TECHNOLOGY IN HEALTH: KNOWLEDGE DISCOVERY IN PUBLIC HEALTH DATABASES: STUDY OF VIRAL HEPATITIS IN THE STATE OF PARANÁ, BRAZIL

Carla Machado da Trindade[1], Claudia M. Cabral Moro[1], Márcia Gil Aldenucci[2], Júlio César Nievola[1], Deborah Ribeiro Carvalho[1], Samuel Jorge Moysés[1]

[1]Pontifical Catholic University of Paraná - Rua Imaculada Conceição, 1155 - Prado Velho - Curitiba – PR

[2]Paraná State Health Department - Rua Piquiri, 170 – Rebouças - Curitiba – PR

carlamtrindade@hotmail.com, c.moro@pucpr.br

***Abstract**: This paper shows some benefits that the methodology of Knowledge Discovery in Databases (KDD), using the data mining technique, can bring when used in databases that store data on the health of population, describing the analytical process applied in the identification of the behavior of viral hepatitis. The KDD process involved the application of the classification technique on 2003 data, stored in the Notifiable Diseases Information System of Paraná Health Department. Sixty-five characteristics of 3.063 investigation forms were analyzed, resulting in 4 decision trees and 99 classification rules. Of these rules, 60 were analyzed and the other ones were discarded because they did not contemplate enough examples to be considered valid or they had a high number of errors. The method enabled the database to be explored thoroughly, as well as enabling an increased number of appraised characteristics, the identification of problems relating to the quality of the data and to information routinely used by the epidemiological surveillance service. It was also possible to discover the occurrence of hepatitis B in its chronic form in children under 13 years old. This knowledge, unperceived in the original database, can help in the formulation of new policies, suggesting the importance of this method as an form of scaling up routine strategies, with the aim of reducing and controlling diseases. This technology contributes towards the exploration of the data stored by the various Health Information Systems.*

***Keywords:** Public Health Information Systems, Knowledge Discovery in Databases, Data Mining, Epidemiological Surveillance, Viral Hepatitis*

1. INTRODUCTION

Disease control and prevention, harm reduction, increased health promotion and increased life expectancy are amongst the most important tasks attributed to Public Health services.1 For this to be possible there must be continuous enhancements to areas such as epidemiological surveillance, whether it be through the advancement of knowledge on the health-disease process, or through the scaling up of methodological and technical information resources that contribute towards the identification of behaviors and how the main problems that daily affect individuals and groups are distributed. This process enables the establishment of actions that improve both health and quality of life standards [1].

Currently in Brazil an important source of the knowledge used in the process of epidemiological surveillance is the information captured and stored by means of several different Health Information Systems (HIS), used by the Brazilian Ministry of Health (MoH) and the State and Municipal Health Departments [1]-[2]-[3]. Since 1994 the system that officially holds data on compulsorily notifiable diseases throughout the entire country is the Notifiable Diseases Information System (known as SINAN in Portuguese).1 This system provides the information used in formulating strategic policies, at Municipal, State and Federal, level for the control and prevention of notifiable diseases [2].

The problem approached by this study was the identification of the behavior of viral hepatitis, by using the Knowledge Discovery in Databases – KDD method, and the data mining technique. The purpose of the study was to contribute towards epidemiological surveillance efforts to enhance information available for decision-making and the implementation of better-founded actions.

Of the diseases that are notified and constantly investigated through the SINAN system and which are subject to epidemiological surveillance, viral hepatitis is the one that most occurs in the population in the State of Paraná, with approximately 5,000 new cases each year [4]. Hepatitis is currently a disease that the World Health Organization (WHO) is closely accompanying, since in some of its forms it can evolve into chronic liver cancer or even lead to death, in addition to its high transmission potential. It is believed that 3% of the world’s population is infected with hepatitis C and 5% with hepatitis B, so that hepatitis accounts for more cases of infection than the AIDS virus [2]-[5].

The extraction of information from databases, such as those relating to hepatitis on the SINAN system, is carried out routinely using manual procedures that require large operational efforts whilst not being efficient. A further procedure involves the use of structured queries or the use of TABWIN (data tabulation software supplied by the Ministry of Health). These procedures are self-limiting with regard to the amount of characteristics assessed, thereby restricting the acquisition of knowledge and requiring long processing times.

In order to improve the efficiency of the extraction of relevant and, possibly, innovative data contained in large databases, such as the SINAN system, the use of KDD as an alternative is proposed [6]-[7]. In terms of health services, KDD has been used with the aim of extracting useful knowledge which may assist in disease prevention, obtaining more accurate medical diagnoses, treatments, prognoses, detection of anomalies, control of hospital infections and epidemiological research.[8]-[9]-[10]-[11]-[12]

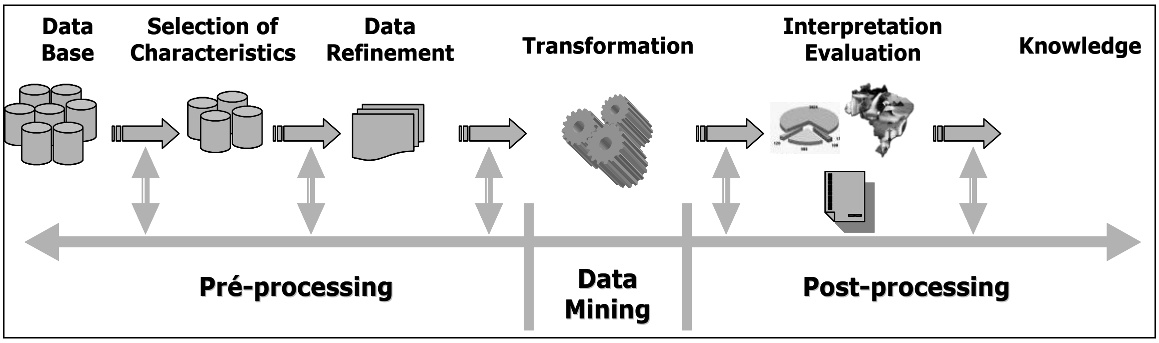

Figure 1 KDD Process

Generally speaking the KDD process is divided into three stages: pre-processing, data mining and post-processing, as shown in Figure 1 [7].

Pre-processing involves the acquisition of databases that may contain useful information on the domain in question. Later, the characteristics which represent this domain are selected and the data refined, so that duplicated records are removed, as are absent data, values are normalized, outliers (values that do not follow general data distribution patterns) are also removed, in order for the data to be prepared so that it can be processed in line with the defined target [11].

In the data mining stage the data is transformed into information by means of finding regularities, restrictions, patterns and significant relationships [7]-[11]. Three possible techniques can be carried out during the KDD process: clustering, association and classification. In this study, with the aim of keeping the description as brief as possible, only the technique that was selected will be dealt with, since it appeared to be the most coherent and robust in relation to the objectives of the study, namely the classification technique. This consists in discovering the relationship between possible predictive attributes and the target attribute. These relationships are represented by a decision tree and can be described through rules of classification. Some mined attributes generate a complete tree, containing all the relationships encountered, whilst simplified trees are less complex [7]-[11]

During post-processing the results obtained using the mining algorithm are represented graphically, in addition to being analyzed and interpreted, thereby concluding the KDD process and generating the obtainment of knowledge [7]-[11]

Despite the use of KDD in health services, no published studies were found in the literature that relate the exploring of databases used by the public health service in Brazil as sources of information on viral hepatitis. However, it is a recognized fact among the community of those interested in the KDD method that databases relating to viral hepatitis represent a challenge in terms of generating good results. As such, during a world conference held in Italy in 2004 the challenge was made to discover relevant patterns through the use of a database containing information on hepatitis patients [13].

Therefore the objectives of this study were defined as being: (a) to explore the data stored on the SINAN system, whereby the researchers took the epistemological stance of suspending any pre-established concepts on viral hepatitis; (b) to discover and portray the epidemiological profile of viral hepatitis, using KDD, verifying the relationships between the diverse attributes of the database in the identification of the aetiological classification of hepatitis, through trees and rules generated by a mining algorithm; (c) to identify the attributes contained on the hepatitis investigation forms that are most relevant for defining their aetiological classification; and (d) to evaluate the quality of the stored data.

2. METHODS

During the pre-processing stage, data was acquired from the viral hepatitis investigation from records on the SINAN system, relating to the year 2003 in Paraná. The system held 5,063 investigated cases. Some of the 134 attributes existing on the spreadsheet that was generated were not taken into consideration, as exemplified in Table I. Specifically these attributes were: patient identification data (27), data on the hospital (10), data on dates (7), data with a low percentage of replies (25), as well as attributes that could not be identified using the data dictionary or the investigation form dictionary, or those without adequate values, such as: profession – which is a field that is not filled in properly.

Table I Example of removed attributes

Field	Description	% Filled in
in_aids	Not in the data dictionary	0.00%
imunihres	Immuno-Histo-Chemical	0.12%
parceiros	Not in the data dictionary	0.26%
sanaguaout	Not in the data dictionary	3.69%
genotipo	HCV RNA / PCR /genotype	4.74%

As such, 65 attributes remained in the database and were effectively analyzed. In order to verify hepatitis behavior, only confirmed cases of the disease were analyzed: therefore, only 4,738 of the initial 5,063 records were included.

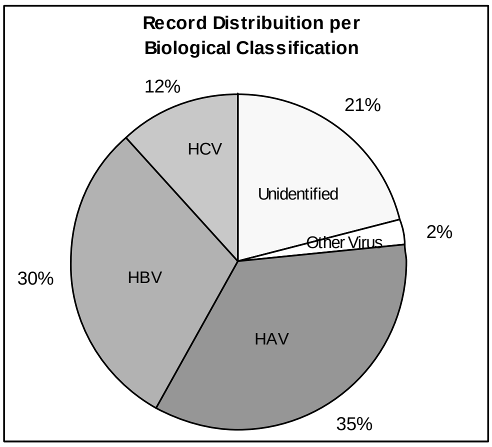

Figure 2 Distribution of confirmed cases per aetiological classification

The KDD technique used was that of classification. When using this technique it is desirable that there be a balance between the classes of aetiological classification so as not to compromise the induction capacity of the classifying algorithm. Therefore, in order to ensure this balance, only confirmed hepatitis A and B cases were selected, being equivalent to 3,063 records, of which 1,628 refer to hepatitis A and 1,435 to hepatitis B. The records relating to other aetiologies of the disease were eliminated owing to the large difference encountered when comparing the percentages of the number of recorded cases of HAV and HBV with the number of confirmed cases, as can be seen in Figure 2.

During the data mining stage the C4.5 algorithm was used. It enables comprehensible results to be generated, in addition to being a tool that is widely used by KDD researchers [14]. This algorithm was initially used using the wrapper method in order to obtain a set of attributes best able to distinguish between the behavior of hepatitis A and B. This method involves the addition or removal of attributes used as predictors of a given aetiological classification, thereby enabling the calculation and analysis of error rates resulting from the application of the algorithm. The attributes resulting from each iteration (application of the algorithm) are used as the input for the next application and this process is repeated in a cyclical manner until the criterion for concluding the application of the algorithm is reached. In this study, the use of all the input attributes in the formation of both complete and simplified decision trees was the criterion for determining the final iteration.

The data mining stage was carried out twice when selecting the attributes, once using the wrapper methodology and a second time using the characteristics routinely assessed by a specialist (epidemiology professional who routinely works with the SINAN system).

During the wrapper method stage 64 prediction attributes were used initially in addition to the target attribute, defined as the aetiological classification. In the first iteration the algorithm used 18 and 12 characteristics (attributes), respectively for the complete and simplified trees, distinguishing between hepatitis A and B. The second iteration used only the 18 characteristics of the complete tree as predictors. This process resulted in the use of 17 and 12 attributes, respectively. This process was repeated six times.

During a second stage of data mining, conceived as an alternative means of contents validation, the algorithm was applied to the characteristics routinely assessed by the specialist with the aim of comparing the specialist’s results with the results relating to the set of attributes obtained using the wrapper method. The set of attributes indicated by the specialist included aetiological classification, age of the patient, and vaccination against hepatitis B, its form, evolution and confirmation of the diagnosis. In this case the set of attributes selected using the wrapper method as being the most representative in distinguishing hepatitis A and B included: AgHBs and Hav-IGM testing, the evolution of the disease, previous suspected infection, form of the disease, municipality of residence and of notification and the patient’s age, vaccination against HBV and three or more sexual partners.

During the post-processing, in order to facilitate the analysis of the results in the most objective manner with the help of the specialist, statistics and graphic representation were used, by means of decision trees, graphs and maps. Classification rules were followed in the production of the text.

3. RESULTS AND DISCUSSIONS

In all, four decision trees were generated, being two complete and two simplified trees. One complete tree and one simplified tree resulted from iteration with those attributes identified by the wrapper method as being the most representative, whilst the others resulted from the application of the data miner algorithm to the attributes indicated by the specialist. As the graphical representation method used was the same for all the trees, in this study only the graphic representation of the complete tree will be presented, having been generated from the attributes identified by the wrapper method as being more representative. This tree resulted in 49 classification rules for distinguishing hepatitis A and B. The lowest number of decisions, or tests, needed to classify both aetiologies was six, whilst the highest number of tests in this structure was seven. The graphic representation of the complete tree generated using the wrapper method can be seen in Figure 3.

Exemplifying the interpretation of the tree, the first node represents whether the HAV-IgM test is reactive or not. Therefore, in the case of rule 1, once the HAV-IgM test is proven to be reactive, the form of the disease is checked to determine whether it is chronic or fulminant hepatitis, whether the patient or the infection is asymptomatic, thus classifying the target attribute as hepatitis B; the remaining rules can be interpreted in a similar manner.

The structures relating to the simplified tree obtained from the fifth iteration and the trees resulting from the attributes indicated by the specialist were represented in a similar manner [6].

The process of transforming the structure of the decision tree into classification rules generated 31 rules, of which only 18 were evaluated, as 5 did not contemplate any of the records and the remainder presented a probability of accuracy of less than 1%, being nearly 100% incorrect, and were therefore discarded.

Table II Tree structures

Number of Attributes	**Tree**		**No. of attributes used in the best classification**		**Highest Level**	**Number of Rules**		
	Structure	Nodes	HAV	HBV		Generated	Evaluated	Zeroed
11	Simplified	31	5	3	6	19	11	3
	Complete	49	6	6	7	31	18	5
6	Simplified	14	3	1	4	8	7	1
	Complete	74	4	2	12	41	24	7

Figure 3 Complete tree using the wrapper method

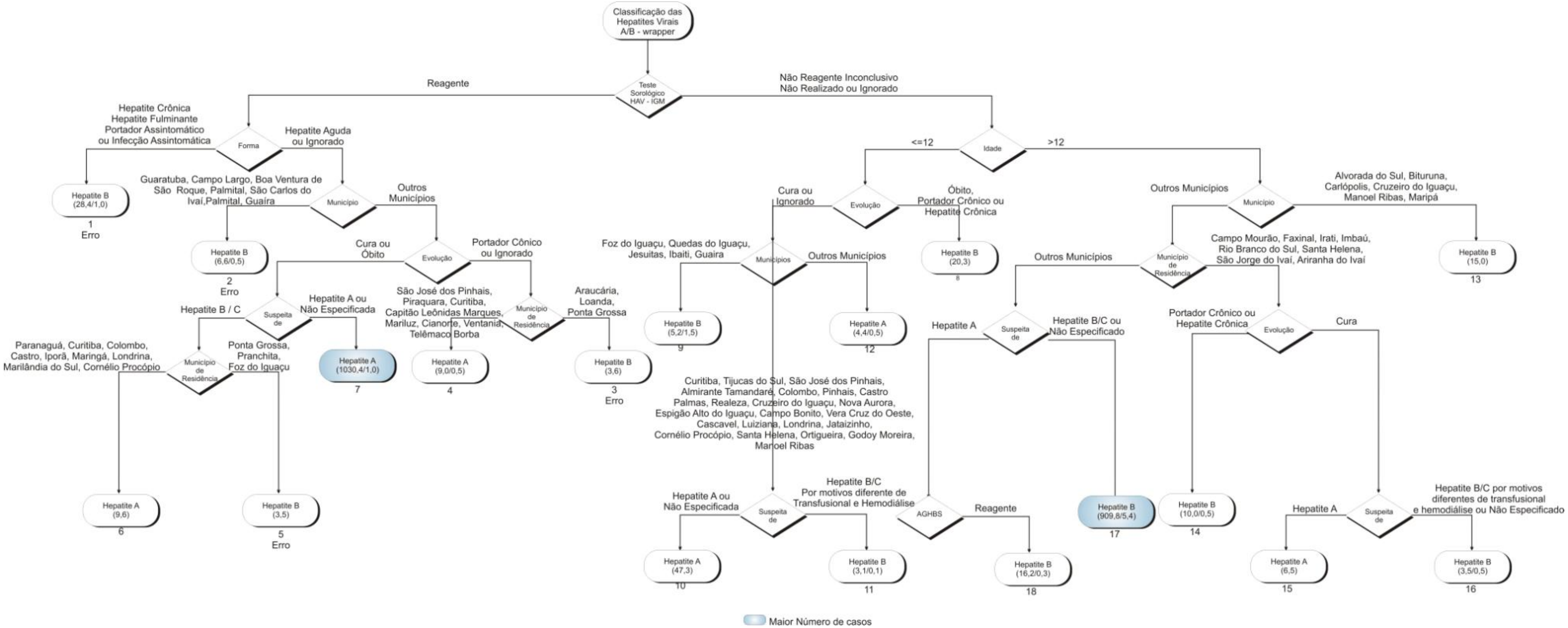
Classificação das Hepatites Virais A/B - wrapper
Teste Sorológico HAV - IGM
Reagente
Não Reagente Inconclusivo Não Realizado ou Ignorado
Hepatite Crônica Hepatite Fulminante Portador Assintomático ou Infecção Assintomática
Forma
Hepatite Aguda ou Ignorado
Hepatite B (28,4/1,0)
1 Erro
Guaratuba, Campo Largo, Boa Ventura de São Roque, Palmital, São Carlos do Ivaí,Palmital, Guaíra
Município
Outros Municípios
Hepatite B (6,6/0,5)
2 Erro
Cura ou Óbito
Evolução
Portador Cônico ou Ignorado
Hepatite B / C
Suspeita de
Hepatite A ou Não Especificada
São José dos Pinhais, Piraquara, Curitiba, Capitão Leônidas Marques, Mariluz, Cianorte, Ventania, Telêmaco Borba
Município de Residência
Araucária, Loanda, Ponta Grossa
Paranaguá, Curitiba, Colombo, Castro, Iporã, Maringá, Londrina, Marilândia do Sul, Cornélio Procópio
Município de Residência
Ponta Grossa, Pranchita, Foz do Iguaçu
Hepatite A (1030,4/1,0)
7
Hepatite A (9,0/0,5)
4
Hepatite B (3,6)
3 Erro
Hepatite A (9,6)
6
Hepatite B (3,5)
5 Erro
<=12
Idade
>12
Cura ou Ignorado
Evolução
Óbito, Portador Crônico ou Hepatite Crônica
Hepatite B (20,3)
8
Foz do Iguaçu, Quedas do Iguaçu, Jesuitas, Ibaiti, Guaira
Municípios
Outros Municípios
Hepatite B (5,2/1,5)
9
Hepatite A (4,4/0,5)
12
Curitiba, Tijucas do Sul, São José dos Pinhais, Almirante Tamandaré, Colombo, Pinhais, Castro Palmas, Realeza, Cruzeiro do Iguaçu, Nova Aurora, Espigão Alto do Iguaçu, Campo Bonito, Vera Cruz do Oeste, Cascavel, Luiziana, Londrina, Jataizinho, Cornélio Procópio, Santa Helena, Ortigueira, Godoy Moreira, Manoel Ribas
Hepatite A ou Não Especificada
Suspeita de
Hepatite B/C Por motivos diferente de Transfusional e Hemodiálise
Hepatite A (47,3)
10
Hepatite B (3,1/0,1)
11
Outros Municípios
Município
Alvorada do Sul, Bituruna, Carlópolis, Cruzeiro do Iguaçu, Manoel Ribas, Maripá
Hepatite B (15,0)
13
Município de Residência
Campo Mourão, Faxinal, Irati, Imbaú, Rio Branco do Sul, Santa Helena, São Jorge do Ivaí, Ariranha do Ivaí
Outros Municípios
Hepatite A
Suspeita de
Hepatite B/C ou Não Especificado
AGHBS
Reagente
Hepatite B (16,2/0,3)
18
Hepatite B (909,8/5,4)
17
Portador Crônico ou Hepatite Crônica
Evolução
Cura
Hepatite B (10,0/0,5)
14
Hepatite A
Suspeita de
Hepatite B/C por motivos diferentes de transfusional e hemodiálise ou Não Especificado
Hepatite A (6,5)
15
Hepatite B (3,5/0,5)
16
Maior Número de casos

Table III Results of the statistical calculations

Method	Iteration	HAV	HAV Error	HBV Error	HBV	HAV Sensitivity	HAV Specificity	Positive Predictor	Negative Predictor	Accuracy	False Positive	False Negative
Wrapper	1st	469	31	22	397	96%	7%	94%	95%	94%	6%	5%
	2nd	469	31	22	397	96%	7%	94%	95%	94%	6%	5%
	3rd	469	31	22	397	96%	7%	94%	95%	94%	6%	5%
	4th	469	20	22	408	96%	5%	96%	95%	95%	4%	5%
	5th	476	20	15	408	97%	5%	96%	96%	96%	4%	4%
	6th	476	20	15	408	97%	5%	96%	96%	96%	4%	4%
Special-list	7th	447	31	44	397	91%	7%	94%	90%	92%	6%	10%

The rule identified as being the most relevant in this study (rule 15) was the one that enabled the identification of municipalities with cases of chronic HBV in children under 13 years old. This information as well as information on the other structures is available in Table II.

Table III shows the results of prediction and of the statistical calculations used to evaluate the attributes used in each iteration.

In the statistical analysis, it was observed that the first iterations were the ones that resulted in the highest specificity values. In the fourth iteration, an improvement was obtained in relation to the positive predictor value and accuracy. The fifth and sixth iterations generated similar results and produced the best values in relation to the other iterations regarding sensitivity rates, negative predictor values, and rate of correct results and lowest proportion of false negative values.

The fifth iteration using the wrapper method was considered the best iteration as it presented the best statistical results and the smallest tree structure. When comparing the statistical results of the fifth iteration with the results obtained by applying the C4.5 algorithm, in relation to the attributes indicated by the specialist, it was found that the set of attributes indicated by the specialist only presented better results for specificity.

4. Conclusions and Future Perspectives

Considering the limitations of this study, principally in relation to the quality of the data available in the database that was studied, it would appear to be reasonable to admit that it produced relevant results. This allows us to suggest the importance of scaling up the methodological repertoire used with non-standard technologies being incorporated in the production of information that is useful for epidemiological surveillance.

Even at the pre-processing stage, it was possible to identify problems with the stored data, both in relation to some fields not being filled in and others not being filled in properly. In addition, some fields in the database did not exist on the investigation forms, or even in the data dictionary. The lack of the use of database consistency routines was also identified. All these observations are important in the case of an epidemiological surveillance system, in relation to which a high level of reliability and consistency is expected, since it has important repercussions on the health of the population.

In the data mining stage the wrapper method made it possible to verify that, of the set of attributes on the SINAN system investigation form, those that best distinguish between hepatitis A and B are: AgHBs and HAV-IgM testing, evolution of the disease, previous suspected infection, form of the disease, municipality of residence and of notification and the age of the patient. When the attributes selected automatically by the mining tool were compared to the attributes selected by the specialist, coincidence occurred in relation to the patient's age, vaccination against hepatitis B, as well as its form and evolution. Regarding the attributes routinely used by the specialist, confirmation of the diagnosis was the only attribute not identified by the wrapper method.

In the post-processing stage, among the results produced using the C4.5 algorithm, confirmation was obtained of some aspects of hepatitis that are already known and that have already been related in the literature. However, an important contribution originated through the use of KDD, recognized as new knowledge, was the identification of early cases of chronic hepatitis B patients, namely in children under 13 years old. These cases may possibly be related to the vertical transmission (mother to child) of the hepatitis B virus (HBV). There are indications that the risk of the disease evolving into cirrhosis or cancer of the liver is 25% greater in these individuals [5].

The recommended procedure for avoiding this type of infection is the joint use of immunoglobulin and vaccination against hepatitis B in the first 12 hours of the lives of children born to mothers infected with HBV. This practice reduces the possibility of vertical transmission by 90% [5]-[15]. Identification of HBV in pregnant women can reduce potential sources of infection both of the mother, by using correct prevention measures, and of the child, by means of vaccination. People with chronic HBV become reservoirs of the hepatitis B virus and provide continuity to the disease's transmission chain [15].

It is thought that there are currently more than 325 million people with chronic HBV, or approximately 5% of the world population. In addition, this disease is responsible for around one million deaths each year globally [5]. It is believed that the magnitude of the disease is due to it being 51 to 100 times more infectious than AIDS, since HBV can be transmitted both by sexual contact and by contact with the skin. As for cases due to vertical transmission, in Brazil it is thought that the test for detecting HBV in pregnant women should be adopted as a routine part of public health service ante-natal care, similar to the United States, thereby reducing the number of people infected.

Geographic representation of chronic cases of HBV in children under 13 years old enabled it to be verified that, excluding the cities with the largest number of inhabitants (Curitiba and Londrina), the southwest region of the State of Paraná has the highest concentration of patients infected with the disease. This knowledge, as well as other knowledge generated by applying KDD, including knowledge on the quality of the database used in epidemiological surveillance, is important for the State's epidemiological surveillance policy and for the definition of strategic actions for improving the health of the population.

Exploring and analyzing health data usually results in complex implications, since disease behavior depends on a variety of factors, and changes over the years. For this reason, the use of fixed reports, or pre-defined analysis, results in limitations when the aim is to identify new behaviors or characteristics of diseases. The use of KDD has enabled the generation of rules, patterns of regularities and characteristics of the diseases from the data available on the database, and not just from the literature or the routine activities of the specialist. This increase in methodological possibilities not only enabled the confirmation of previously known data, but also new relevant knowledge, as well as inconsistencies, irregularities, exceptions.

In this study it was possible to: a) explore diverse characteristics of viral hepatitis, identifying relationships existing between the predictive characteristics and generating rules for classification for aetiological specification; b) identify individual cases of HBV infection in low age groups, at an advanced stage of the disease, which was a fact that the epidemiological surveillance had not previously observed; c) to verify the existence of discrepancies and/or inconsistencies in relation to the investigation form, the data dictionary and the database available on the SINAN system, in addition to identifying flaws in the filling in of the data.

It was seen that the use of KDD enabled the identification in databases of public health problems that otherwise are hidden and need to be detected. It was also possible to explore data on viral hepatitis in detail, thereby making it possible for the epidemiological surveillance service to have better knowledge of the structure and the stored data on this illness, thereby enabling it to improve its prevention and control actions and policies. Was better harnessed the potential of data from the use of technology.

The recommendation is therefore made to Paraná's State Health Department and to other bodies responsible for epidemiological surveillance to use and explore the data available on other diseases stored on the SINAN system, as well as to apply KDD to other Health Information Systems.

References

[1] Mussi-Pinhata Marisa M. Imunogenicidade da vacina contra hepatite B iniciada precocemente em pré-termos: implicações para a prevenção. J. Pediatr. Porto Alegre. mar/abr. 2004. v.80 n.2. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0021-75572004000200003&lng=pt&nrm=iso%3E&tlng=pt. Accessed on: 13/09/2005.

[2] O'Carroll, P.W.; Yasnoff, W.A.; Ward, M.E.; Ripp, L.H.; Martin, E.L.. Public health informatics and information systems. Health informatics series, Springer – Verlag, New York, p. 16-40, 2002.

[3] Li J.; Wong L. Rule-based data mining methods for classification problems in biomedical domains. In: practice of knowledge discovery from databases, Pisa, Italy. In: A tutorial note for the 15 the European conference on machine learning (ECML) and the 8th European conference on principles and practice of knowledge discovery in databases (PKDD), September, 2004. 32.

[4] Medronho, R.A. Epidemiologia. São Paulo: Editora Atheneu, 2003.

[5] Focaccia, R.; Veronesi, R., Tratado de infectologia: hepatites virais, São Paulo: Atheneu, 1997.

[6] Rouquayrol, M. Z.; Almeida Filho, N. de. Epidemiologia e Saúde. Rio de Janeiro: Medsi. 5ª edição. 1999.

[7] Brossette, S.E.; Sprague, A.P.; Hardin, J.M.; Waites, K.B.; Jones, W.T.; Moser, S.A.. Association rules and data mining in hospital infection control and public health surveillance. American Medical Informatics Association, 1998, v 5, n.4. p. 373-381. Avaialbe at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=61314. Accessed on: 26/11/2004.

[8] Trindade, C.M.. Identificação do comportamento das hepatites virais a partir da exploração de bases de dados de saúde pública Dissertação (Mestrado em Tecnologia em Saúde). Pontifícia Universidade Católica do Paraná. Curitiba. 2005.

[9] Fayyad, U.; Uthurusamy, R. Data mining and knowledge discovery in databases. Association for Computing Machinery. Communications of the ACM. New York. Nov 1996. v39. n11. p. 24-27

[10] Ohsaki, M.; Sato, Y.; Yokoi, H.; Yamaguchi, T. A rule discovery support system for sequential medical data – in the case study of a chronic hepatitis dataset. In: Proceedings of practice of knowledge discovery from databases. Croatia. 2003. p. 22-26. Available at: http://lisp.vse.cz/challenge/ecmlpkdd2003/proceedings/Ohsaki.pdf. Accessed on: 28/08/2005.

[11] Olaru, C.; Wehenkel, L. Data Mining, IEEE Computer Applications in Power. v. 12, n3, p. 19-25, July 1999

[12] Thuraisingham, B. Data Mining: A premier for understanding applying data mining. IEEE IT Pro. 2000. p. 28-31.

[13] Alves, V.; Neves J.; Maia, M.; Nelas, L. A computational environment for medical diagnosis support systems. Proceedings of the Second International Symposium on Medical Data Analysis, Madrid, Spain. p. 44-49. 2001.

[14] Quinlan, J.R. Induction of Decision Trees. Machine Learning. Boston, 1986. V1. p 81-106.

[15] Fayyad, U.; Piatetsky-Shapirio, G.; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. Association for Computing Machinery. Communications of the ACM. New York. Nov 1996. v39. n11. p. 27-34.