19

Results of Data Mining Technique Applied to a Home Enteral Nutrition Database

Maria Eliana M. Shieferdecker, Carlos Henrique Kuretzki, José Simão de Paula Pinto, Antônio Carlos Ligoki Campos and Osvaldo Malafaia
Universidade Federal do Paraná,Curitiba, Brasil

1. Introduction

This chapter presents the results of data mining (ID3 and A PRIORI) techniques applied to a Health (nutrition) database that was originated from a knowledge management point of view.
As participants of a Graduate Program in Surgery, we developed a knowledge management strategy and operationalized it by an information system, called SINPE, which is able to manipulate a large database (actually with 500,000+). Data items are organized into data collecting protocols that store data in a relational database. The system offers tools to retrieve and analyze data, with some basic (descriptive) statistics and graphic charts generated automatically. An interface allows to apply data mining algorithms into collected items. A main feature of the system is that it was developed from the usability point of view, because the primary users are physicians with low domain of information systems. Thus the system is very easy to manage.
The aim of this work was to apply these data mining techniques to a home enteral nutrition database in order to identify features not previously suspected.
The data mining technique was applied to a large health database protocol, with 1592 specific (nutrition) collected items. After the selection of interesting items the ID3 and APRIORI algorithms were applied to 111 patients, 58 females and 53 males, between 19 and 92 years old. These data were analyzed and presented in graphics and tables. Two questionnaires were answered by the users to validate the tool and its results. All operations were performed by physicians with low knowledge of data mining techniques, who were assisted by a BS in Computer Science professional.
After mining the database, obtained results were compared with the international literature and the overall results met our expectations. Among other results, the data mining technique applied to the home enteral nutrition database identified an unexpected high incidence of malnutrition among patients that were receiving home enteral nutrition. Readmission after treatment was also higher than expected, reaching a 50% rate. Physicians who used the system approved it. No discrepancy was observed while using the system, but there are some parameters that must be better explained.
The application of data mining techniques to a large medical (nutrition) database allowed us to identify nutrition features not previously known, which helped to improve public nutrition policies in this specific area. This user friendly system was proved useful when

applied to a large nutrition database. It is necessary to improve the samples, such as confidence and support, in a way to better explain these results.

2. Data collecting protocols

To make a data protocol for a research is similar to build a large questionnaire. It is a simple but methodical task. The questions to be used are defined by a specialist, using his tacit knowledge and some medical references. Each time new research is needed a new protocol is generated. In our approach, the principal items of a health area, like gastroenterology, are first defined and stored in a database. A Master Protocol is organized. Then, when a medical research in gastroenterology is started, say Zenker's diverticulum, the necessary data collecting items are obtained from the master protocol, by choice in a knowledge tree, generating a specific protocol. Master items are never deleted, but may be inactivated. Collected data are linked to a specific protocol that will be analysed in terms of item frequency, patient's genre, age and ascendance, presence or absent of certain symptoms, and so on.

The original idea of this approach was developed over the last 15 years in a Brazilian school hospital, and for the last 10 years applied in clinical researches by graduate (master and doctorate) students while composing their theses. Over the years were organized 500,000 items, in 40 master and 234 specific data collect protocols, arranged in the following areas (Table 1):

Master protocol	# items
Diseases of the Esophagus	2656
Urological Device	1030
Anorectal Diseases	3926
Liver Transplantation	4892
Bowel	2967
Bile Ducts, Extrahepatic	1948
Diseases of the Pancreas	5059
Diseases of the Stomach	2172
Small intestine	9349

Table 1. A brief view of SINPE's database protocols

The basic problems involved with this task are to manage protocols, how to improve data quality and how to build a user friendly system for healthcare professionals that store and help analyse data.

2.1 The SINPE system

SINPE is a Brazilian Portuguese contraction for Sistema Integrado de Protocolos Eletrônicos (Integrated Electronic Protocols System). Developed by medical informatic graduate students nowadays intends by sixth version, presenting a MS-Windows Desktop, a handheld Windows Compact Edition-based (Fischer et al., 2003) and a web interface for data collect. The task of protocol building have only a desktop interface, because our experiences with handheld to large user interaction necessary to make the knowledge tree using this interface was not good; the same occurred with browsers, that forcing to develop an applet or other way for better interaction. Actually there is not a Linux version, because our physicians do not use this system yet.

In Medicine, a common approach to make a research is to build a data collect (clinical) protocol to conduce the observations. A protocol is like a structured questionnaire, consisting in information about diseases, treatments, drugs, patient data, therapies and clinical history (Sackett et al., 2000).

To make a protocol is not a simple task, since there is an extensive number of references about the theme that must be included. Common reference sources are international journals, books, congress proceedings and older studies / researches (Warren & Warren, 1993).

After obtaining the material, the method used is to select the most important and organize then in a logical sequence. This procedures result in a protocol (see Figure 1).

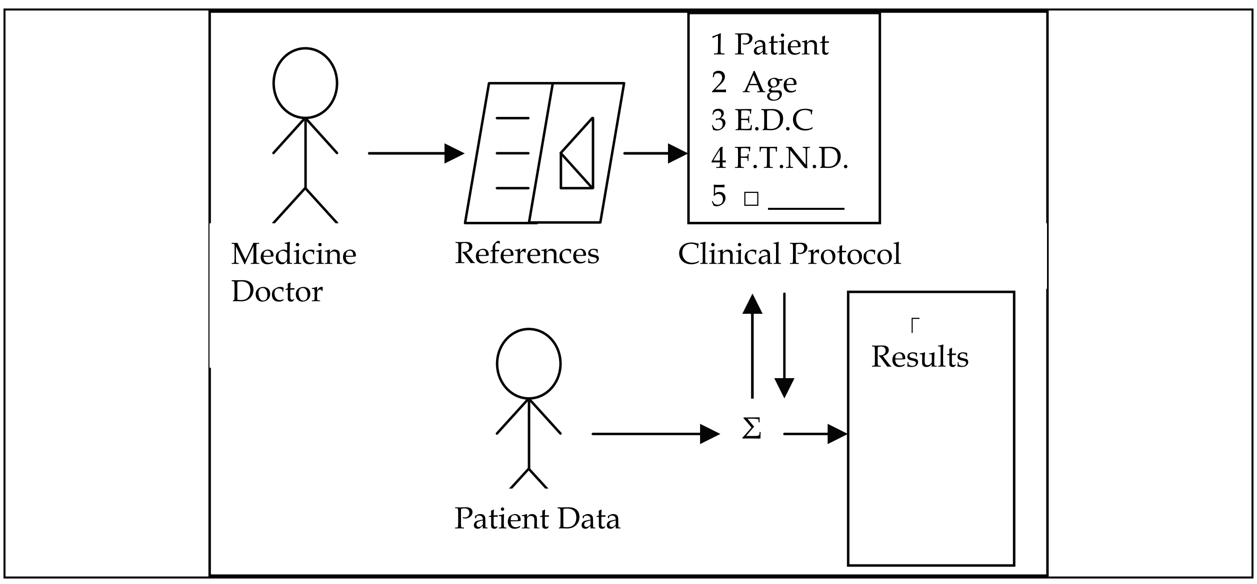

Fig. 1. The process to make a clinical protocol

A first problem related to this way to conduce researches is to make new protocols, according to the above description, because each new detail needs to be tested, and, since too many researches have overlapped metadata, like patient sex or allergy, the proposed approach results in a significant amount of work to make new protocols (see Figure 2).

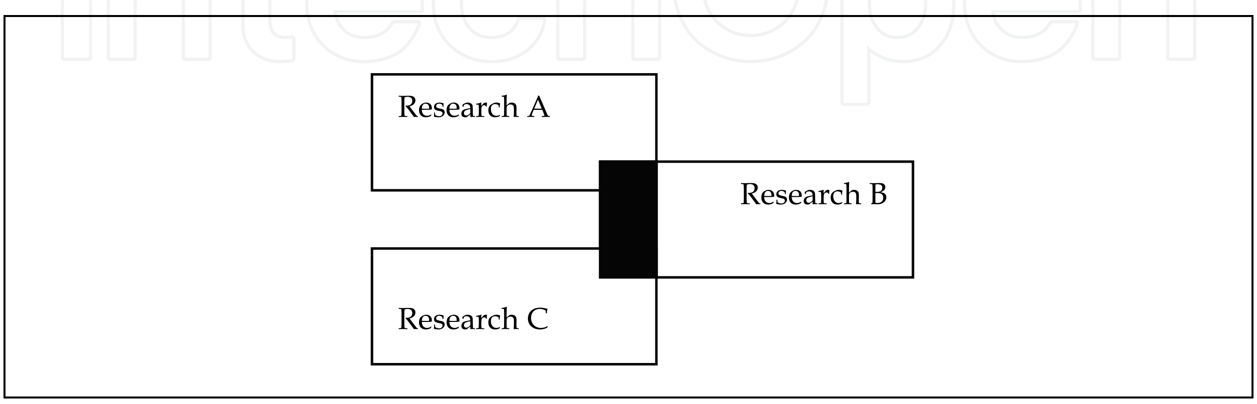

Fig. 2. Overlap of metadata in different researches

In SINPE's, our approach is to categorize the common metadata to an area, gastroenterology by example, in a master protocol, and then build specific protocols from the master. The process to relate data collected in a protocol to a unique master description, makes it easy to proceed with the same or correlated researches in different sites, which improves epidemiologic studies, and the integration process with the pre-existent hospital information systems (Afrin et al., 1997).
Earlier, when the system was idealized, the first prototype was built using object oriented paradigm (Russel, 2000) and SUN's JAVA to make an MS-Windows® stand alone application, with a centralized database and a web applet, but the final result was a problem to manage: it is difficult to physicians to control the JAVA environment variables, like "PATH", "VM Versions", and conflict with other applications (like Oracle® 8.x clients) installed in the computer. The centralized database was not over clinical research department responsibility, and there is not a DBA to manage backup, recovery and other DBMS procedures. Also, the use of JAVA technology forced us to improve the memory capacity of the microcomputers in use. This was built in 99-2000's. The actual system was developed using Microsoft dotNET framework, with C# coding and Access and SQL Server databases, since MS-Windows environment is widely used by Brazilian healthcare professionals. A reduced class diagram is showed in figure 3, in which it is possible to view the classes that define a master and a specific protocol.
A security model was developed to manage the system, producing 4 user types: an administrator, a data collector, a protocol elaborator and a viewer. Differences between them are concerned to a power to create master protocols, new users and sites, new master and specific protocols, only collect data or only view protocols and data results.
This approach was developed as a part of doctorate studies in medical informatics that resulted in an information system called SINPE. The SINPE has been used by master degree students, that build master protocols over a PhD supervision and by doctorate students that are responsible for multicentre studies.

2.2 The SINPE Process

The difficulty related to put a new system in practice is directly proportional to the operational/ behavioral changes that will be made from this application. As a way to put SINPE to work we developed a three phase process that manages activities, as showed in Figure 4.
The phase I is not different from the traditional way to build protocols, since it consists in identifying references about a theme. But here the references and the metadata (patient temperature, temperature unit and temperature limits, for example) are stored in a database and will be accessible by many researches at time. A concept may be related to metadata (like fewer) and may exist as a link to an external dictionary (like ICD-10) or metathesaurus (like UMLS) (Lindberg et al., 1993; Rocha et al., 1994). This phase is very important to insure data quality.
Once finished phase I, in phase II a specific research will be build selecting concepts from master's database. It is possible to use the same concepts (consisting in metadata, description, archives like images or sounds and references) in various protocols. Once a concept is used, it may be actualized but not removed from the database. There is a possibility to invalidate an old used concept, but never to remove it (Sacket & Straus, 1998). With this, we guarantee the knowledge management/ reuse of questionnaires.

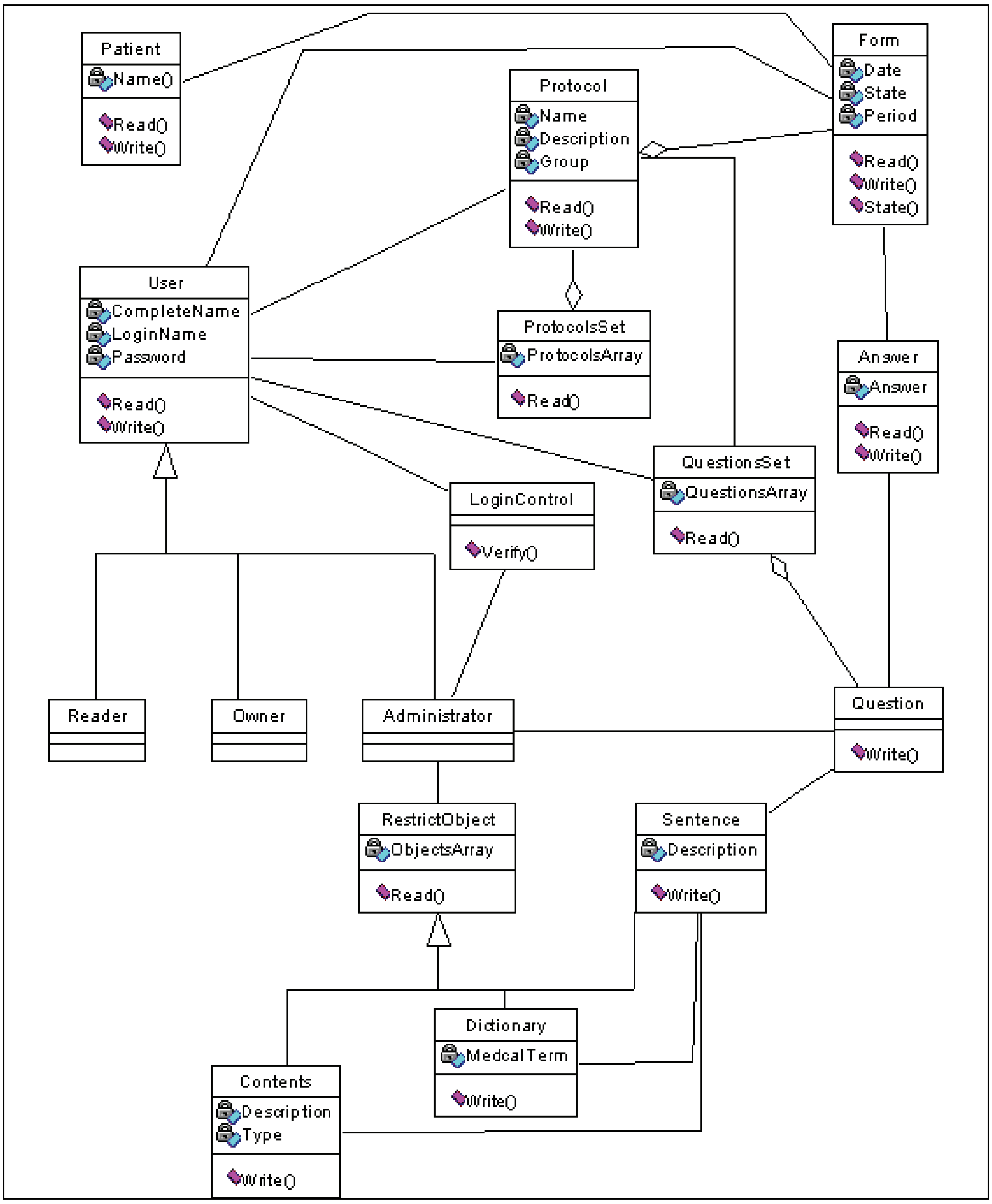

Fig. 3. SINPE's reduced class diagram

When a specific protocol is built it may be distributed for the researches or his or her co-workers that will collect the data. In this phase (phase III), the data may be collected by using a trivial MS-Windows® interface, Web-based interface or by a handheld computer simplified interface (Graham et al., 2002). The database may reside in a local copy or a remote (Internet accessible), decentralized or centralized SGBD.

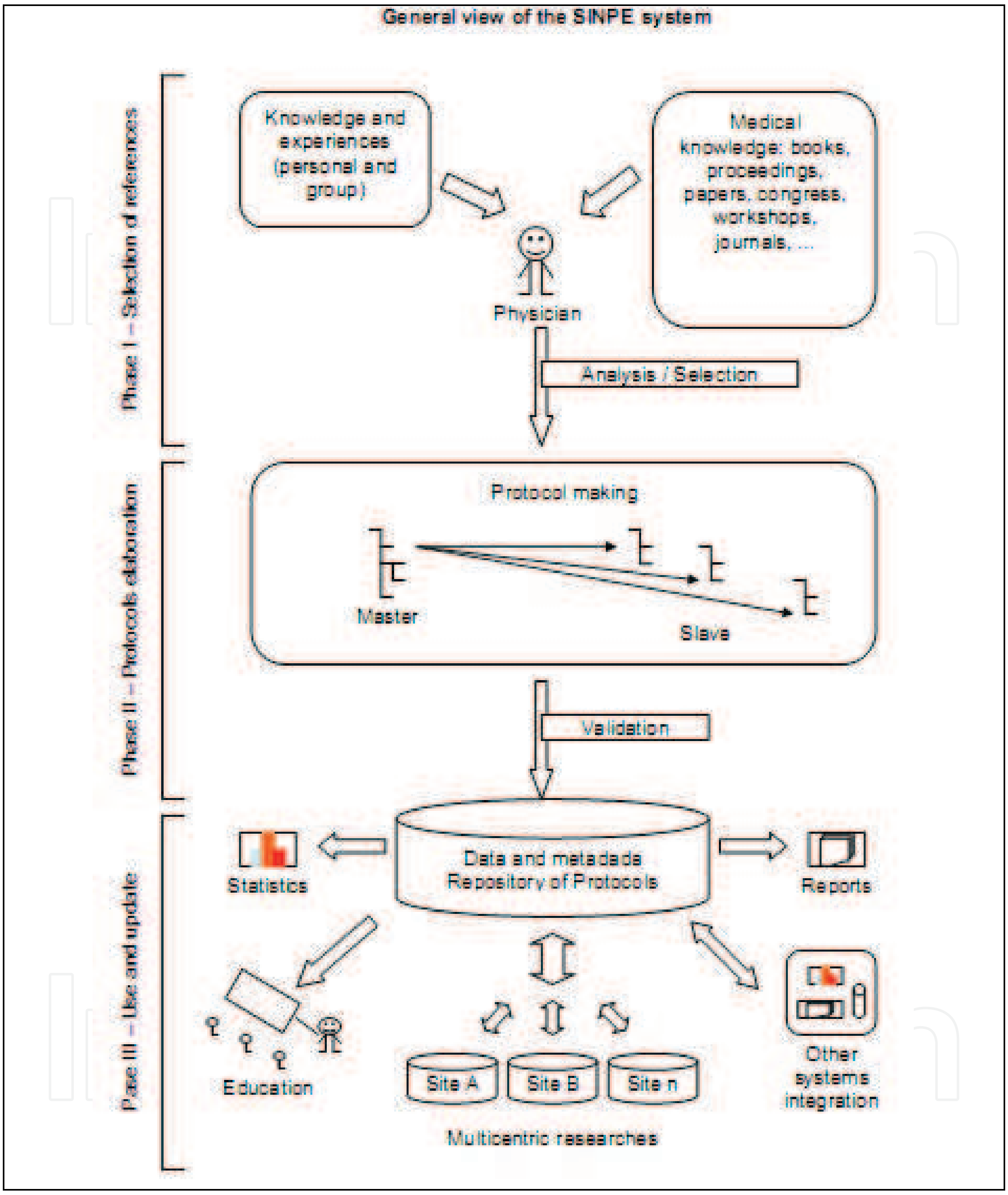

Fig. 3. SINPE's reduced class diagram

2.3 SINPE's analysis tool and use in learning/ education

Once protocols are elaborated and data are collected, SINPE's offers an analysis tool that generates descriptive statistics and some graphs automatically when a specific protocol is selected, and permits to elaborate reports, export files and print results. For each item in a protocol the analyzer will calculate its proportion related to total data collected and some

epidemiological data, like patient's gender or race. To help results evaluation we implemented a data mining tool, that permits to evaluate the collected data using a wizard interface: physicians may not know algorithms or calculus (Ingelfinger, 1993), only the hypothesis to be validated (or not).
Another use of the this system is in training areas, allowing students to access a high quality database to make "ad hoc" questions about diseases and their treatments, negative and positive responses on drug dosages, epidemiologic studies and so on. To enforce this use we are now idealizing a system to make clinical cases based on collected data by the clinical protocols system (showed in Figure 4).

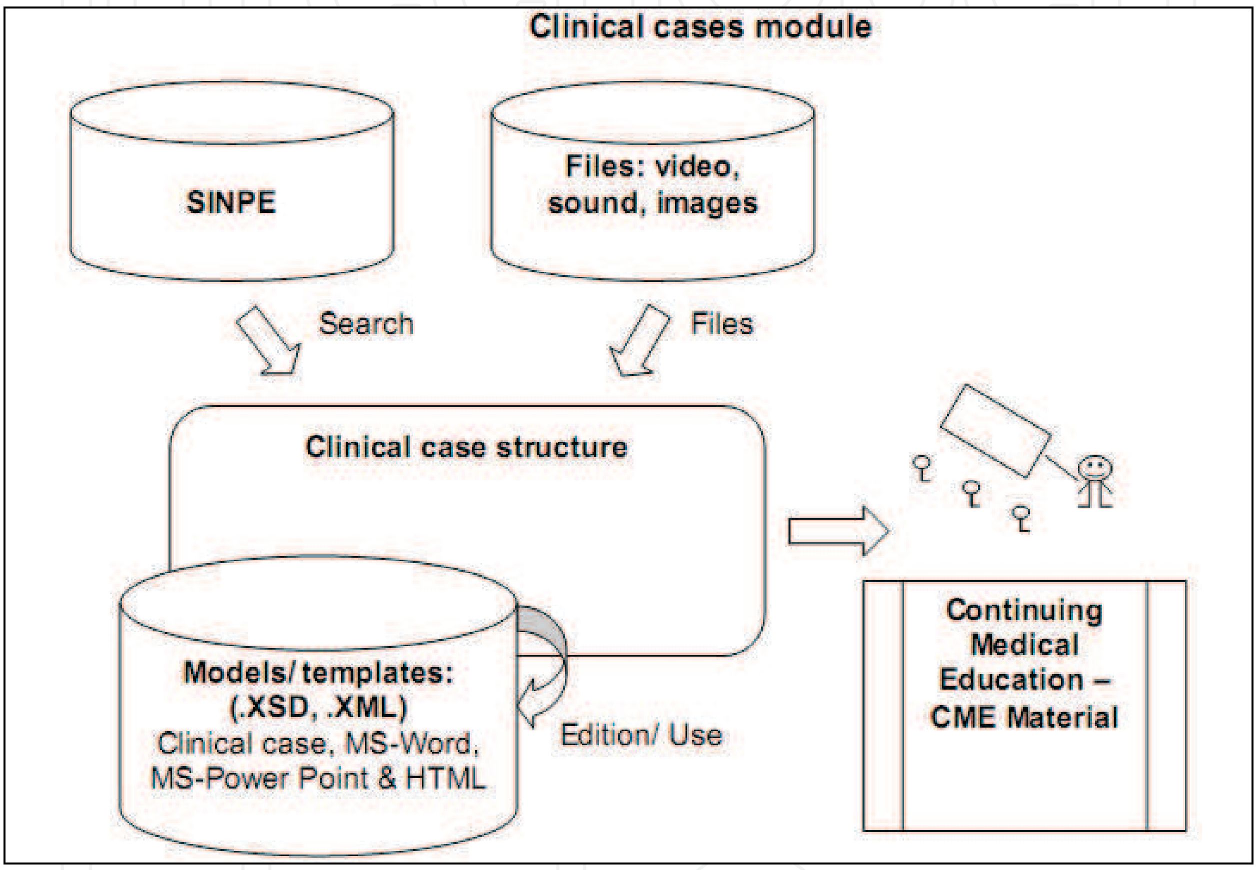

Fig. 4. SINPE's reduced class diagram

An advantage of the use of SINPE's approach is that there are high quality metadata linked with high quality collected data that will allow the use of data mining techniques with adequate precision and considerable importance in their results.

3. Home enteral nutrition

The use of computational resources, especially regarding the capture, storage and retrieval of clinical data, has been important to the production of relevant and reliable clinical studies (Haux et al., 2002; Doebbeling et al., 2006). These databases allow the collection of structured clinical information to the analysis and production of prospective studies in large series of patients. It contributes to the quality of healthcare, development and solidification of technical and scientific knowledge. The integration of technological advances in information and on health sciences allows the production of relevant and reliable clinical studies.

The home care is growing, it is applied to 460 patients per million of the United States of America's population, and in 40 patients per million habitants in Spain (Moreno Villares, 2004; Planas, et al., 2006). Having the records of patients in nutritional therapy at home is important, because this information enables the coordination of resources and improvement of health care (Planas et al., 2003).

The study "Elaboration and Validation of an Electronic Protocol for Homecare Enteral Nutritional Therapy in Patients Attended by the Municipal Health Unities of Curitiba" applies an electronic protocol for homecare enteral nutrition patients (TNED) through a research performed in 111 home visits of Curitiba´s health units users. The development of the research evolved the following stages: elaboration of an electronic protocol for TNED with 1793 sub items grouped in nine main items: identification, nutritional evaluation, diet prescription, enteral nutrition indications, enteral tube feeding access, enteral nutrition composition, administration systems, complications and re-hospitalization. This theoretical base was created as an electronic protocol called Informatized Master Protocol of TNED using the integrated System of Electronic Protocols (SINPE©). A specific electronic protocol was then created from this master protocol, which was applied for the evaluation of homecare users of the municipal health units. The demonstration of the results was produced using the visualization interface module of information. There were 58 females and 53 males, between 19 and 95 years of age, whose data were analyzed and presented using graphics. These images could be saved, copied to the computer transference area (memory), allowing their exportation to other softwares or insertion in analysis files. The Electronic Protocol creation and its application on TNED municipal health units patients was possible. A number of useful information emerged from this research, such as: most of caregivers are family members; most of the patients are malnourished; neurological diseases are those that predominates as indication for homecare enteral nutrition, and between those, stroke was the most prevalent, gastrostomy was the most used tube feeding method, the most frequent complications of the nutritional therapy were the gastrointestinal ones; more then half of the patients needed to be re-hospitalised after the beginning of home enteral nutritional therapy. The database analysis resulted in important information that contributed for research and creation of public nutrition politics.

3.1 Results on data mining in a health database in nutrition

A general view of applied data mining allowed the creation of algorithms. A critical view over results versus application versus reality was therefore possible. The Apriori method rule showed that a BMI ≤ 22 kg/m^2 implied the intake of a homemade diet. The patients with IMC ≤ 22 kg/m^2 are characterized as malnourished and were receiving a homemade enteral diet. This relation may indicate that quality of the infused diet is related with the nutritional state of the patient.

These results have a key relevance on clinical decisions and contribute to the improvement of public policies in healthcare. If the malnourished patients are those who receive the homemade diet it can indicate that this diet is not meeting the nutritional needs of the patients, and therefore the prescription must be modified.

According to Van Bemmel (VAN BEMMEL, 1997), techniques of Data Mining and Knowledge Discovery in Databases processes were originally developed for corporate sales and production data, but they are also relevant to health care settings.

To Witten and Frank (WITTEN, 2005) Data Mining is defined as the process of discovering patterns in data. These processes should be automated or semi-automatic. But the patterns discovered must be meaningful and that they lead to some benefit.

"Many times a Data Mining is a part of the discovery of knowledge as one of the most important fields on knowledge management. [...] Techniques such as Bayesian models, the decision tree, artificial neural networks, association rules and genetic algorithms are generally used in the discovery of patterns that are known or previously unknown to the system and to users. Data mining can be used: applications, marketing, customer relationship, engineering, medicine, analysis of crimes, prognosis specialist, Web mining and mobile computing, among others." (Chen, 2005).

Data mining is also used to extract rules from health care data. For example, it has been used to extract diagnostic rules from breast cancer data (Chen, 2005).

Malnutrition is associated with poor clinical outcomes such as delayed recovery from illness, longer length of hospital stay, increased occurrence of complications and reduced quality of life (RUSSELL, 2007). The data mining enables better practices by presenting the relation link between these items, allowing the identification of related factors with malnutrition.

Techniques such as data mining and text mining need to be used with great care in the biomedical applications, in view that medical data are often sensitive and involves private and confidential information. Errors and incorrect associations could be rapidly propagated through electronic media, especially when large databases and powerful computational techniques are involved (Chen, 2005).

The Apriori method of data mining has been widely used on supermarkets sales databases, but despite this application, the Apriori method can be easily applied on health areas in order to develop rules involving possible items. The application of specific databases on health care has also great commercial potential.

4. Conclusion and future works with the use of SINPE in nutrition related studies

When observing healthcare and its guidelines we have perceived the need for quality data and strong processes as a way to assure efficient guidelines. Some data mining tasks conduced over a database with high quality data are applicable to various areas of the medical sciences. This study is an adequate demonstration of that. With the SINPE we were able to obtain improvements in data organization and quality, and also usability enhancement when considering the use of statistic functions.

5. References

Afrin LB, Kuppuswamy V, Slater B, Stuart RK. Electronic Clinical Trial Protocol Distribution via the World-Wide Web: A Prototype for Reducing Costs and Errors, Improving Accrual, and Saving Trees. *JAMIA* 1997; 4(1):25-35.

Doebbeling, B. N.; Chou, A. F.; Tierney, W. M. *Priorities and strategies for the implementation of integrated informatics and communications technology to improve evidence-based practice*. Journal of General Internal Medicine, v. 21, n. 2, p. S50- S57, 2006.

Fischer S. (2003), Stewart TE, Mehta S, Wax R, Lapinsky SE. Handheld Computing in Medicine. *JAMIA* 2003; 10: 139-149.

Graham M, Chignell M, Harumi T. Design of Documentation for Handheld Ergonomics: Presenting Clinical Evidence at the Point of Care. *Proceedings of the 20th annual international conference on Computer documentation*, October, 2002.

Haux, R.; Ammenwerth E.; Herzog, W.; Knaup P. *Health care in the information society. A prognosis for the year 2013*. International Journal of Medical Informatics, v.66, p. 3 – 21, 2002.

Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. *Biostatistics in Clinical Medicine* 3rd ed. New York: McGraw-Hill; 1993.

Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods of Information in Medicine* 1993; 32(4):281-91.

Moreno Villares, J.M. *La práctica de la nutrición artificial domiciliaria en Europa*. Nutrición Hospitalaria, v.19, p. 59-67, 2004.

Planas, M.; Castellá, M.; García Luna, P.P. et al. *Nutrición enteral domiciliaria (NED): Registro Nacional del año 2000*. Nutrición Hospitalaria, v. 18(1), p. 34-38, 2003.

Planas, M.; Lecha, P. P.; García Luna, P.P. et al. *Registro Nacional de la Nutrición Enteral Domiciliaria (NED) del año 2003*. Nutrición Hospitalaria, v. 21 (1), p. 71-4, 2006.

Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a Controlled Medical Vocabulary Server: The VOSER Project. *Computers and Biomedical Research* 1994; 27(6):472-507.

Russel, LT. Clinical Trial Remote Data Capture Client/Server Solution. *Addendum to the 2000 proceedings of the conference on Object-oriented programming, systems, languages, and applications*, January, 2000.

Russell CA."Impact of malnutrition on healthcare costs and economic considerations.," Clinical Nutrition Supplements, vol. 2, pp. 25–32, 2007.

Sackett DL et al.. *Evidence-Based Medicine*. How to Practice and teach. EBM. 2nd Ed., Edinburgh: Churchill Livingstone, 2000.

Sackett DL, Straus SE. Finding and appling evidence during clinical rounds: the evidence cart. *JAMA*, 1998.

Van Bemmel JH, Mussen MA. Handbook of Medical Informatics: Springer, 1997.

Warren JR, Warren DE. A knowledge-Based Patient Data Acquisition System for Primary Care Medicine. *Proceedings of the second international conference on Information and knowledge management*, December, 1993 .

Witten IH, Frank E. Data Mining: practical machine learning tools and techniques, San Francisco: Elsevier, 2005.