Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 196 (2022) 692–698

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

Data warehouse for machine learning: application to breast cancer diagnosis

Marwa Ben Ammar[a,]*, Faten Labbene Ayachi[a], Riadh Ksantini[b], Halima Mahjoubi[c]

[a]Higher Institute of Medical Technologies of Tunis (ISTMT) Tunis El Manar University, Tunisia
[b]Digital Security Research Unit - Higher School of Communication of Tunis (SUPCOM), University of Carthage, Tunisia
[c]College of IT, University of Bahrain, Bahrain

Abstract

The diagnosis and treatment of breast cancer are very much studied in the field of medical research. Indeed, the early detection of this pathology allows for the best possible treatment and minimization of the after-effects.

The business process "Diagnosis of breast cancer in women" is perfectly mastered by the practitioner and the medical profession and is widely documented in the literature. We propose a vision of “analysis of the existing” which allows to discover this process to the computer engineer in charge of the implementation of a system for aid in diagnosis and prognosis of the disease. This step is indispensable to discover the data generated during the business process and to discover in a later step all the characteristics that govern the disease. We define a system of diagnosis and prognosis of the disease Diaprog which uses the data of the classified and archived care records. We say that we use the historical data generated during the main process. We distinguish for this purpose between a production database and a data warehouse. The latter is a database fed from the various production databases accessible and made available in our system. In the case of our work, we call this data warehouse DDA (Diagnosis Data Archive). The data warehouse thus defined is a component of the medical diagnosis aid system Diaprog whose functional architecture also implements a learning module that allows the discovery of relationships between the characteristics that govern the disease.

Keywords: Breast cancer diagnosis, Patient career, data warehouse, Structured Data.

* Corresponding author.
E-mail address: marwa.ammar@istmt.utm.tn

10.1016/j.procs.2021.12.065

1. Introduction

Female breast cancer has now surpassed lung cancer as the leading cause of global cancer incidence in 2020, with an estimated 2.3 million new cases, representing 11.7% of all cancer cases , followed by lung cancer (11.4%), colorectal cancer (10%), prostate (7.3%), and stomach (5.6%). It is the fifth leading cause of cancer mortality in the world, with 685 000 deaths [1].

In Tunisia, breast cancer ranks first and represents 34.5% of all cancers incidents in women with 3092 new cases registered in 2020 [2]. The male sex represents less than 1% of the registered cases [2]. According to the World Health Organization (WHO), breast cancer is the second leading cause of death (8.3%) with a high 5-year prevalence around 163.63 per 100 thousand women [2]. The average clinical diameter of breast cancer discovery is about 5 cm [3]; with local, regional and distant extension in 30%, 40% and 15% of cases [3]. This complicates the healing process and increases the risk of mortality. Breast cancer develops in 70% of cases in women over 50 years of age and in 11% of cases in women under 35 years of age [3]. However, this cancer remains relatively frequent in women over 40 years of age in the world [2].

This pathology is still diagnosed late with a high frequency, locally advanced and aggressive stages. Long delays in diagnosis leads to a bad prognosis, often mutilating, responsible for morbidity and significant consequences. This constitutes a major concern both for the patient himself and for the health professionals and actors of the health system in general. Many studies have been devoted to the factors of variation of breast cancer. So, not all women with breast cancer automatically get the same cancer; therefore, they do not get the same treatment. The doctor offers each patient a treatment adapted to her situation. This gives the patient the best chance of recovery [4].

The actors of the health system in general have recalled the need to update the approach of early diagnosis both to improve the prognosis, increase the chances of recovery and to respond to the needs of patients for a better quality of life. Such is the context of our study which is inscribed in the field of the investigation of the medical pathology breast cancer. In fact, our investigation is conducted on women who have an incidence of breast cancer and who are continuing their diagnostic journey at the Salah Azaiez Institute (ISA). The Salah Azaiez University Hospital is a Tunisian public health institution that constitutes a reference center in the country for the monitoring, diagnosis and treatment of various types of cancer.

Our objective is to provide ICT solutions to satisfy the need for early detection of cancer and to guide the prognosis of the doctors towards a better medical management of women with breast cancer.

We structure this article as follows. Section 2 frames our approach in the definition of the data warehouse and presents the support architecture for a medical diagnosis aid system. Section 3 develops the conceptual schema of the data warehouse. The conclusion defends the pertinence of the data warehouse.

2. Methodology and Proposed Framework

In our approach, we ignore the actual deployment of any existing health information system. The data warehouse [5] stores the original archive of data collected during the diagnostic process. For this purpose, we distinguish between a production database and a data warehouse. The latter is a database fed from the various production databases accessible and made available to our system. The following figure illustrates the Main Process (PP) defined at the beginning and the different Sub-processes (SP).

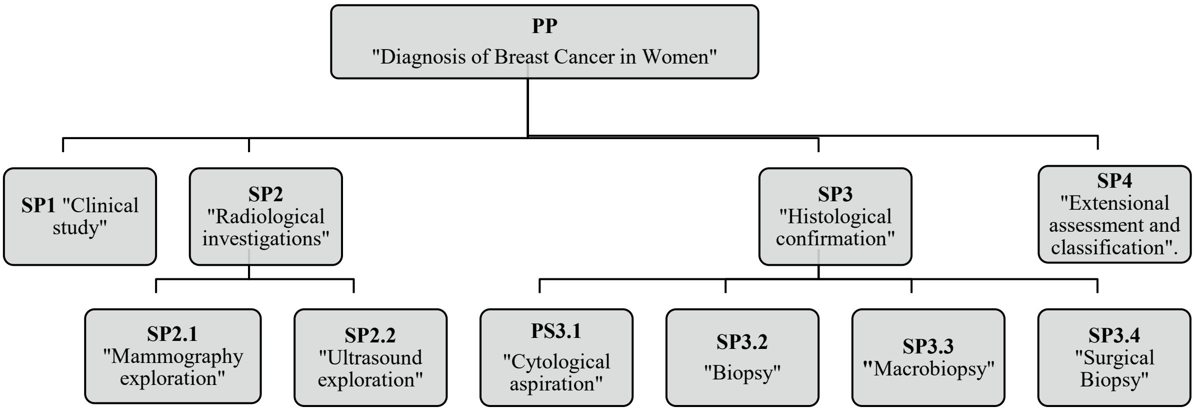
PP
"Diagnosis of Breast Cancer in Women"
SP1 "Clinical study"
SP2 "Radiological investigations"
SP3 "Histological confirmation"
SP4 "Extensional assessment and classification".
SP2.1 "Mammography exploration"
SP2.2 "Ultrasound exploration"
PS3.1 "Cytological aspiration"
SP3.2 "Biopsy"
SP3.3 "Macrobiopsy"
SP3.4 "Surgical Biopsy"

Fig.1. "Diagnosis of breast cancer in women" Process.

We have named our warehouse Diagnosis Data Archive (DDA). The data warehouse is centralized and is a valuable resource to support the diagnostic process. It ensures the availability and sharing of useful information that aids decision making in patient management and continuity of care. The history of care over long periods of time makes it possible to verify certain hypotheses and to refine them on a larger number of patient files.

We chose to proceed in four main phases for the design of our data warehouse. (1) understand and deconstruct the disease diagnosis process (2) identify for each decision step, the material support of the decision, the decision process and the risk of errors (3) identify the exhaustive list of elementary data constituting the medical record and assign them to finite value domains (4) model the data and identify the entities and the associations between the entities. This paper covers steps 3 and 4.

The data warehouse thus defined is the main component of a Diaprog medical diagnosis aid system whose functional architecture implements three main components.

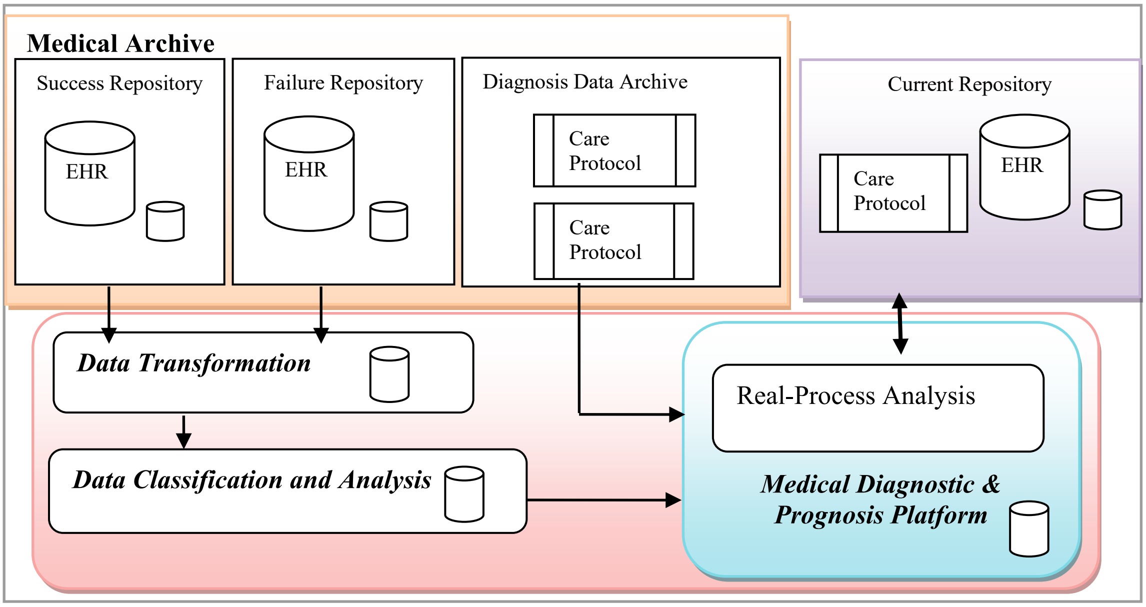

Fig.2.Diaprog Architecture.

- **Data transformation:** This is a phase of transformation of the digitized data into a format pre-exploited by the two other components of classification and prognosis. The digitized data are of different types (structured data, medical images, free text, etc.). Some archived data are in paper form.

- **Classification and data analysis:** This is an intelligent component of learning and correlation between medical data. Several algorithms are used, taking into account the nature of the input data: XML files, images, etc.

- **Medical prognosis platform:** A real process is an instantiation of the conventional process describing the treatment path followed by the patient. The medical prognosis platform implements the algebra of real processes and exploits the process archive to perform the following operations: reading of archived processes, detection of intersection and difference between real processes. This platform also allows the management of patient processes under elaboration, such as process initialization, process modification.

3. Data understanding and data modeling

Each phase of the breast cancer diagnostic process allows to characterize the detected lesions according to an appropriate approach and technique. The archiving of all the observations makes it possible to trace the evolution of the lesions over time and during the different phases of the same diagnostic process. To facilitate the investigation of the data and the realization of the objectives pursued, we have chosen to organize the data in advance in a structure adapted to a scientific project: a non-evolving database strongly oriented towards data processing. We present in the following the structure and content of this database.

Each exam is dated and refers to the administrative file of the patient and the organ studied. The doctor's conclusion is information that allows the different exams to be classified according to whether or not they have reached a certainty.

In some cases, an ultrasound exam may be solicited directly after the clinical exam and without going through a mammographic exam. We will materialize these two decision points in the future database.

The suspicion, during the clinical exam, of the existence of a lesion must be indicated by the mention of the estimated mass of the largest lesion as well as its estimated position on the breast map. Clinically, the breast is divided into four quadrants: superior-external, superior-internal, inferior-external and inferior-internal.

The mammographic exploration will confirm or deny the existence of lesions. The medical supporting document is also referenced and assigned the appropriate classification. If the mammogram confirms the existence of lesions, it is important to describe each lesion, specifying its type and position on the breast map. To each lesion, an order number will be assigned according to the exam in question (clinical, mammographic or ultrasound): Lésion_ECL, Lésion_EM, Lésion_EC. However, the memorization of the total number of lesions identified is important for improving the quality of the search and navigation in the data.

If a lesion is suspected during the mammographic exam, then an ultrasound exploration is recommended and should report all observed lesions. Ultrasound exploration allows to confirm or deny the existence of suspected lesions. The medical supporting document is also referenced.

The description of the lesions detected is similar to that adopted in the mammographic examination: indicate for each lesion its type, its diameter and its position on the breast map. To each lesion a relative order number will be attributed which takes into account the importance of the lesion.

A diagnosis of cancer can only be made after studying at a microscopic level the nature of the cells composing the lesion identified through imaging techniques [6]: this is the histological confirmation stage during which the physician uses another exploration technique and researches other types of information.

There are four types of sampling, including aspiration or cytological puncture, microbiopsy (commonly referred to as "biopsy"), macrobiopsy. The last two types are called percutaneous biopsies as opposed to surgical biopsy. The sampling techniques are different in terms of the type of needle used in the cytological exam, the size and location of the tumor in the breast, the amount of tissue removed and the nature of the suspected lesion, as well as the quality of the histological result.

Cytopuncture is performed by using a syringe and a finer needle than those used for blood tests. This technique is very simple, not very painful and quick and does not require local anesthesia or hospitalization.

Whatever the technique chosen, it allows in one way or another to confirm a diagnosis of cancer in order to affirm or not the cancerous nature of the lesion and its degree of local extension [7].

The conceptual diagram associated with the DDA warehouse is introduced in the figure below and highlights the three weak entities Lesion_ECL, Lesion_EM and Lesion_EC. In a real deployment solution, the DDA warehouse must be linked to the medical data production sites, in particular the information systems of the health care institutions. Finally, the information deduced from the interrogation must be synthesized and materialized in the database. This information is used to calculate the risk incurred whether or not the disease is diagnosed. This information relates either to the patient's career or to the career of her close family.

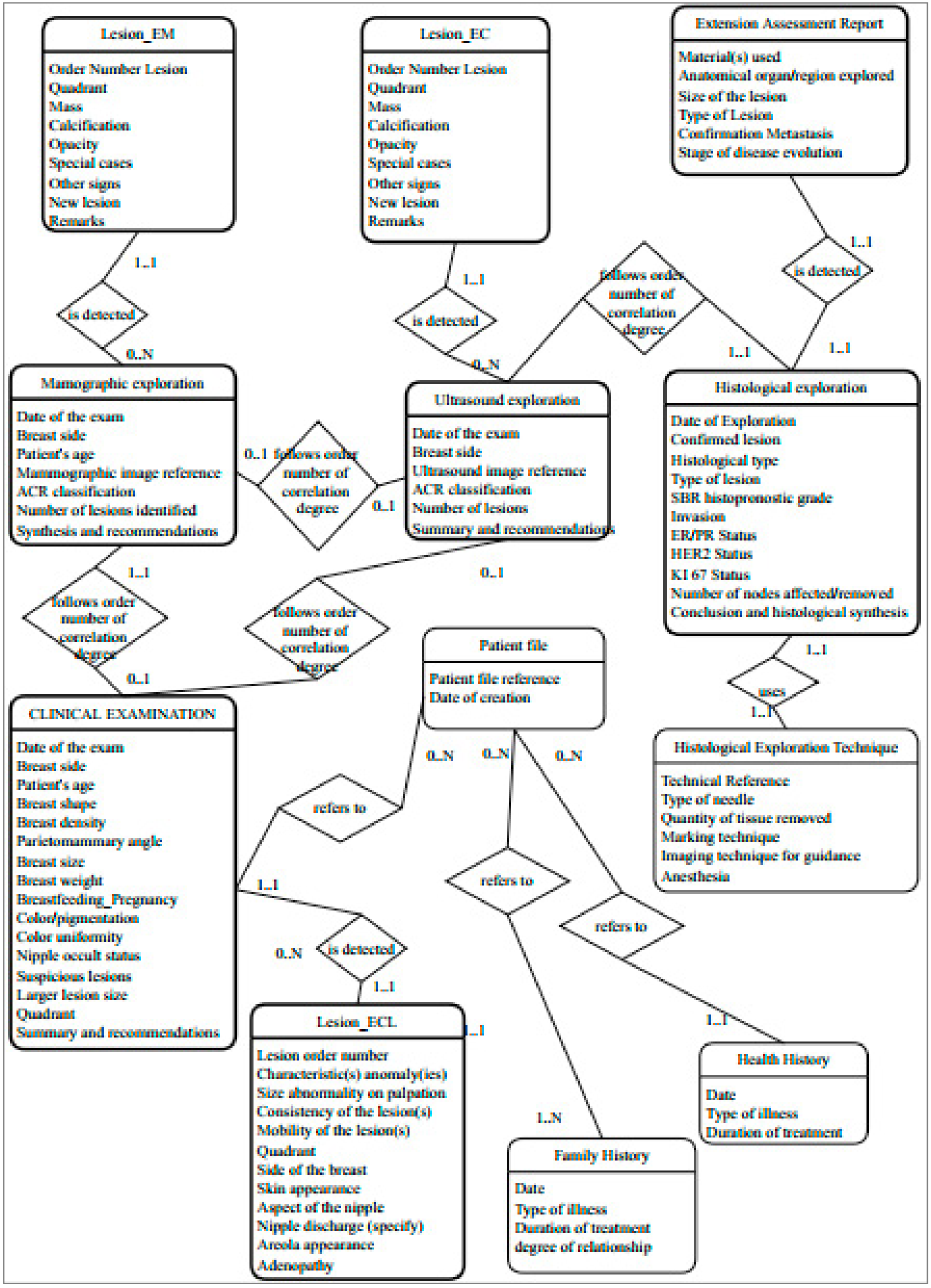

Fig.3. Data warehouse DDA (Diagnosis Data Archive) Conceptual Diagram.

4. Conclusion

The conceptual schema obtained captures without redundancy the data generated during the breast cancer diagnosis process. The representation in the relational model makes it possible to highlight the links between the data and to easily identify the requests that can be launched on this database. In the following, we mention the main points of the use of this database.

Our database allows us to conduct studies over long periods of time to explore the medical data and identify the benefits and harms [8] of organized mammography screening.

Interrogating the data warehouse allows us to perform general statistics:

- Disease incidence by country/region/age range;
- Number of new breast cancers/year.

Interrogating the data warehouse allows us to obtain knowledge on the disease:

- Risk factors for recurrence [9];
- Correlation between the chances of cure and the detection phase of the disease.

Interrogating the data warehouse allows us to evaluate the performance of the screening process:

- Efficiency of the screening process: discovery of the disease before symptoms set in: usually screening is performed by mammography;
- Rate of missed screening mammograms: this is the rate of breast cancers that occur between two mammograms. Usually because the cancer grows very quickly;
- Incidence of breast cancer: number of positive cases out of the total screenings carried out;
- Average delay between the initial treatment and the appearance of metastases. Metastases can occur in these organs years or decades after diagnosis and initial treatment;
- Risk of invasive breast cancer being diagnosed late. This is the problem with the timing and frequency of screening mammograms;
- Characteristics of breast cancer that will never progress;
- False-positive rate;
- Rate of over-diagnosis.

Thus, a good identification and structuring of the data allows to explain and understand the causes of cancer, to discover the cumulative, digressive or compensatory factors that can modify the conclusions.

Acknowledgements

The author thanks SALAH AZAIEZ Hospital staff for providing the research grant to support this research work. The corresponding author wishes to thanks Doctor Olfa Jaïdane and Data analyst Nejla Hamda for the great cooperation and motivation for this research.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." CA CANCER J CLIN 2021;71:209–249. doi: 10.3322/caac.21660. Available online at ca cancer journal.com.

[2] Tunisia - Global Cancer Observatory, March, 2021. Available online at https://gco.iarc.fr/today/data/factsheets/populations/788-tunisia-fact-sheets.pdf.

[3] M Ben Abdallah, S Zehani, M Maalej, M Hsairi, M Hechiche, K Ben Romdhane, H Boussen, A Saadi, N Achour, F Ben Ayed. "Breast cancer in Tunisia: epidemiologic characteristics and trends in incidence." Tunis Med. 2009 Jul;87(7):417-25.

[4] Daniela Rodriguez-Rincon, Brandi Leach, Camilla D'Angelo, Amelia Harshfield and Catriona Manville. (2019).Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK. "Factors affecting access to treatment of early breast cancer: Case studies from Brazil, Canada, Italy, Spain and UK Implications for future research, policy and practice."

[5] Kristina K Gagalova, M Angelica Leon Elizalde,Elodie Portales-Casamar, and Matthias Görges. (2020)."What You Need to Know Before Implementing a Clinical Research Data Warehouse: Comparative Review of Integrated Data Repositories in Health Care Institutions."PMC7484778 v(4)8.doi: 10.2196/17687.

[6] K. Venkata Ratna Prabha,D. Vaishali, Pallikonda Sarah SuhasiniK,SubhashiniP. Ramesh.(2021)"Different Diagnostic Aids and the Improved Scope of Establishing Early Breast Cancer Diagnosis." Springer.book series (LNNS, volume 179).

[7] Laurentius O. Osapoetra, Lakshmanan Sannachi, Daniel DiCenzo,Karina Quiaoit,Kashuf Fatima, and Gregory J. Czarnotaa.(2020)"Breast lesion characterization using Quantitative Ultrasound (QUS) and derivative texture methods." doi: 10.1016/j.tranon.2020.100827.v.13(10).

[8] Screening programmes: a short guide. Increase effectiveness, maximize benefits and minimize harm (2020). World Health Organization. ISBN 978 92 890 5478 2.

[9] Polton, Dominique. (2018)"Les données de santé." Med Sci (Paris), Vol. 34, N° 5; p. 449-455; DOI: 10.1051/medsci/20183405018.