012	Female	Carpenter Route, 4515
16	Elisabeth	Widdows	1/25/1994	Female	Burton Road, 9965
17	Angelique	Amstead	5/14/1944	Female	Cato Grove, 5062
18	Kurt	Hammond	08/11/1987	Male	Ashley Alley, 2702
19	Eden	Redden	10/16/1974	Female	Belgrave Drive, 1496
20	Irene	Nobbs	9/15/1953	Female	Sherlock Road, 147
21	Lindsay	Hepburn	08/03/1956	Female	Cave Pass, 8808
22	Nick	Jobson	10/25/1947	Male	Chester Alley, 5113
23	Dorothy	Thomson	1/23/1979	Female	Blandford Road, 2248
24	Celina	Gordon	11/24/1947	Female	Chestnut Rise Tunnel, 3854
25	Rose	Flynn	05/11/1965	Female	Gautrey Way, 9329
26	Matthew	Dobson	3/30/1965	Male	Elba Lane, 8679
27	Daniel	Irwin	05/12/2006	Male	Heritage Crossroad, 2643
28	Evie	Tyrrell	1/19/1953	Female	Abbotswell Avenue, 6314
29	Johnny	Dale	2/23/2006	Male	Blackpool Rue, 4965
30	George	Bell	12/16/1946	Male	Marischal Hill, 702
31	Louise	Shields	4/17/1961	Female	Camdenhurst Grove, 3524

FIGURE 4.13 – Samples of our data after the pretreatment.

Data Integration & Treatment

After the data was collected and pre-processed, we integrated it using Talend open studio (1.2), the input data was the pretreated data we introduced it in the previous section, each entry represent an medical source, we used Talend tmap component to build our output data.

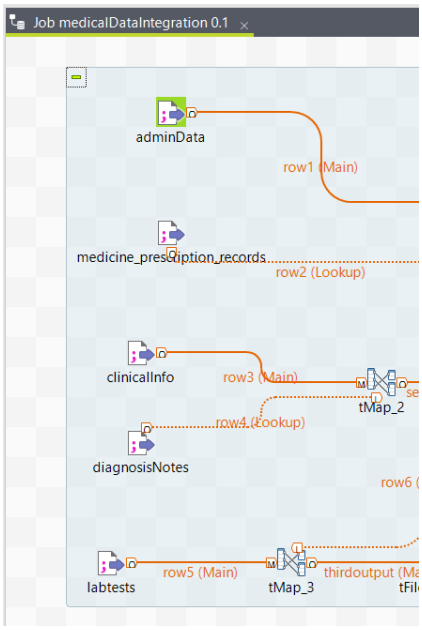


FIGURE 4.14 – Input data.

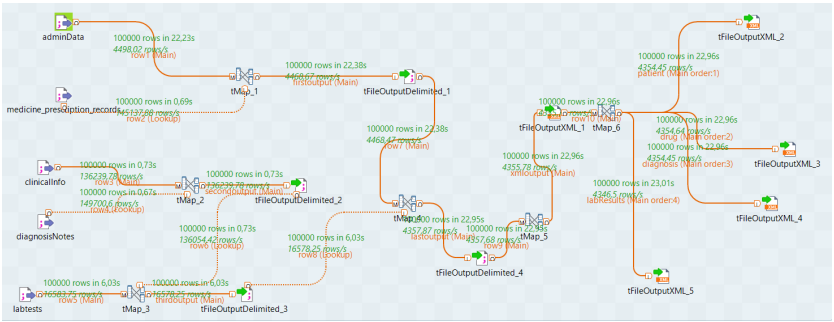


FIGURE 4.15 – Talend job execution.

Output Data

Our main objective was to extract the following data (each head represents an XML node) :

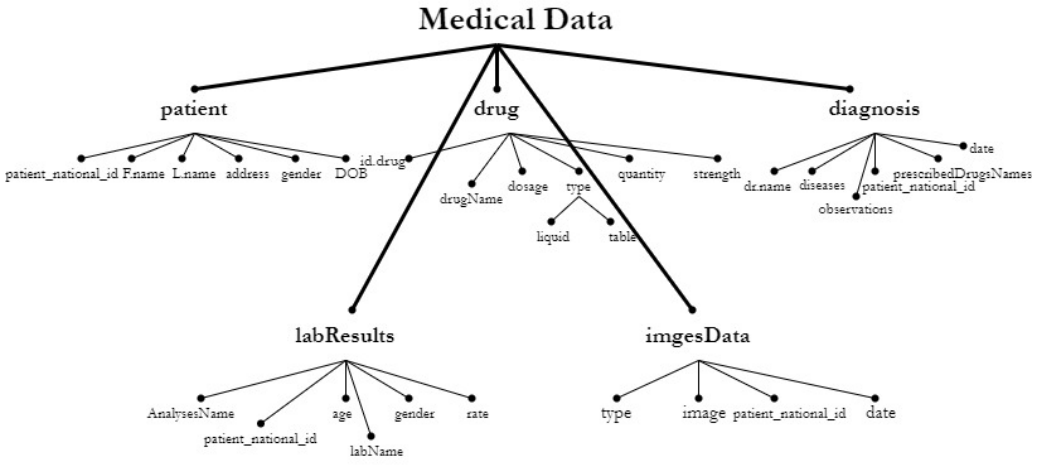


FIGURE 4.16 – The corresponding tree of XML schema.

- **Patient :** That presents the administrative information about the patient x, including his full name, national ID, gender, date of birth, address.
- **Drug :** Presents the list of drugs in the patient's prescription, it includes : the drug's name, its dosage, strength usually mentioned by the doctor, the quantity, and its type(liquid or table).
- **Diagnosis :** Includes the doctor's name, the diseases name, and the observations taken by the medical actor in charge, it also contains the patient's national ID, the name of the prescribed medications and the date of diagnosis.
- **LabResults :** It presents the medical biology results that includes the analysis's name, the patient rate saved, the gender and the age for the comparison (a predefined high/low rate is defined corresponding to each analysis), attached with the laboratory name and the patient's national ID.
- **ImagesData :** Present the data of the image obtained by the patient, it includes the image name and the image itself, the day it was taken, and the patient's national ID.

The resulting data will be used as a source to feed all the components of the digital marketing reporting applications that are built in many visualization tools(1).

5.2 Data Rendering

Once the xml output data is prepared, the data rendering phase is started, it consists of 3 stages as illustrated in the following figure :

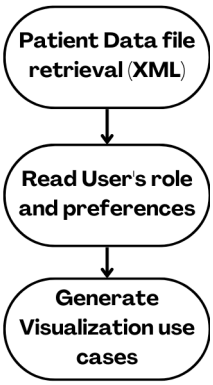


FIGURE 4.17 – The data rendering flowchart.

Patient Data file retrieval (XML)

This step consists of integrating the XML files into the visualization script. In our project process, we generated 4 files that revolve around the patient record and integrated them :

- **patient.xml :** Contains patient information.
- **diagnostic.xml :** Contains the patient's diagnosis.
- **drug.xml :** Contains the drugs prescribed to the patient.
- **labResults.xml :** Contains patient lab test results.

Read User's role and preferences

This step consists of analyzing the user's role and need in order to display the appropriate visualization graphs. For this project, we have created three predefined users : the patient, the doctor, and the administrator, each one has his own preferences.
The data will be displayed according to that.

Generate Visualization use cases

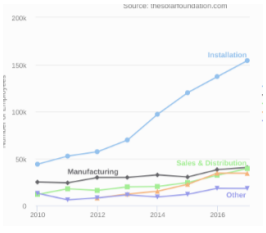
For the purpose of generating the appropriate graphs we needed to go through several analyzing steps :

- Analyzing the input data usually depends on the data properties.
- Analyzing the data type (quantitative (continuous or Discrete quantitative) or qualitative (ordinal or nominal)).
- The questions to be answered.
- How the information should be presented.
- The size of the dataset.
- And the audience (actors).

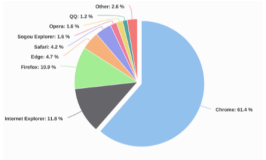
Depending on the libraries we integrated in this project, we used different types of graphs to present the quantitative data, each based on the factor mentioned previously and the frequent questions asked from our point of view.

- To express comparison between parts of a bigger set of data, highlighting different categories, or showing change over time, we used : bar charts type.
- To show relative proportions and percentages of a whole dataset or to compare the effect of ONE factor on different categories or to express nominal and not ordinal data, we used : pie charts type.
- To express a continuous dataset that changes over time or to display multiple series for the same timeline or to visualize trends instead of exact values, we used : line charts type.
- To show correlation and clustering in big datasets or the dataset containing points that have a pair of values, we used : Scatter Plot type.

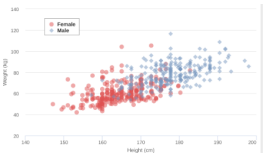
(a) Bar chart.



(b) Line chart.



(c) Pie chart.



(d) Scatter plot.

FIGURE 4.18 – The different used charts.

5.3 Image Data

The rendering data produced from the previous phase will be displayed in personalized dashboard in the next chapter.

6 Conclusion

In this chapter, we presented the proposed system process, which goes through three phases : data preparation, data rendering, Image data. The data preparation phase consists of choosing the row data and pretreating it then integrating it. The data rendering phase consists of analyze the output data to display it in the last next phase, this phase consists of three stages : Patient Data file retrieval (XML), Read User's role and preferences, then Generate Visualization use cases , and lastly the image data phase that will be explained in the next chapter.

Chapitre 5

Implementation

In this chapter, the process of implementation will be covered, starting by the architecture and the tools to the results and how everything fits together.

1 Tools

1.1 Python

Python[18] is a programming language that can be used in many contexts and is suitable for any type of use thanks to specialized libraries. However, it is particularly used as a scripting language to automate simple but tedious tasks. It is also used as a prototype development language when a functional application is needed before optimizing it with a lower level language. It is particularly widespread in the scientific world, and has many libraries optimized for numerical calculations[16].

Pandas : is a library[14] written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical arrays and time series. Pandas is free software under the BSD license.

The main data structures offered by Pandas are series (to store data according to one dimension - size according to an index), DataFrames (to store data according to 2 dimensions - rows and columns), Panels (to represent data according to 3 dimensions, 4D Panels or Data Frames with hierarchical indexes also called Multi Index (to represent data according to more than 3 dimensions - hypercube))[25].

FIGURE 5.1 – Python & Pandas Logo.

1.2 Talend

We used Talend Open Studio to create and develop
ETL processes. It is a tool based on Java and with an interface derived from that of Eclipse (Figure 5.2). It allows you to design ETL processes visually, and offers more than nine hundred components (the following list is not exhaustive)[17] :

- Connect to different data sources for reading and writing :
  - — Flat files, .xml, .csv, xls, etc.
  - — Relational databases (Postgresql, MsSql, etc) and Nosql.
- To manipulate the data, namely :
  - — Filter them.
  - — Apply aggregate functions on them.

— Sort them.
- To organize data flows.

These components are then assembled as needed to design ETL processes[50].

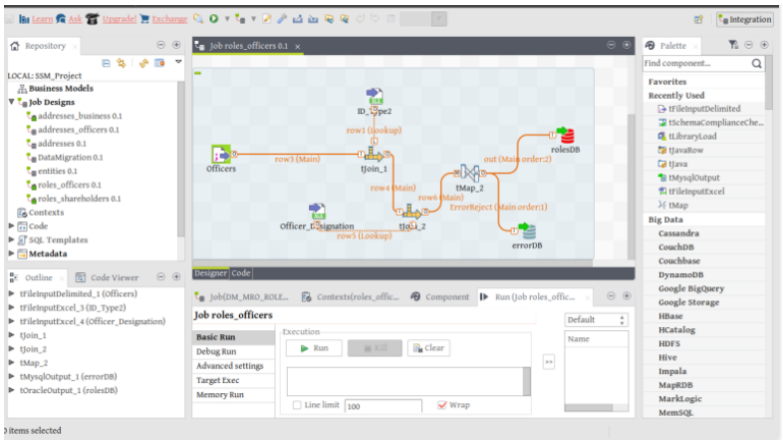


FIGURE 5.2 – Talend Interface.

1.3 Visualization Tools

1.3.1 Tableau

Tableau is an excellent data visualization and business intelligence tool used for reporting and analyzing vast volumes of data. It helps users create different charts, graphs, maps, dashboards, and stories for visualizing and analyzing data, to help in making business decisions. Tableau supports powerful data discovery and exploration that enables users to answer important questions in seconds, it can connect to several data sources that other BI tools do not support. Tableau enables users to create reports by joining and blending different datasets and it supports a centralized location to manage all published data sources within an organization[22].

FIGURE 5.3 – Tableau Logo.

1.3.2 Microsoft Power BI

Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence. It is part of the Microsoft Power Platform. According to Microsoft : "Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights"[4].

FIGURE 5.4 – Microsoft Power BI Logo.

1.3.3 Tableau Vs. Power BI

Microsoft Power BI and Tableau are two of the top business intelligence (BI) and data analytics platforms in the market. As two highly regarded analytics platforms, users often are forced to choose between Power BI and Tableau. There are arguments for and against each tool[54].

- Microsoft wins in terms of breadth of service due to its ecosystem of integrated platforms. However, Tableau perhaps comes out ahead when it comes to depth of analysis and the kind of robust, intuitive features that data scientists and analysts need for competitive advantage.
- Power BI wins on broad usage by a non-technical audience whereas Tableau has the edge with technical users.
- Power BI is that it supports various data sources but has limited access to other databases and servers compared to Tableau, while Tableau Software has access to numerous data sources and servers such as Excel, Text File, PDF, JSON and others but it does not connect natively to XML files.
- Both are not free, the difference is only in the prices

For the last two reasons we decided to create our own visualization tool, using HTML, CSS frameworks, and java script libraries.

1.4 CSS & Bootstrap

Css (Cascading Style Sheets) is a language for specifying how documents are presented to users[19], it is a cornerstone technology of the World Wide Web, alongside HTML and JavaScript. CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts. This separation can improve content accessibility; provide more flexibility and control in the specification of presentation characteristics[26]. It has various frameworks comprising several CSS stylesheets ready for use for standard web design functions.

Bootstrap : Twitter introduced the framework in 2011, Bootstrap[] is an open-source framework containing CSS and JavaScript-based templates for interface components. It is known for popularizing the focus on responsive design among web developers. It promoted the now-ubiquitous concept of mobile-first and provided the right tools for its easy implementation. Bootstrap did so by introducing a grid – partitioning the screen into columns (invisible to the end user's eye).

FIGURE 5.5 – Bootstrap Logo.

1.5 Javascript & Highcharts

JavaScript[11] is a dynamic programming language that's used for web development, in web applications, for game development, and lots more. It allows users to implement dynamic features on web pages that cannot be done with only HTML and CSS. It offers various libraries that are pre-written JavaScript code, containing multiple functions, methods, or objects to perform practical tasks on a webpage or JS-based application.

Highcharts : Highcharts is a software library for charting written in pure JavaScript meant to enhance web applications by adding interactive charting capability. It has all the tools needed to create reliable and secure data visualizations by providing a wide variety of charts. For example, line charts, spline charts, area charts, bar charts, pie charts and so on. They offer wrappers for the most popular programming languages (.Net, PHP, Python, R, Java) as well as iOS and Android, and frameworks like Angular, Vue, and React[9].

FIGURE 5.6 – JS & Highcharts Logo.

2 Result

To validate the proposed solution, we visualize the data using various technologies highlighted in previous sections, in this section we will go through the results that were obtained by this investigatory effort.

2.1 Processed Data Result

In the data preparation stage, the data is output from the integration stage using Talend in this format :

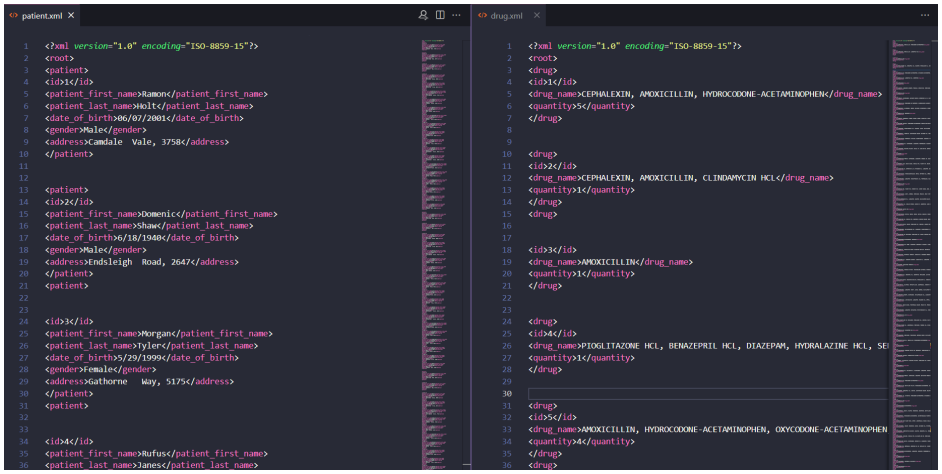


FIGURE 5.7 – Patient Informations and the related drugs.

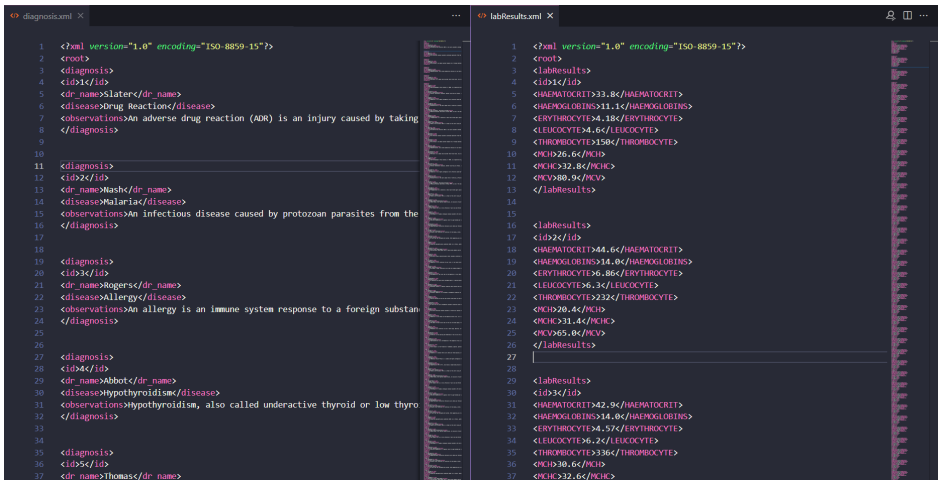


FIGURE 5.8 – Patient's diagnosis and the his Lab tests results.

2.2 Dashboards

[44] Defines a dashboard as :

> "A visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance".

HealthViz dashboard application (illustrated in the following pictures), is a web application developed using bootstrap[10] framework and the Highcharts library in order to display the right insight for each user.
This application represent the image data phase we already mentioned.

- **Administration Dashboard :**

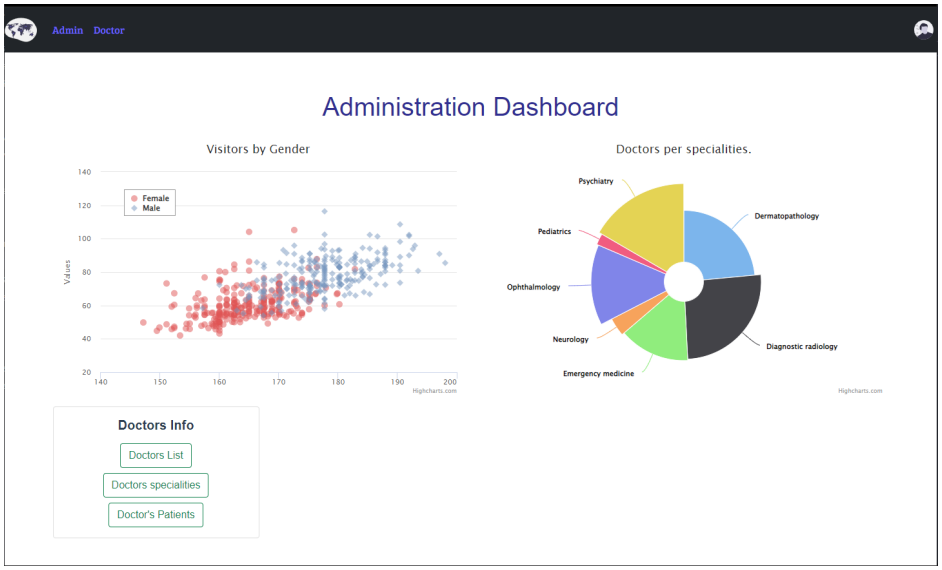


FIGURE 5.9 – Administration Dashboard illustrates some of Viz results.

- **Doctor Dashboard :**

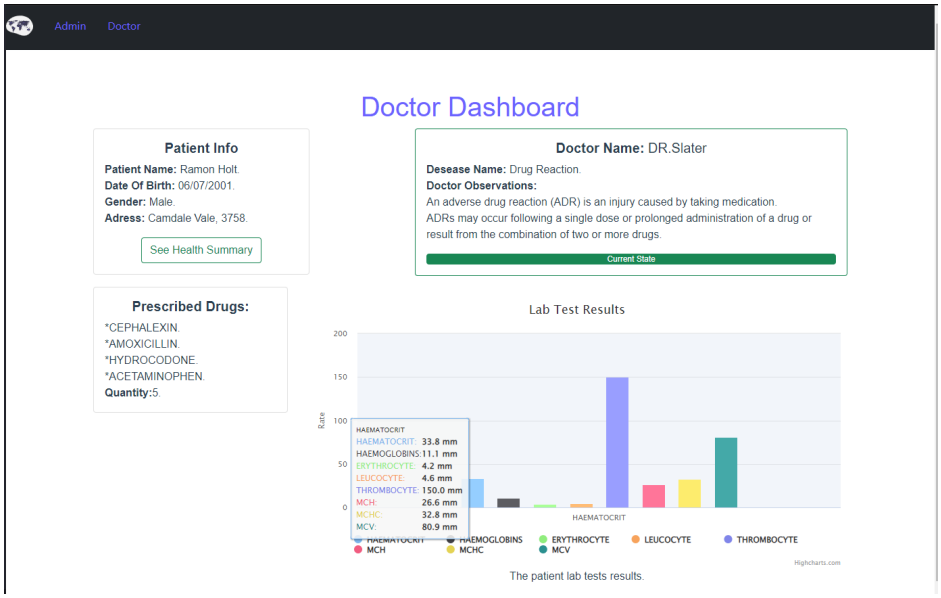


FIGURE 5.10 – Doctor Dashboard.

Chapitre 6

Conclusion & Future work

During this graduation project, we touched on many areas in the world of medicine and informatics. Many technologies and tools have been explored over time, starting with data integration systems and analytics to visualization. Each technology represents a universe to be discovered and many others still remain to be explored in the field of data integration and analyses.

In this thesis, we introduced medical informatics presenting all the data types and actors contributing to it. We also presented the visualization field, the visualization pipeline and we have seen the data warehouse definition, and the various data integration approaches.
In our work, we presented a visualization system that deals with multiple data sources and produces a structured data form presented in an XML form that could be easily visualized to the multiple medical actors through personalized dashboards. However, it will be necessary to ensure to have the right tools and good practices. In our case, we were limited to open source tools and hardware, which severely limited us.

This work remains a first version made using the relatively limited means at our disposal. However, it remains a good starting point that would be interesting to develop. Several points could improve this system, we mention :

- Automate the rendering process.
- Introduce continuous data integration.
- Mix the processing step by machine learning systems.
- Get the system protected using some security systems.
- Offer different outputs depending on what the actor needs.

Bibliographie

[1] ATIH : Agence technique de l'information sur l'hospitalisation. https ://atih.sante.fr/.

[2] Data Analysis - Process. https ://www.tutorialspoint.com/excel_data_analysis/data_analysis_pr

[3] Data Analysis, Visualization and Interpretation | MEALD Pro Starter. https ://mealdprostarter.org/n-data-analysis-visualization-and-interpretation/.

[4] Data Visualisation | Microsoft Power BI. https ://powerbi.microsoft.com/en-au/.

[5] Electronic Health Records | CMS. https ://www.cms.gov/Medicare/E-Health/EHealthRecords.

[6] Electronic Medical Record Systems | Digital Healthcare Research. https ://digital.ahrq.gov/electronic-medical-record-systems.

[7] Health Data Management : Benefits, Challenges and Storage.

[8] Healthcare data volume globally 2020 forecast. https ://www.statista.com/statistics/1037970/global-healthcare-data-volume/.

[9] Interactive javascript charts library.

[10] Introduction · Bootstrap v5.0. https ://getbootstrap.com/docs/5.0/getting-started/introduction/.

[11] JavaScript.com. https ://www.javascript.com/.

[12] Kaggle : Your Home for Data Science. https ://www.kaggle.com/.

[13] The Key to Maintaining Medical Records | Smartsheet. https ://www.smartsheet.com/medical-records-management.

[14] Pandas - Python Data Analysis Library. https ://pandas.pydata.org/.

[15] Paper-based versus computer-based records in the emergency department : Staff preferences, expectations, and concerns - Haleh Ayatollahi, Peter A. Bath, Steve Goodacre, 2009. https ://journals.sagepub.com/doi/10.1177/1460458209337433.

[16] Python (langage) — Wikipédia. https ://fr.wikipedia.org/wiki/Python_(langage).

[17] Talend Open Studio for Data Integration. https ://www.next-decision.fr/editeurs-bi/etl/talend-open-studio.

[18] Welcome to Python.org. https ://www.python.org/.

[19] What is CSS? - Learn web development | MDN. https ://developer.mozilla.org/en-US/docs/Learn/CSS/First_steps/What_is_CSS.

[20] What is Data Filtering? - Displayr. https ://www.displayr.com/what-is-data-filtering/.

[21] What is Information Visualization? https ://www.tibco.com/reference-center/what-is-information-visualization.

[22] What is Tableau : The Ultimate Guide To Know All About Tableau in 2021. https ://www.simplilearn.com/tutorials/tableau-tutorial/what-is-tableau.

[23] What Is the Data Analysis Process? 5 Key Steps to Follow. https ://www.g2.com/articles/data-analysis-process.

[24] Wheel | What are Electronic Medical Records. https ://www.wheel.com/companies-blog/what-are-electronic-medical-records.

[25] Pandas. *Wikipédia*, June 2020.

[26] CSS. *Wikipedia*, May 2022.

[27] Data analysis. *Wikipedia*, May 2022.

[28] Kaggle. *Wikipedia*, May 2022.

[29] Noura Mahmoud Abd Elazeem, Nevine Makram Labib, and Aliaa Kamal Abdella. A proposed data warehouse framework to enhance decisions of distribution system in pharmaceutical sector. *Egyptian Computer Science Journal*, 43(2) :43–60, 2019.

[30] Frederic Achard, Guy Vaysseix, and Emmanuel Barillot. Xml, bioinformatics and data integration. *Bioinformatics*, 17(2) :115–125, 2001.

[31] Yasser K Alotaibi and Frank Federico. The impact of health information technology on patient safety. *Saudi medical journal*, 38(12) :1173, 2017.

[32] Majedah Mohammad Alrehiely. Evaluating Different Visualization Designs for Multivariate Personal Health Data. page 461.

[33] Yaser A Alsahafi and Valerie Gay. An overview of electronic personal health records. *Health Policy and Technology*, 7(4) :427–432, December 2018.

[34] Javier Andreu-Perez, Carmen CY Poon, Robert D Merrifield, Stephen TC Wong, and Guang-Zhong Yang. Big data for health. *IEEE journal of biomedical and health informatics*, 19(4) :1193–1208, 2015.

[35] Sarah NAIT BAHLOUL. *Les entrepôts de données pour le décisionnel : Concepts et notions de base*, pages 33–36. USTO University, 2019.

[36] Petros Belsis, Apostolos Malatras, Stefanos Gritzalis, Christos Skourlas, and Ioannis Chalaris. *Pervasive Secure Electronic Healthcare Records Management*. January 2005.

[37] SAS bOracle France. Detection of adverse drug events : proposal of a data model. *Detection and Prevention of Adverse Drug Events : Information Technologies and Human Factors*, 148 :63, 2009.

[38] D Brailer. The decade of health information technology. *HHS Report, July*, 21, 2004.

[39] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare : management, analysis and future prospects. *Journal of Big Data*, 6(1) :1–25, 2019.

[40] Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, and Cees De Laat. Addressing big data challenges for scientific data infrastructure. In *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, pages 614–617. IEEE, 2012.

[41] George Demiris, Lawrence B. Afrin, Stuart Speedie, Karen L. Courtney, Manu Sondhi, Vivian Vimarlund, Christian Lovis, William Goossen, and Cecil Lynch. Patient-centered Applications : Use of Information Technology to Promote Disease Management and Wellness. A White Paper by the AMIA Knowledge in Motion Working Group. *Journal of the American Medical Informatics Association*, 15(1) :8–13, January 2008.

[42] Emmanuel Chazard. Réutilisation et fouille de données massives de santé produites en routine au cours du soin. page 174, July 2017.

[43] Reinhold Haux. Medical informatics : past, present, future. *International journal of medical informatics*, 79(9) :599–610, 2010.

[44] Jiří Hynek and Tomas Hruska. *Automatic Evaluation of Information Dashboard Usability*. April 2015.

[45] William H. Inmon. *Building the Data Warehouse*. Wiley, Indianapolis, Ind, 4th ed edition, 2005.

[46] Ralph Kimball and Joe Caserta. *The Data WarehouseETL Toolkit : Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. John Wiley & Sons, Inc., Hoboken, 2011.

[47] Shixia Liu, Weiwei Cui, Yingcai Wu, and Mengchen Liu. A survey on information visualization : Recent advances and challenges. *The Visual Computer*, 30(12) :1373–1393, December 2014.

[48] Par Niels Martignene. visualisation unifiée de données cliniques temporelles, hétérogènes, hiérarchiques. page 66.

[49] Izet Masic. The history and new trends of medical informatics. *Donald School J Ultrasound Obstet Gynecol*, 7(3) :301–302, 2013.

[50] Benmohamed Aek El Mehdi. Intégration continue dans un projet decisionnel au sein de la cnl. pages 1–73, 2018.

[51] Kenneth Moreland. A Survey of Visualization Pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 19(3) :367–378, March 2013.

[52] Somayeh Nasiri, Farahnaz Sadoughi, Mohammad Hesam Tadayon, and Afsaneh Dehnad. Security requirements of internet of things-based healthcare system : a survey study. *Acta Informatica Medica*, 27(4) :253, 2019.

[53] Mpho Ngoepe, Lebohang Mokoena, and Patrick Ngulube. Security, privacy and ethics in electronic records management in the South African public sector. *ESARBICA Journal : Journal of the Eastern and Southern Africa Regional Branch of the International Council on Archives*, 29, March 2011.

[54] Drew Robb. Power BI vs. Tableau : 2022 Software Comparison. https ://www.eweek.com/big-data-and-analytics/power-bi-vs-tableau/, February 2022.

[55] Pekka Ruotsalainen. A cross-platform model for secure Electronic Health Record communication. *International journal of medical informatics*, 73 :291–5, April 2004.

[56] Abderrazak Sebaa, Fatima Chikh, Amina Nouicer, and AbdelKamel Tari. Medical big data warehouse : Architecture and system design, a case study : Improving healthcare resources distribution. *Journal of medical systems*, 42(4) :1–16, 2018.

[57] Osama El-Sayed Sheta and Ahmed Nour Eldeen. Building a health care data warehouse for cancer diseases. *International Journal of Database Management Systems*, 4(5) :39–46, October 2012.

[58] Robert Spence. *Information visualization*, volume 1. Springer, 2001.

[59] Paul C. Tang, Joan S. Ash, David W. Bates, J. Marc Overhage, and Daniel Z. Sands. Personal Health Records : Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption. *Journal of the American Medical Informatics Association*, 13(2) :121–126, March 2006.

[60] Alexandru Telea. *Data Visualization : Principles and Practice*. January 2008.

[61] Pham Thi Thu Thuy, Young-Koo Lee, and Sungyoung Lee. S-trans : Semantic transformation of xml healthcare data into owl ontology. *Knowledge-Based Systems*, 35 :349–356, 2012.

[62] Teh Wah and Ong Sim. Development of a data warehouse for Lymphoma cancer diagnosis and treatment decision support. 6, March 2009.

[63] James G. Williams and And Others. Visualization. *Annual Review of Information Science and Technology (ARIST)*, 30 :161–207, 1995.

[64] REX Wong and Elizabeth H Bradley. Developing patient registration and medical records management system in ethiopia. *International journal for quality in health care*, 21(4) :253–258, 2009.

[65] Jian Xu, Laiwen Wei, Wei Wu, Andi Wang, Yu Zhang, and Fucai Zhou. Privacy-preserving data integrity verification by using lightweight streaming authenticated data structures for healthcare cyber–physical system. *Future Generation Computer Systems*, 108 :1287–1296, 2020.

[66] Sonja Zillner, Tilman Becker, Ricard Munné, Kazim Hussain, Sebnem Rusitschka, Helen Lippell, Edward Curry, and Adegboyega Ojo. *Big Data-Driven Innovation in Industrial Sectors*, pages 169–178. Springer International Publishing, Cham, 2016.