Abstract—Data warehouses have been developed to integrate and restore data from heterogeneous sources to provide better analysis for decision making. The objective of this paper is to present an overview of various works dedicated to medical data warehouses and a comparison of these approaches. Our comparative study is based on a number of criteria analysis dimensions, and the fact schema of the data warehouse.

Keywords— medical data warehouse, medical data, simple data, complex data.

I. INTRODUCTION

Any organization can be described by its decision-making system dedicated by means of an approach of decision-making. The decision-making system assures the extraction, the transformation, the storage, the analysis and the restoration of the decision-making data of an organization.

The data managed by this type of system, are scattered and can result several heterogeneous sources and do not allow having a synthesized outline of the activity of the organization. Consequently, it is important to make them homogeneous and integrated to facilitate any treatment to help the decision-makers in their decision-making.

The role of the data warehouse is recognized as the heart of a decision-making information system: he integrates the data stemming from various domains to make them accessible to the process of decision-making analysis.

The data warehouse is defined as «a collection of data, integrated, not volatile and historiated for decision making». [8]

Decision systems revolve around data warehouses and allow online analysis called OLAP (On Line Analytical Processing).

The systems of data warehouses based on paradigm multidimensional make use of the multidimensional representation of the data as well as the operators of on-line analysis OLAP to lead a successful and efficient analysis on intelligent data.

Since the 90s, the data warehouses became strategic components of the decision-making computing. They were successfully set up in diverse sectors as the transport, the telecommunication, the distribution, the business, financial services, the insurances and the medicine.

For example, the medicine aims at preventing diseases by using medical proofs based on the clinical experience and the preferences of the patients to be able to help a doctor to diagnose better a disease and to propose her an adequate treatment thanks to a relevant identification of the disease on the basis of the list of symptoms [12].

In this article, we concentrate on the data warehouses dedicated to the medicine allowing to find a suited treatment to identify a given disease based on the tests. We present in this work an outline of the diverse approaches dedicated to the medical data warehouses. A comparative study underlining the strengths and the inconveniences of every work will be performed.

The rest of this article is organized as follows: in the section 2, we present the various research works bound to the medical data warehouses. In the section 3, we propose a comparative study between these works. Finally, in section 4, a conclusion and some perspectives will be presented.

II. EASE STATE OF THE ART

The medicine based on the tests is necessary when:

(i) A diagnosis must be made: when a clinician overwhelmed by the variety of symptoms should consult the rules based on evidence to determine the correct diagnosis.

(ii) A patient needs to get treatment guidelines based on tests generally offer few treatments, but only those that do not conflict with the preferences of the patient and the health risks are taken into consideration.

(iii) A prognosis is required: when the external knowledge about the state of health of the patient is ensured.

(iv) An etiology must be clarified when the drug based tests provides expert knowledge on the possible causes of a given disease. For example, how and which proportion of nicotine increases the risk of a heart attack [13] [14].
The medical data can be considered as simple or complex data. On the basis of this criterion, we classify the medical projects of the data warehouses according to two trends: medical warehouse based on simple data, medical warehouse based on complex data.

A. Medical Data Warehouse based on Simple Data

The implementation of a data warehouse is a medical complex task given the objective it seeks. Three projects were launched in this framework: (i) federation project [1], (ii) clinical data warehouse [6], [20], [21], (iii) project ADELEM (Help Logistics and Medical Decision) [15], [16].

1) Federation Project: warehouse federation for health care insurance

The data warehouse federation for health insurance organizations practicing evidence-based medicine is obtained by merging multiple data warehouses of different insurance organizations in Austria. [1]

It is created from heterogeneous data warehouse and integrated into one unity of conceptual standpoint. The data model is based on three actions: meetings PatientEncounter patients, therapies SubstanceAdministration, procedure and requirements.

Banek et al. [1] describe only the facts in the data warehouse schema designed star. The fact ordinance is analyzed in eight standard dimensions. The authors stressed structures and dimensions Patient InsuranceOwner that are identical because they both provide a different view of personal data (if the treatment of a patient is carried by its own policy, the content of these two dimensions are the same).

The data warehouse can be used to find correlations between symptoms and diagnostic results provide the best therapy appropriate given the full status of a patient or to compare the efficacy of different drugs against the same disease. On the other hand, economic various drugs have been taken into account and promoting products with lower costs, since insurance organizations pay diseases therapeutic procedures and medications. The project aims to find potential costs (for example the drug therapies used for a long time with little or no effect) and to promote therapies that provide the best possible return for equal costs.

Data warehouses related to patient billing and care in the management of clinical settings are as follows:

- The systems of invoicing of the patients and the management of the financial information supply mainly information for the financial optimization, the management of the costs and the coordination of the persons receiving benefits of health insurance. They usually contain demographic information about the patient, duration of hospital stay, charges, costs and detailed transactions for each billable service as a surgical procedure performed, medications given, etc.

- The systems of care of patients manage generally information on the individual patients, including the location of the patient at the hospital, the prescribed medicines, etc. They allow the optimization of the administrative efforts, in a hospital.

3) Project ADELEM (Help Logistics and Medical Decision)

The project ADELEM admits the development of software required to support logistical and medical decision, within the SROS (Schemes Health and Social Organization) managed by the ARH (Regional Hospitalization Agencies).

The schema of the data warehouse is in constellation with three fact tables and dimensions.

The fact table Prise_MCO (Medicine-Surgery-Obstetrics) has been designed for the management of hospital stays for making short stays. (ii) The fact table Population demographic processes, (iii) the third table Prise_SSR (After Care and Rehabilitation) manipulate the data for long stays.

We can decompose this warehouse in snowflake schemas as follows:

(i) Fact Table Prise_MCO allows decision aid in the analysis of stays from measurements as AccountDuration_stay, SumDuration_stay, which can be analyzed according to the following dimensions: Establishment, CIM10, Time, Mode_output, Age, Area_geo, and Weight_birth.

(ii) Table of facts Population treats the demographic data in the process of analysis to help the decision-maker in the decision-making from the axes of analysis such as: the table Zones Geo and the table RP99 as the dimensions and Accountfem_proc as measure.

(iii) Table of facts Prise_SSR allows the decision-making support during the manipulation of the long stays from the measures such as AccountDaysinwk and AccountsDayshorswk, who can be analyzed according to the following dimensions: CIM10, Mode, Time, Age, Area_geo, Week the Beginning, Week the End and the establishment.

B. Medical Warehouse based on Complex Data

1) Clinical Data Warehouse

The clinical data warehouse [6], [20], [21] is implemented to reduce medical errors, increase the efficiency and quality of medical care and reduce treatment costs.

The data warehouse is deployed to generate evidence-based medicine.

---

1 The formal concept lattice is a mathematical structure to represent disjoint classes not underlying a set of objects.
According to Darmont et al [4] [5], medical data may contain complex data based on the following characteristics:

- Multi format: The data are represented in different formats (databases, texts, images, sounds, videos ...);
- Multi structure: The data is structured (relational databases, XML documents repository ...);
- Multiple sources: The data are originated from different sources (distributed databases, web ...);
- Multimodal: The data are described through several canals or points of view (radiographies and in the audio diagnosis of a doctor, the data expressed in various scales or languages...);
- Multi-version: the data are changeable in terms of definition or the value (temporal databases, periodic inquiries).

For example, we quote in the following paragraphs What were launched six contributions in this framework: (i) the contribution of Pederson [9] [11], (ii) the contribution of Teste [12] [18] [19], (iii) project MAP (Anticipation Personalized Medicine) [2] [3], (iv) the contribution of Hussain [7], (v) the contribution of Midouni [4], (vi) contribution of Khalid et al [10].

1) The Contribution Of Pederson

The authors [9] [11] proposed to integrate XML and relational data in multidimensional databases to perform OLAP analysis.

Nine requirements of the multidimensional data model are fixed; each is illustrated with a real world and a clinical case study.

A survey of existing models revealed that the requirements do not support many-to-many relationships between facts and dimensions for the treatment of changes over time, and the contribution of the uncertainty, as well as various levels of granularity in the data. A multidimensional data model range, which caters to all nine requirements, is introduced. The first requirement is designed for a non-strict hierarchy is to consider the partial order of dimension values.

The second requirement takes into account the many-to-many relationships between facts and dimensions and handling different levels of granularity are supported in the third requirement.

In addition, the fourth time requirement is managed by adding valid time and transaction time to the base model. The fifth requirement is based on the uncertainty that is treated by adding probabilities to the base model.

The advanced features of current models are also supported in the sixth requirement.

The seventh requirement designed for explicit hierarchies and multiple dimensions are supported by the lattice structure and how to treat all data as dimensions, while enabling the calculation of symmetric treatment of dimensions and measures. The aggregation mechanism presented in the eighth requirement allows the user to ensure that data is properly aggregated.

Requirement addresses the ninth on algebra model objects showing that it is closed and at least as strong as the relational algebra with aggregation functions.

2) The Contribution Of Teste

Teste et al, described the decision in the medical field by using the extension of the data model for the analysis and control of health care spending. This model considers only the data in the patient record and ignores the complexity of medical data.

Teste defined manipulation language and representation models [12], [18], [19], which are dedicated to data warehouses and stores data based on complex and evolving data based on object-oriented models. It has built in his model the temporal dimension to account for the evolutionary aspect of the data in a meaningful way.

3) Map Project

The objective of MAP (Medical Anticipation Custom) [2], [3] is to consider that the data warehouse should be used as a structure using the diagnosis layer in the field of high performance sport. He also used as the data warehouse storage structure from several static views, so they can be used, if necessary, to bring the proof.

The data warehouse store only affects biological data, biometric and cardiovascular disease. The data warehouse is medical MAP organized as a collection of data stores. Each magazine contains specific data relating to the medical specialty. These data stores are defined by a set of facts shared with other repositories.

A Meta model medical was used to build a set of these "concepts" in a certain area). This model is consistent with the MAP Meta model. Two other Meta data store models for cardiovascular disease are proposed.

4) The Contribution of Hussain

Hussain [7] tried to propose a new model of data warehouse cardiovascular analysis. He used three fact tables:

(i) Table «Exam»: This table stores data on examinations that contains information on tests that prescribed by the doctor and also the conclusion of these tests. It is also linked to other fact tables as Document_Cardio, Exam_Results and other dimension tables as table Individual, Doc_Card, Type_Exam, Machine....
(ii) Table «Document_Cardio»: This table is used to store references to documents examinations medicine. It also stores measurements on the size and type of documents. It is also linked to the examination table in a relationship M: N.

(iii) The third table «Exam_Results» presents the fact table. It stores the results of the examinations prescribed. It is also linked to the dimension table to classify the results.

5) **The Contribution of Midouni**

The project of Midouni [4] proposed a multidimensional model of complex medical data, specifically for the cardiovascular module data. Its objective consists in generalizing the current model in a meta-model for the storage of medical data. The role of this warehouse is to store and integrate information necessary for doctors MAP project and keep the history to support analysis required for decision making.

6) **The Contribution of Khalid and al**

The approach Khalid presents the methodology of data warehousing to manage data DCM (Dementia Care Mapping) which allow a large number of users (mapper, director and researcher) the collection, storage, sharing and analysis of relevant and consistent information from data warehouses that serve to improve the quality of dementia care nationally, locally and internationally.

DCM is a tool for monitoring complex to assess the quality of care for people with dementia over a long period often than 6 consecutive hours. The observations are made by a trained observer recorded the behavior and mood levels and commitment based on two types of codes: category code of behavior (BCC) and code of mood and engagement (ME).

### B. Comparative Studies

In this section, we propose to compare the research work related to medical data warehouses. Our comparative study is based on the technical aspects such as data type (simple or complex), the data warehouse schema what were designed to make the data ready for analysis (Star schema, Snowflake schema, Constellation schema) the number of dimensions, the number of facts and modeling data warehouses.

The functional aspect concerns the field and topic analysis. The comparison is summarized in the following table (see **TABLE I**)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Data type</th>
<th>Schema of data warehouse</th>
<th>Modeling</th>
<th>Number of dimensions</th>
<th>Number of fact</th>
<th>Topic</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrib. of PEDERSON [9], [11]</td>
<td>Complex</td>
<td>Snowflake schema</td>
<td>Data warehouse</td>
<td>Several</td>
<td>One</td>
<td>Clinical case</td>
<td>Medical</td>
</tr>
<tr>
<td>Contrib. of TESTE [12], [18], [19]</td>
<td>Complex</td>
<td>Star schema</td>
<td>Data warehouse</td>
<td>Several</td>
<td>One</td>
<td>Control health spending</td>
<td>Medical</td>
</tr>
<tr>
<td>MAP project [1],[3]</td>
<td>Complex</td>
<td>Constellation schema</td>
<td>Data Marts</td>
<td>Several</td>
<td>Several</td>
<td>Diagnosis player in the field of sports</td>
<td>Sport</td>
</tr>
<tr>
<td>Contrib. of HUSSAIN [7]</td>
<td>Complex</td>
<td>Constellation schema</td>
<td>Data warehouse</td>
<td>Several</td>
<td>Several</td>
<td>Cardio-vascular</td>
<td>Medical</td>
</tr>
<tr>
<td>Contrib. of MIDOUNI [4]</td>
<td>Complex</td>
<td>Constellation schema</td>
<td>Data warehouse</td>
<td>Several</td>
<td>Several</td>
<td>Cardio-vascular</td>
<td>Medical</td>
</tr>
<tr>
<td>Contrib. of KHALID et al [10]</td>
<td>Complex</td>
<td>Star schema</td>
<td>Data warehouse</td>
<td>Several</td>
<td>One</td>
<td>Quality of care</td>
<td>Medical</td>
</tr>
<tr>
<td>Federation Project [1]</td>
<td>Simple</td>
<td>Star schema</td>
<td>Data warehouse</td>
<td>Several</td>
<td>One</td>
<td>Health insurance</td>
<td>Medical</td>
</tr>
<tr>
<td>Clinical data warehouse [6],[20], [21]</td>
<td>Simple</td>
<td>Constellation schema</td>
<td>Data warehouse</td>
<td>Several</td>
<td>Several</td>
<td>Efficiency and quality of medical care</td>
<td>Medical</td>
</tr>
<tr>
<td>Project ADELEM [15], [16]</td>
<td>Simple</td>
<td>Constellation schema</td>
<td>Data warehouse</td>
<td>Several</td>
<td>Several</td>
<td>Management of hospital stays</td>
<td>Medical</td>
</tr>
</tbody>
</table>

The works [1], [10], [12], [18] and [19] were interested in the star schema data warehouse, describing each made eight dimensions [1] to make the warehouse ready to analysis and decision making.

These works were related to medical field to monitor the health costs and quality of care, based on analysis axes.

The works [9], [11] were interested in the pattern snowflake data warehouse based on an explicit hierarchy in dimensions. The many-to-many are supported by facts-dimensional relationships. The modeling is based on the analysis of data stores and data warehouses.
The works [2], [3], [4], [7], [6], [20], [21], [15] and [16] were interested in the constellation schema what was designed to facilitate decision-making in the medical field. The model has several dimensional models that share the same dimensions. These works are based on complex data to be modeled by data stores and data warehouses to facilitate data acquisition.

III. CONCLUSION

In this article, we presented the various research works which concerned the medical data warehouses. This works were classified according to two trends: medical warehouse based on simple data and medical warehouse based on complex data. Our comparative study is based on two technical and functional aspects. In the future, we suggest studying the modeling problem of data warehouses in the medical domain.

REFERENCES