APPLYING A KDD PROCESS IN AN ORAL PUBLIC HEALTH UNIT AT RIO GRANDE DO SUL, BRAZIL

Luciano Costa Blomberg[1], Rodrigo Coelho Barros[2], José Antônio Poli de Figueiredo[3] and Duncan Dubugras Alcoba Ruiz[1]

[1]*Graduate Program in Computer Science – Pontifical Catholic University of Rio Grande do Sul (PUCRS) Porto Alegre/RS - Brazil*

[2]*Graduate Program in Computer Science and Computational Mathematics - Institute of Mathematics and Computer Science - University of São Paulo (USP), São Carlos/SP - Brazil*

[3]*Graduate Program in Dentistry, Pontifical Catholic University of Rio Grande do Sul (PUCRS) Porto Alegre/RS - Brazil*

ABSTRACT

This study aims at conducting and documenting a complete Knowledge Discovery in Databases (KDD) process for inducing predictive models in order to investigate new hypotheses about the causes of dental diseases such as periodontal and dental caries. We analyze data concerning the dental records of 598 low-income patients, treated by the Brazilian Unified Health System (UHS) in a collaboration work with Pontifical Catholic University of Rio Grande do Sul (PUCRS). Predictive models induced suggest there is a genetic origin in the incidence of new cases of periodontal disease, as well as a possible influence of hypertension and high-cholesterol diseases in the incidence of new caries cases.

KEYWORDS

Databases; data mining; data warehousing; Knowledge Discovery in Databases; dental diseases.

1. INTRODUCTION

Despite the significant improvement of the caries index in the Brazilian child population for the last two decades, dental caries still constitute a public health problem directly related to the patient's life conditions. Tooth decay is the most common chronic childhood disease, five times more common than asthma and seven times more common than hay fever (Montenegro, 2008). According to the Brazilian epidemiological study of 2003 on oral health (Brazilian Ministry of Health, 2006), periodontal diseases was highly incident in all age groups, given that less than 22% of the adult population and less than 8% of elderly patients presented healthy gums.

Considering the incidence of malocclusion in the population, the prevalence data at age 5 revealed moderate or severe occlusal problems in 14.5% of the population, ranging from a minimum of 5.6% in the north region to a maximum of 19.4% in the south region. Furthermore, there was also an increase in mortality caused by mouth cancer in the years 1979 to 1998, from 1.32 to 1.82 per 100,000 citizens, a fact mainly observed in males (Brazilian Ministry of Health, 2002). Given the severity of the epidemiological scenario, it is extremely important that researchers and public administrators adopt new methods for investigating the factors that cause these diseases, in order to plan suitable policies aimed at improving public oral health.

This work applies and documents a complete KDD process in the oral health area, exploring its benefits in the generation of predictive models related to oral diseases. Related work usually perform a straightforward performance comparison of data mining techniques applied to oral health, neglecting important steps of the KDD process, such as the validation and interpretation of the discovered knowledge. Through the KDD process, we verified that the quality of the generated predictive model is directly related to the quality of the collected data, i.e., there is a strong influence of data quality on the performance of the predictive models.

Considering the importance of information and knowledge management as an instrument of diagnosis and planning, the goal of this study is to conduct and document a complete KDD process for inducing predictive

models that may help to identify new hypotheses on the causes of dental diseases such as periodontal and dental caries. We intend to encourage the application of a complete KDD process, providing to researchers and managers a comprehensive view of steps typically not discussed in the literature, but which are fundamental to the analysis of the obtained results.

2. KDD OVERVIEW

The rapid advance in data collection and storage technologies has allowed organizations to accumulate large amounts of data over the past decades. However, human beings cannot naturally analyze a large amount of data, resulting in the need of techniques that can automatically analyze data in an intelligent way (Tan, 2006). With that in mind, the KDD process has emerged as a great approach to identify valid, new, potentially useful and understandable patterns in data (Fayyad, 1996).

2.1 Steps of the KDD Process

According to (Fayyad, 1996), the KDD process comprises five steps, but we can summarize them in three major steps: pre-processing and data cleaning, data mining, and interpretation/evaluation data.

i) Pre-processing and data cleaning: consists of performing operations such as exclusion of outliers, handling missing and unknown data. Many authors consider it to be the most time-consuming step in the KDD process;

ii) Data mining: step in which the actual search for patterns of interest in a particular form of representation is performed.

iii) Interpretation/Evaluation: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users.

As earlier described, data mining is an important step in the KDD process, traditionally applied in the search for useful information in large data repositories. According to (Tan, 2006), we can divide it into two broad types of tasks: descriptive and predictive tasks. Descriptive tasks are derived patterns that summarize the underlying relationships in the data. Some well known descriptive strategies are association analysis, cluster analysis and anomalies detection. Predictive tasks predict the value to a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the dependent variable or target, while the attributes used to forecast are known as independent variables. According to (Tan, 2006), predictive tasks can be used to assign objects into pre-defined categories. This kind of algorithm uses the induction principle to learn classification models from a set of records with known class variable (training set), and apply them to the deduction of other records, where the variable class is unknown (test set). Among the several classification techniques, decision-tree induction is widely employed in several application domains. A decision tree is comprised of a hierarchy of nodes connected by branches. Each internal node (or non-terminal) denotes a logical test on an attribute (independent variable), in which branches represent the test outcome, and each leaf node stores a label for the class (the dependent variable).

3. RELATED WORK

For the past two decades, a number of studies have been performed in order to identify the dental profile of citizens. In a broader scenario, epidemiological surveys conducted by the Brazilian Ministry of Health in 1986, 1996 and 2003 have served as information source to new researches in the area, whose aim are to explore the association between dental diseases and oral health data (eg: anamnesis and dental charts). Among the most relevant researches in state of art, we have highlighted the application of well-known statistical and machine learning-based algorithms for the study of oral diseases, as for example, caries (Stewart, 1991), (Baldani, 1996), (Powell, 1998), (Zhang, 2006), (Montenegro, 2008), (Tamaki, 2009) (Ito,

2011) and oral cancer (Majumder, 2005), (Shuang, 2011), (Sharma, 2011). Additionally, the performance comparison of traditional machine learning methods also has been common in these studies, such as decision trees, neural networks, support vector machine and k-nearest neighbor (Oliveira, 2005), (Montenegro, 2008) and (Sharma, 2011).

However, few papers in oral health literature have documented the whole steps and benefits of the "mother process" named KDD. Regarding this issue, the closest studies have been done by (Gansky, 2003) and (Lin, 2009). In (Gansky 2003), data of 466 children up to twenty-four months of age were analyzed in order to predict the caries risk. Logistic regression, classification and regression trees, and neural networks were used and compared. Although fundamental KDD concepts have been introduced, this is still a work eminently oriented to performance's comparison of machine learning methods. (Lin, 2009), on the other hand, applied a complete KDD process in order to help the third-party payer to prevent fraudulent claims and overcharges against insurance programmes. For that, SOM neural network unsupervised clustering approach was used in conjunction with domain experts, although more details about the preprocessing were desirable.

In this paper, we seek to document a complete KDD process, investigating the use of predictive models for analysis of oral diseases. By surveying similar works that attempt to use dental diseases data for performing knowledge discovery, we concluded that:

a) Typical statistical approaches that exploit the predictive analysis of dental diseases are particularly limited, since they are focused in the validation of pre-defined hypotheses (instead of allowing the discovery of new hypotheses), and do not offer a simple, intuitive and easy-to-interpret predictive model;
b) Approaches related to this study are excessively oriented to the strict comparison of data mining techniques. They tend to ignore important steps in the process of knowledge discovery, especially the evaluation and interpretation of the models that are generated.

This study differs from others since it promotes a deeper discussion regarding the KDD process in oral health, thus providing to domain specialists a broader view of its important role in quality control of collected data, and also allowing the validation of new hypotheses concerning the incidence of oral pathologies.

4. APPLYING A KDD PROCESS IN AN ORAL PUBLIC HEALTH UNIT

The School of Dentistry at PUCRS has over 50 years of experience in teaching, developing research and supporting social in Porto Alegre, Brazil. Together with the Undergraduate Program, the Graduate Programs with areas of concentration in Surgery and Buccomaxillofacial Traumatology, Restorative Dentistry, Endodontics, Clinical Stomatology, Dental Materials, Orthodontics and Facial Orthopedics, and Dental Prosthesis have all combined efforts to developing Dentistry in Brazil. With this broad area of activity, a large quantity of clinical data, laboratory exams, image exams (radiography, tomography and echography) and auxiliary charts are created daily, making the manual analysis of the collected data an extremely exhausting process (Blomberg, 2009).

In order to improve and automate the information and knowledge management in oral health area, the School of Dentristy at PUCRS developed in partnership with the Graduate Program in Computer Science – PUCRS an environment for storage, organization and retrieval of data from dental records.

Due to the diversity of areas involved in the partnership and the short period of time to collect data, we selected a single area (Social Dental Care – Vila Fátima CEU) as the unit of study, given its social relevance. This choice is justified as the Vila Fátima CEU develops a social assistance service and dental care through the Brazilian UHS (Unified Health System) to a low-income population, with more than 8000 inhabitants located in the East of Porto Alegre, Brazil.

The solution developed in this study aims to completely document the laborious KDD process, providing to oral health managers a systematic approach to support decision-making and formulation of new hypotheses on the influence of socieconomics and pathological factors in oral diseases. In Figure 1, we present the KDD process proposed for this work, dividing it into three major steps: i) Step A - Data Collection, ii) Step B – Analytical data pre-processing, and iii) Step C – Dental Data Mining and Evaluating Models.

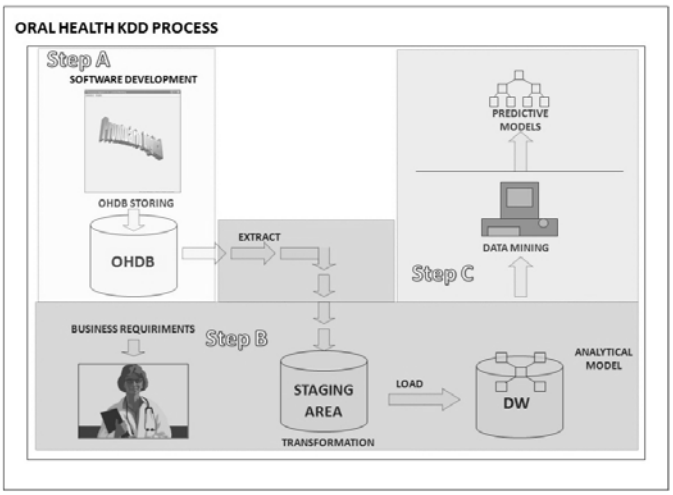

Figure 1. Oral health KDD Process

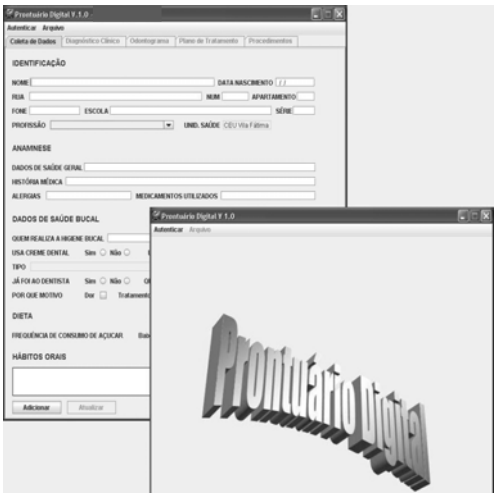

Figure 2. Instrument to data insertion in the OHDB

4.1 Step A – Data Collection

We developed an application in JSE (Java Standard Edition) client/server (see Figure 2) to collect and store the data into the Oral Health Database (OHDB), which is the operational database established in (Blomberg, 2009).

After developing the data collection application, we performed a survey of family files collected at Vila Fátima CEU, in which we identified 675 family files, arranged sequentially by order of creation. We planned a protocol for collecting dental records from the files so we started the data insertion activity through the developed application. In 2009, we had the collaboration of five dentistry students to the data collection step, ending with a population of 642 dental records collected, as shown in Table 1.

Table 1. Results of Data Collection Step

Total amount of family files: 675 Totals:				
Patients	Dental Records	Dental charts	Dental Visits	Dental Procedures
598	642	235	2031	3103

With the conclusion of the data collection step, we observed some peculiarities regarding the poor quality of data recorded in the dental records. Particularly, the number of missing fields of information was quite high, as presented in Table 2.

Table 2. Fields with Missing Data

	Anamnesis		Dental Chart	
Total data	642	100%	642	100%
Complete data	343	53.4%	235	36.6%
Missing data	299	46.6%	407	63.4%

Similarly, we found that the absence of information not only limited the understanding of the patient's dental history (performed through the matching of oral epidemiological factors captured in the anamnesis and oral health fields), but also affected the analysis of dental diseases.

4.2 Step B – Analytical Data Pre-processing

We performed the second step (B) on the data population previously detailed. As an alternative to the (Fayyad, 1996) original process, we chose to build a data warehouse to perform analytical activities. In this kind of environment (see (Han, 2001)), the data is stored in a large historical repository (called Data

Warehouse, DW) and organized under the so-called multidimensional data model. Multidimensional modeling can be understood as a technique for analytical model conception, where typically the data are summarized and presented from different points of observation (dimension tables). Moreover, these observation points are measured by numerical quantities related to facts (fact tables) that we want to investigate.

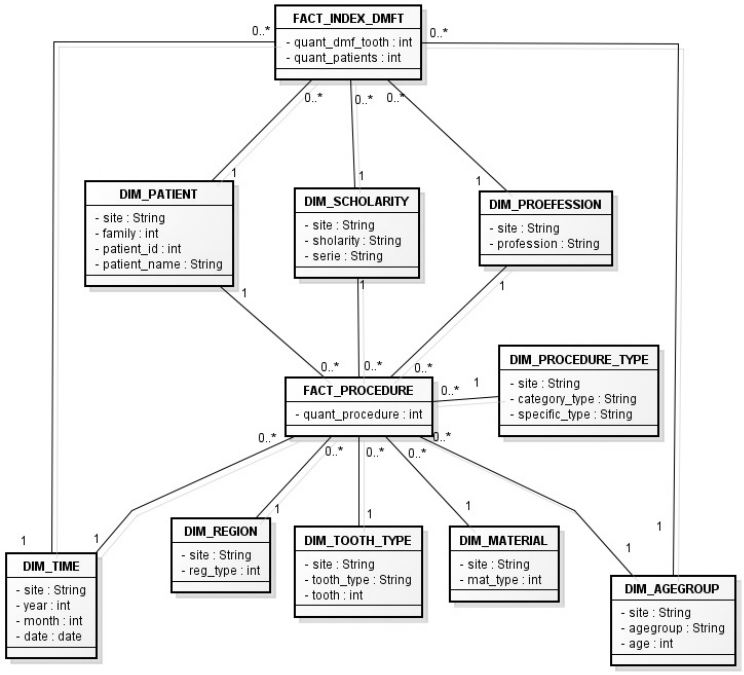

Figure 3. Oral Health Analytical Model

In Figure 3, we illustrate the analytical model we generated in this work. It is composed of nine dimension tables with socioeconomic and oral health data (dim_patient, dim_scholarity, dim_profession, dim_procedure_type, dim_time, dim_region, dim_tooth_type, dim_material, dim_agegroup), and two fact tables (fact_index_dmft and fact_procedure) for generating the amount of caries and enabling dental procedure analysis, such as restorative, endodontic, surgery and periodontal treatments.

In this change of the KDD process proposed by (Han, 2001), the analytical models within the DW are used to identify new opportunities for data mining, as well as to increase the potential for discovery of patterns of knowledge, through the multiple combinations of dimensions in varying levels of granularity. Through the implementation of this analytical model, the process of building the DW also involves the implementation of other operations such as data cleaning, integration and implementation of ETL steps (extraction, transformation and loading) within the DW. These operations (particularly data cleaning, data integration and data transformation) can be understood as the first pre-processing step of the original KDD process proposed by (Fayyad, 1996).

Motivated by the potentials benefits to identify new opportunities for data mining, we include a DW in the KDD process, using the Oracle SQL Developer tool to extract a dataset containing data from 19 tables of the transaction OHDB, as described in Table 3.

Table 3. Transactional OHDB extracted data.

Tables	Information
Patient, Profession_Data, Profession_Type, Scholarity	Socioeconomic
Dental_Record	Dental History
Oral_Health_Data, Oral_Habits, Diet	Anamnesis
Diagnostic	Periodontal Disease Incidence
Dental_Visit	Dental Visits
Dental_Chart	Caries Disease Incidence
Procedure_Data	Dental Procedures
Procedure_Categ	Dental Procedures Categorical Types
Procedure_Specif	Dental Procedures Specific Types
Material	Medications and Restorative Materials

After the extraction step, the data were transferred to a staging area to perform cleaning (missing values treatment, typing errors correction and standardization of nomenclature) and transformation (summarization and aggregation of data). These pre-processing steps are laborious though essential to achieving solid results in data mining. Once extracted and transformed, we loaded the data stored in the staging area to the DW. For this purpose, we used the SQL Developer tool to import data, yielding a total of 6 tables (fact_index_dmft, dim_patient, dim_ scholarity, dim_ profession, dim_time and dim_agegroup).

4.3 Step C – Dental Data Mining and Evaluating Models

In the last step of our KDD process, we selected data from 6 tables loaded from the DW and started to generate the predictive models related to the incidence of dental caries. We also selected and pre-processed data from 4 transactional tables in order to generate predictive models related to periodontal diseases. Table 4 depicts the complete dataset for the data mining step.

Table 4. Dataset for the Data Mining Step

Table	Model	Information
fact_index_dmft	analytical	caries disease incidence
dim_patient, dim_ scholarity, dim_ profession, dim_time, dim_agegroup	analytical	socioeconomic
oral_health_data, oral_habits, diet	transactional	anamnesis
diagnostic	transactional	periodontal disease Incidence

Based on these models, the aim of this study was to identify new hypotheses on the influence of socioeconomic factors and related diseases in the incidence of caries and periodontal diseases. These new hypotheses could be used by managers in oral health so they are able to plan preventive actions directed to specific risk groups.

We used the selected data in CSV format (comma-separated values) and applied the C4.5 algorithm (Quinlan, 1992) (java version implemented in the Weka) data mining tool for inducing predictive models in the form of decision trees. Among the reasons for choosing a decision-tree induction algorithm, we highlight the fact that decision trees offer a graphical representation that resembles the human reasoning, so it makes it easier for the domain specialist to interpret the results and make decisions accordingly (Freitas, 2010) and (Barros, 2012).

We performed 10-fold cross-validation for evaluating the effectiveness of this approach. Also, we had the support of the domain expert to verify the usefulness of the hypotheses generated by the data mining algorithm.

In Figure 4, we illustrate a predictive model for classifying patients according to the incidence of caries, where Y (leaf node presented as a rectangle) means the positive cases and N negatives cases.

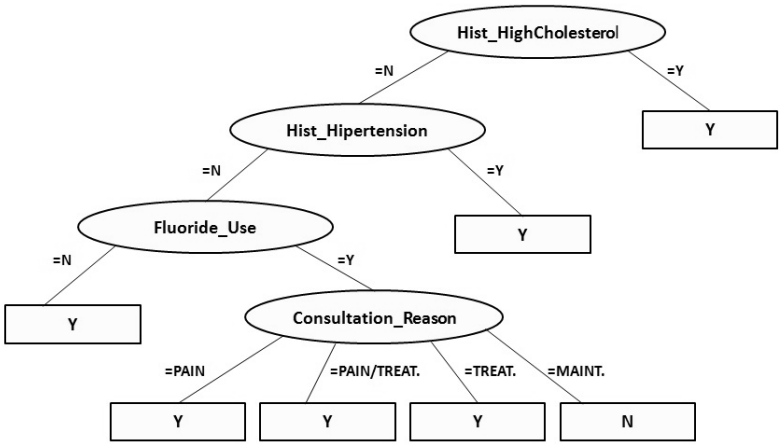

Figure 4. Predictive Model of Carie Incidence

This model could correctly identify the positive cases with an accuracy of 91.95%, even though it presented limitations on the prediction of negative cases. The predictive model suggests that high-cholesterol (Hist_HighCholesterol = Y) and hypertension (Hist_Hipertension = Y) diseases are related to incidence of caries, which is in fact could be a consequence of the type of food involved (excess of salt and sugar).

Similarly, the model suggests that caries are usually verified in dental visits that occurred due to pain or treatment, being less common in maintenance dental visits.

We generated a second predictive model that provides information on periodontal disease incidence, where Y (leaf node) means positives cases and N negatives cases. By analyzing the model illustrated in Figure 5, there seems to be a strong influence of genetic factors, for instance "Periodont_Family" node, in the incidence of new positives cases. Still according to the model, other situations of periodontal disease cases are related to advanced age (over 42 years) or for those patients between 11 and 42 years old with bad brushing habits (Brushes_More1x = N) and high sugar consumption (Sugar_Intake + 4x).

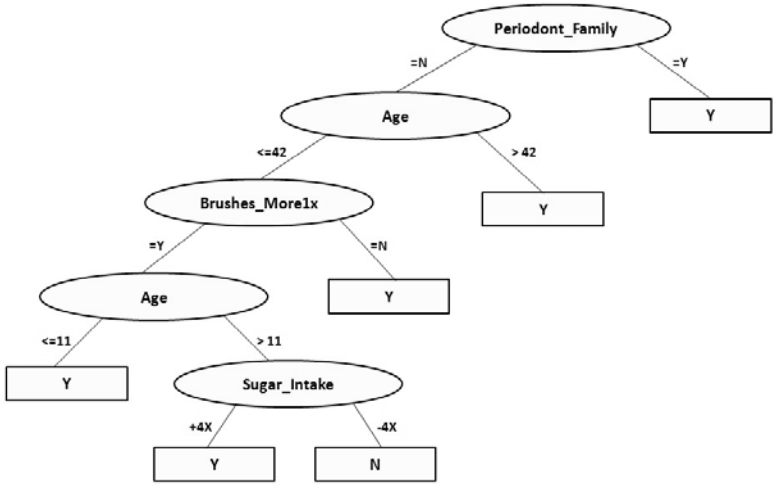

Figure 5. Predictive Model of Periodontal Disease

As we have seen along this paper, data mining should not be viewed as an isolated procedure, but as a step within the process of knowledge discovery. Results of the data mining step are directly influenced by previously-performed steps, such as data preprocessing, which receives as input raw data and provide as output structured data, suitable for being processed by a mining algorithm.

A typical example of this influence can be seen in Figure 6, which illustrates a prediction model induced from data that did not undergo pre-processing steps. In this model, we observed some unexpected distortions

in the relationship between oral health habits, such as the floss use, sweets intake, and fluoride use, to the dental caries indicator (considering "Very low", "Very High" and "Moderate" as levels of severity).

Among the distortions mentioned, observe the classification of "Moderate" for patients that did not consume sweets but used fluoride, whereas those that did not consume sweets but did not use fluoride were classified as "Very Low". This may be seen as a contradiction, since it is common sense to expect the rankings to be reversed. We believe that such distortions have been caused by factors related to data quality (i.e., missing fields) and hence the importance of performing pre-processing steps prior to data mining.

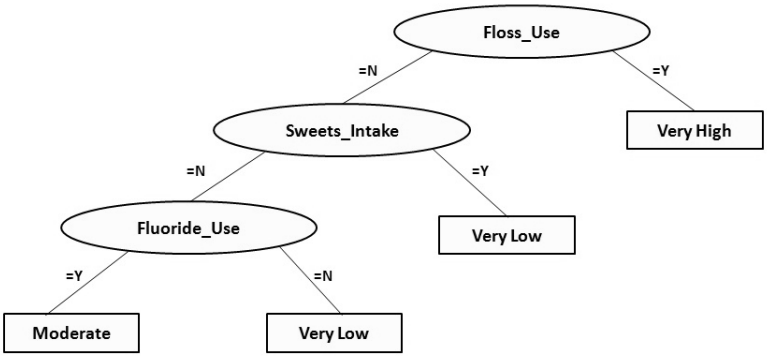

Figure 6. Model induced from Data That Did Not Undergo Pre-Processing Steps

5. CONCLUSION

The main contribution of this study is focused in the developed and documentation of a full KDD process, providing the following advantages:

a) Solution robustness: Through the construction of a data warehouse environment, we add robustness to the solution. The DW enables the analysis of large volumes of data, which is a fundamental requirement for public data management scenarios. Conversely, a DW environment may not be acceptable for analysis of small volumes of data, since it demands a considerable effort to be implemented.

b) Comprehensibility of models: Although the poor quality of data directly influenced the quality of the generated models, we believe that the KDD process we have developed is able to generate models easily understandable by health managers, enabling the generation of new hypotheses and providing a better basis for decision-making. A future solution to the data quality problem would be the adoption of computer systems for recording information of the dental records, instead of manual writing. Thus, we could delegate to the computer system tasks such as information consistency-check, significantly increasing the quality of the data and of the models that are generated.

c) Unconventional data processing: Among the typical characteristics of this type of scenario in healthcare, we highlight the importance of adopting techniques for the treatment of unconventional data such as textual descriptions and temporal data. For the treatment of temporal data, we modeled dimensions that allowed us to observe the factors related to different periods of time. However, for the extraction of predictive models for dental caries, we adopted a selection criterion based on the choice of the latest dental chart of each patient. For extracting the data used for inducing the predictive models, we adopted preparation practices based in the stemming technique. However, we understand that new software based on natural language processing can increase the quality of the processing of unstructured text.

As future work we intend to continue the study at the Vila Fátima CEU by exploring new areas of work with the Faculty of Dentistry, such as the recognition of new patterns from the analysis of medical images.

ACKNOWLEDGEMENT

This study received support from the National Council of Scientific and Technological Development (CNPq), in accordance with the research grant MCT/CNPq70/2009.

REFERENCES

Baldani M.H. et al, 1996. Associação do Índice CPO-D com indicadores socioeconômicos e de provisão de serviços odontológicos no Estado do Paraná, Brasil. *In Cad Saúde Pública*, Vol. 20, Nº 1, pp. 143-152.

Barros, R.C. et al, 2012. A Survey of Evolutionary Algorithms for Decision-Tree Induction. *In IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, Vol. 42, Nº 3, pp. 291-312.

Brazilian Ministry of Health, 2002. *Falando Sobre Câncer da Boca*. INCA, Rio de Janeiro, Brazil.

Brazilian Ministry of Health, 2006. *A Política Nacional de Saúde Bucal do Brasil: Registro de uma Conquista Histórica.* José Felipe Riani Costa, Luciana de Deus Chagas, Rosa Maria Silvestre (Orgs.). Organização Pan-Americana da Saúde, Brasília, Brazil.

Blomberg, L.C. et al, 2009. Development of an oral health database for the management of clinical records. In *Revista Odonto Ciência*, Vol. 24, Nº 3, pp. 249-253.

Chuang, L.I. et al, 2011. Support Vector Machine-based Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes. *Lecture Notes in Engineering and Computer Science*, Vol. 1, pp. 426-429.

Fayyad, U. et al, 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. In *Communications of the ACM*, Vol. 39, Nº 11, pp. 27-34.

Freitas, A. et al, 2010. On the Importance of Comprehensible Classification Models for Protein Function Prediction. *In IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7, Nº 1, pp.172-182.

Gansky, S.A., 2003. Dental data mining: potential pitfalls and practical issues. *In Advances In Dental Research*, Vol. 17, pp.109-114.

Han, J. and Kamber, M., 2001. *Data Mining: concepts and techniques*. Morgan Kaufmann, San Francisco, USA.

Ito A. et al, 2011. Risk assessment of dental caries by using Classification and Regression Trees. *In Journal of dentistry*, Vol. 39, Nº 6, pp. 457-463.

Lin, C. et al, 2009. The development of dentist practice profiles and management. *In Journal of Evaluation in Clinical Practice*, Vol. 15, pp. 4-13.

Majumder, S.K. et al, 2005. Support vector machine for optical diagnosis of cancer. *J Biomed Opt*. Vol.10, Nº 2, pp.1-14.

Montenegro, R.D. et al, 2008. A Comparative Study of Machine Learning Techniques for Caries Prediction. *Proceedings of 20th IEEE International Conference on Tools with Artificial Intelligence*. Washington, USA, pp. 477-481.

Oliveira, A.L.I. et al, 2005. A comparative study on machine learning techniques for prediction of success of dental implants. *Proceedings of the 4th Mexican international conference on Advances in Artificial Intelligence*. Springer-Verlag, Berlin, Heidelberg, pp. 939-948.

Powell, L.V., 1998. Caries prediction: A review of the literature. *In Community Dentistry and Oral Epidemiology*, Vol. 26, Nº 6, pp. 361-371.

Quinlan, J.R., 1992. Learning with Continuous Classes. *Proceedings of 5th Australian Joint Conference on Artificial Intelligence*, Singapore, pp. 343-348.

Sharma, S., 2011. Comparing the performance of data mining techniques for oral cancer prediction. *Proceedings of the 2011 International Conference on Communication, Computing & Security*, New York, USA, pp. 433-438.

Stewart, P.W. and Stamm, J.W., 1991. Classification tree prediction models for dental caries from clinical, microbiological, and interview data. *J Dent Res*, Vol. 70, pp. 1239–1251.

Tamaki, Y. et al, 2009. Construction of a Dental Caries Prediction Model by Data mining . *In Journal of Oral Science.* Vol. 51, Nº 1, pp. 61-68.

Tan, P. N. et al, 2006. *Introduction to Data Mining.* Addison-Wesley, Boston, USA.

Zhang, Q. and Helderman, V.P.W.H., 2006. Caries experience variables as indicators in caries risk assessment in 6-7-year-old Chinese children. *Journal of Dentistry*, Vol. 34, Nº 9, pp. 767–781.