NEURALTB WEB SYSTEM
Support to the Smear Negative Pulmonary Tuberculosis Diagnosis

Carmen Maidantchik, José Manoel de Seixas, Afrânio Kritski, Fernanda C. de Q Mello
Rony T. V. Braga, Pedro H. S. Antunes
Federal University of Rio de Janeiro, Cidade Universitária, C.P. 68504, 21945-970, Rio de Janeiro, Brazil

João Baptista de Oliveira e Souza Filho
Federal University of Rio de Janeiro, Cidade Universitária, C.P. 68504, 21945-970, Rio de Janeiro, Brazil
Graduated Dept., Celso Suckow Technological Education Center, Av. Maracanã 229, 20271-110, Rio de Janeiro, Brazil

Keywords: Decision support systems, neural networks, web technology, SNPT diagnosis.

Abstract: The World Health Organization estimates that one third of the world population is infected by *mycobacterium tuberculosis*. Tuberculosis (TB) affects mainly poor health places in developing countries. Therefore, it became mandatory to develop more efficient, fast, and inexpensive analysis methods. This paper presents a decision support system that uses neural networks to sustain TB diagnosis. The output is the probability that a patient has or not the illness and an assigned risk group. The NeuralTB system encapsulates the knowledge needed for efficient anamnesis interview integrated to demographic and threat factors typically known for tuberculosis diagnosis. It was developed with the Web technology and data were described with a markup language to enable an efficient communication and information exchange among experts. Data collected during the whole process can be used to identify possible new factors or symptoms, since the infection transmission may evolve. This information can also support tuberculosis control governmental entities to define effective actions to protect the health and safety of the population.

1 INTRODUCTION

Although effective antimicrobial therapies and suitable diagnosis tests are already available, the number of tuberculosis cases increases each year. In particularly, the smear negative pulmonary cases (SNPT) are hardly diagnosed. The World Health Organization (WHO) estimates that one third of the world population is infected by *mycobacterium tuberculosis* (WHO 2002). So, there are approximately already 2 billion infected persons. Every year, 8 to 9 million new cases appear and 1.7 million individuals die. Therefore tuberculosis (TB) is still a serious public health problem worldwide.

The rapid growth of the disease is related to several factors, particularly the HIV epidemic, the increase of social differences in many countries, and the deterioration of health services mainly among poverty population. These problems occur more frequently in urban areas due to migrations that have been happening over the last decades. Additionally, the TB vaccine is not very effective.

Diagnostic tests either fail to identify at least half of cases or are accurate but expensive, and it is often difficult for patients to complete the necessary six-month course of treatment, which contributes to new drug-resistant strains of the disease. Currently, the pulmonary tuberculosis diagnosis is still based on bacilloscopy directly from the sputum smear, which lacks sensitivity (around 50%). Besides that, this method has no utility in the diagnosis of extra-pulmonary TB. On the other hand, the *mycobacterium tuberculosis* culture that presents a higher sensitivity (80%) requires 4 to 6 weeks for the result. Such long period for the confirmation of the infection delays the beginning of the treatment and allows the contagion among other people. The smear-negative transmission rate of *mycobacterium tuberculosis* corresponds to 17% among exposed individuals (Sarmiento *et al*, 2003). Moreover, in deprived countries, only some control programmes permit culture performance in their primary-care diagnostic (Santos *et at*, 2006). Consequently, fast and accurate diagnosis of SNPT could provide lower

Maidantchik C., Manoel de Seixas J., Kritski A., C. de Q Mello F., T. V. Braga R., H. S. Antunes P. and Baptista de Oliveira e Souza Filho J. (2007).
NEURALTB WEB SYSTEM - Support to the Smear Negative Pulmonary Tuberculosis Diagnosis.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - AIDSS*, pages 198-203
DOI: 10.5220/0002366401980203

morbidity and mortality, and case detection at a less contagious grade.

The culture result for the *mycobacterium tuberculosis* can be obtained by automated diagnoses methods, commercialized in the health care area. Besides being expensive, those methods have not been validated in different epidemic situations. Their use in routine conditions is restricted to reference or research laboratories (Perkins and Kritski, 2002). New diagnosis tests as well as the use of statistical models to support the SNPT analysis constitute a real challenge. To predict the patient probability on having TB researchers employ neural network (El-Solh *et al.*, 1999) or multivariate logistic regression and classification tree (Mello, 2001). Santos (2003) uses neural networks and classification trees to identify patients with clinical-radiological suspicion of SNPT. When formulated in a systematic way and implemented with high qualified data, statistical models can be representative of the clinical problem under evaluation and could be useful for physicians in their clinical routine, as well as for public health policy administration (Castelo *et al.*, 2004).

This paper presents the NeuralTB Web system that aims at supporting the SNPT diagnosis in health care units of limited resource areas. The system comprises artificial neural networks as a model for diagnosing the infection. The software was developed using the Web technology and offers user-friendly and intuitive interfaces for symptoms registering, patient monitoring, and result retrieval. The data stored in each health care unit are easily merged in a central database.

This paper is organized as follows. Section 2 describes the artificial neural network diagnosis model and the data set in study. Section 3 presents the NeuralTB Web system and Section 4 explains implementation details. Conclusions and future work are described in Sections 5.

2 THE DIAGNOSIS MODEL

In order to avoid TB becoming a pandemic that would cause serious illness in people and spread quickly throughout populations, the fight against the disease includes discovering new tools for prevention, diagnosis support, and treatment.

Software engineering and the Internet have an important role in the battle against infirmities that are quickly spread among different countries. Computing programs register diseases, symptoms, and locations where infected people live. Internet sites communicate new drugs, treatments, and risk factors related to maladies, allowing an exchange of expertise. Searchable indexes provide access to medical directories and research programs.

We propose a decision support system that health and medical professionals may use to sustain the diagnosis of SNPT under routine conditions in the hospitals and health care units. It should be clearly stated that the purpose of the system is not to replace physicians. The proposed model suggests that mathematical modeling for classifying SNPT cases could be an useful tool for optimizing the utilization of expensive tests, and to avoid costs of unnecessary anti-TB treatment. The diagnosis corresponds to an ongoing process that requires accurate investigation and, therefore, the system output should be analyzed together with interviewing, inspection, auscultation, and examination of the laboratory results.

One concern of the project was to develop a tool that would be suitable for areas of limited resources. Therefore, the requirements of low cost, easy access, and user-friendliness were considered. Since the target disease is geographically spread among different places, our group decided to implement the system using the Web technology.

SNPT experts defined a set of symptoms that would determine whether a patient would have or not the disease. Based on this an artificial neural network model was developed. The network output corresponds to the probability that a patient have or not SNPT and the risk group (low, medium, high level risk) for which the patient would belong to. The developed Web system registers the input information, executes the neural networks code, stores the result, monitors the patients data, and manages data files. TB experts supported the project development process, validating each step to guarantee that the resulting system would achieve the project goals.

2.1 Data Set for Modelling

In order to determine the set of symptoms and characteristics that would indicate the infection, 136 patients agreed to participate. They were referred to the University Hospital of Federal University of Rio de Janeiro, from March, 2001 to September, 2002, with clinical-radiological suspicion of SNPT.

The input data set corresponds to information from anamnesis interview integrated to demographic and risk factors typically known for tuberculosis diagnosis. Forty three per cent of the patients actually showed TB in activity. Initially, clinical variables were considered: age, coughs, spit, sweat, fever, weight loss, chest pain, shiver, dyspnea, diabetes, alcoholism, and others.

2.2 The Artificial Neural Network

The artificial neural network model is fed from data collected from the questionnaires filled by patients in the health care units. The dichotomy variables were codified as -1 and 1, representing the absence or the presence of a symptom, respectively. Three categories were allowed for qualitative variables: -1 (lack of an indication), 1 (presence of the symptom), and 0 (ignored). In model development, relevance of variables was also addressed, which allowed more compact network designs. Starting from 26 variables, the relevance analysis (Seixas *et al*,. 1996) showed that models could be developed considering 12 or just 8 variables. Such variable suppression was also validated by TB experts.

With respect to network topology, a fully-connected multilayer feedforward architecture trained with backpropagation algorithm was designed. Input nodes varied, according to data compaction scheme, from 26 to 8. The network has a single output neuron, and training targets were defined as 1 (active TB) and -1 (otherwise). The number of neurons in the single hidden layer also varied according to model complexity, from 3 to 4 neurons. The hyperbolic tangent is the activation function for all neurons.

The risk group assignment was obtained by means of a modified- ART clustering procedure (Vassali *et al.*, 2002). Risk group assignment was certified by TB experts as symptoms identified in each risk group are also considered by the TB experts in a detailed exam.

Due to restrictive statistics of the database, cross validation (Kohavi, 1995) was used for defining both training (network design) and testing (performance evaluation) tests. For each cross validation test, the training set comprised 80% of the patients and the remaining 20%, formed the test set. Performance was evaluated in terms of sensitivity and specificity for the testing set. Considering twelve input variables, it was possible to obtain both high sensitivity (100%) and specificity (80%).

3 THE NEURALTB WEB SYSTEM

Within this project, our group aims at providing an open and secure platform that supports an efficient and fast distribution of collaborative applications. An untied architecture allows the integration with other systems, dynamic processes, and heterogeneous data repositories. The computing solution must also provide a good connectivity to any data placed anywhere. These requirements guarantee the accessibility of the system either in health care units or hospitals, independently of their location. The software group used interoperable technologies for the system development.

Initially, a system version that could work over the Internet was developed. The health care units would only need to have a browser and an Internet connection. All data and processing would be respectively stored and performed in a central server. For the units that do not have Internet access, a local version is used. The neural network program is also locally executed and the result is placed together with the patient data. The information that is stored in the computer of all health care units is periodically transferred to a central server that collects the data into a main repository.

The local version does not require an Internet access for its execution. However, it is more laborious to update and maintain the system due to its geographical distribution. On the other hand, a version that works over the Internet avoids compatibility problems since an unique version runs in the server. It also facilitates the data transfer from the health care units to the central repository.

Within the system, there are three user categories: administrator, attendant, and physician. Administrators can insert new users, modify user attributes, and perform actions related to data files, and system installation. Attendants may include and edit patient personal data and his/her symptoms. Physicians perform actions on patient data and are the only ones to have access to the network output.

3.1 Input Data Form

In order to fill the questionnaire, attendants are taught to analyse individual's physical condition, to give further information about each question, and to explain the importance of providing the correct answer. Therefore, in order to facilitate the data input into the system, a hypertext form was designed. The form is composed by text fields to include the patient name and date of birth. Other items allow the selection of only one option among three available alternatives (lack of an indication, presence of the symptom, and the patient do not know the answer). In case of relationships among items, the choice of an option automatically obliges the selection of the respective option in another item. For example, if hemoptysis (coughing up blood) is chosen as an existing symptom than the existence of cough has also to be selected.

Two kinds of support needed during the questionnaire filling were implemented. The first group is related with typing errors, data mismatch, and correlations between items. As an example, empty data is not accepted. Concerning the date of

birth, the system automatically validates the days and months, bissextile years, etc. Then, the system calculates the patient age that can be confirmed right away. In case of errors, an alert window comes out informing the mistake.

The second group of help appear as an alert window with further information about the item, explaining how to make a question, and how one can interpret the answer. The specification of this type of support required the extraction of the knowledge used during an anamnesis interview. Hendriks and Vriens (1999) and Probst *et al* (1999) suggest a basic set of fundamental activities to systematically manage knowledge: identify important knowledge that can be used; capture and store useful knowledge in a repository; maintain knowledge in the storage area through update or removal of outdated information.

Subsequent to the inclusion of a new patient, the system automatically executes the neural network program and stores the output together with the information that was entered through the form. Later on, physicians may analyze the result together with other clinic and laboratorial information.

3.2 Patient Monitoring

In order to recover data from the system repository, the user may define one or more attributes, such as cough, sputum, fever, etc and associate with a specific value. The attributes operate as filters that trigger the patients which data fits the query. The "+" option allows the definition of other conditions in the query, i.e., the inquiry can combine several attributes using logical operators (and, or). One condition within a query can be removed by selecting the "-" option.

In case the option "Search for" is selected without specifying values to an attribute, all patients and respective data are presented, ordered by name, in a table format. The attribute names are placed in the heading of the table. The system also provides a facility through which a physician can set the patient as already analyzed. Therefore, it is also possible to search for patients which data were not investigated yet, supporting the information management.

3.3 Probabilities and Risk Groups

When the neural network is fed from a new patient data, it provides as an output the classification probability of the patient to have or not the TB. In case of TB identification, the system also provides the risk group to which the patient belongs.

In case the neural network classifies the patient as having the TB, the output will be presented as the sentence "the patient has P% of having the TB", where "P%" represents the probability for active TB according to the artificial neural model. On the other hand, in case the neural network classifies the patient as not having TB, the output will be presented as the sentence "the patient has P% of not having the TB", where "P%" represents the probability for no active TB, according to model.

The risk groups are presented in a graphical way like a car traffic light using a universal color code. A patient fits in only one of the three risk groups that are drawn as circles painted with red, yellow, and green colors to symbolize, respectively low, medium, and high risk, as presented in Figure 1. The patient is represented in the figure as the "x" letter and the closer he/she is to the center, higher is the probability that the patient belongs to the group.

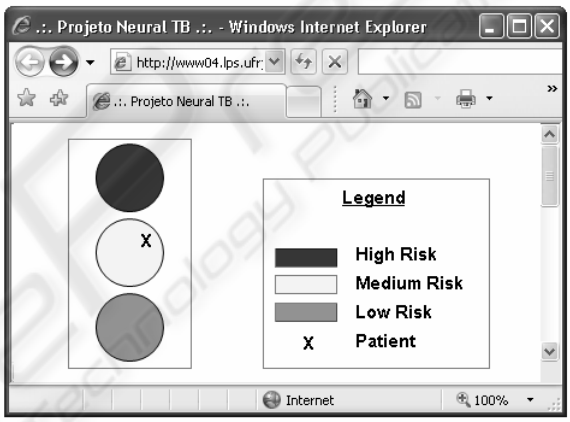

Figure 1: Risk group representation (in grey scale).

3.4 File Management

The NeuralTB Web system was designed to operate in geographically distributed environments. Therefore, the information obtained during the anamnesis interview together with the neural network output is stored in an archive. The files from different units are transferred to a central repository were all patient information are inserted into a database. In case of changes in the patient data, the system manages the records to be copied or sent again in order to update the main database. For security reasons, NeuralTB also provides a backup functionality.

The central repository stores the name of the health care units, associating the patients data with the units where their information was collected. Further information about the units region, such as geographical relation of health conditions to socio-economic status and poverty rates, may be specified, structured and integrated into the database.

Medical and health professionals of TB research groups can access the central repository to extract information that can help on the development of new disease analysis methods and the identification of new demographic and risk factors that can be used later for the tuberculosis diagnosis.

4 THE IMPLEMENTATION

The data representation format is an important aspect that was considered to efficiently manage the whole information. Markup languages, as XML, can be used to describe knowledge structures and to support institutional memory development (Rabarijaona *et al*, 2000, Cook, 2000). XML may provide a standard structure to communicate and interchange data and knowledge among diverse systems. The language allows the creation of multiple visions of the same item and also provides an easy mechanism to capture, store, present and recover information. Considering these benefits, we developed an XML-based approach to describe the different types of information manipulated during the whole process of the SNPT diagnosis.

The group identified three stages where data had to be properly represented: during the anamnesis interview, for describing the patient data, and to extract statistical information within research activities. TB specialists warned that risk factors, the questions made to the patients, and relationships among the stored records may vary according to locations or other factors, such as multidrug resistance (MDR) that is one of the main causes of ineffective treatment of new TB cases. Therefore, the use of XML facilitates the maintenance of the knowledge represented in the three stages. The tags identify the current data and new tags can be easily defined. The language also allows the definition of associations among diverse types of information.

In order to assure the compatibility between the data structure and the system functionalities, the NeuralTB interface and operations were conceived and designed in a way to guarantee its correct execution independently on both the way information is organized and the kind of records that are manipulated. This requirement is achieved by creating the interface with the system operations in the moment the application is executed. The interface reads the XML and presents all commands associated with the tags. So, in case one record type is excluded, the system will do not perform any operation related to this information. On the other hand, in case a new record type is included, it is mandatory to define both the tag that identifies the data and the corresponding operation.

Another advantage of using XML is that it facilitates the integration among data that comes from different health care units and hospitals. Markup languages make easy the combination of heterogeneous records. The use of XML also allows uniform systems interoperability and offers efficient mechanisms for information recovery.

The system was designed in modules to facilitate its integration with other applications. The interface between the system and the neural networks program is also defined through a XML file. This archive describes the name of the application, the neural network weight vector to be used, and the output. This approach facilitates when users want to execute a different neural networks program or update the weight vector.

4.1 Computing Requirements

The NeuralTB Web System runs over the Apache HTTP Server for both UNIX and Windows XP operating systems. The system provides a shell executable of setup programs that automatically install a directory structure and respective files in the computer of the health care unit or hospital. The hardware requirements are: PC computers with USB driver for file transfer (in case of local version) or an Internet connection, and with 128 MB, or preferentially, 256 MB RAM memory.

The system operations were implemented as CGI (Common Gateway Interface) programs, using the C language. The Javascript language is used to write functions embedded in HTML pages and interact with the Document Object Model (DOM) of the page to perform tasks not possible in HTML alone. The Cascading Style Sheets (CSS) language is used to style the web pages written in HTML and format the XML documents.

In order to draw the risk group representation, the GD graphics library was used. GD is an open source code library for the dynamic creation of images, allowing programmers to easily generate PNG, JPEG, GIF (among other images formats), from many different programming languages (C, Perl, and PHP).

The central repository was implemented using MySQL, an open source relational database management system (RDBMS) that uses Structured Query Language (SQL).

5 CONCLUSIONS

Decision support systems can be considered as useful elements for helping physicians on the tuberculosis diagnosis. The application can be used

as a learning tool since it gathers information, defined by experts, that is needed for the tuberculosis diagnosis.

The NeuralTB system can be easily installed in hospitals or health care units and can also be executed in portable computers that are carried to different regions. The approach to incorporate the knowledge into the system, allowing an easy maintenance of the information, guarantees the lifetime of the proposal.

Currently the NeuralTB system is being installed in health care units in the Rio de Janeiro, the number one city for TB cases in Brazil. This effort will facilitate the implantation of a network to integrate diverse professionals and specialists in tuberculosis. During the system operation we will be able to validate the impact of this initiative.

As next steps, we intend to integrate the NeuralTB input data form with other questionnaire items used during an anamnesis interview. Actually, the proposal is to integrate the input form with the system that is used in the hospital reception. As a result, the attendance will use a single environment to register all data related to patients. Another enhancement is to develop queries in the central database to extract the information that comes from the various health care units. The knowledge of which information should be extracted can also be modelled and incorporated into the repository. Data quality metrics (Chapman, 2005) will also be applied to ensure network information quality. This is quite important as network performance relies on the accuracy of questionnaire answers. The continuous update of the neural model with incoming new data is also being developed. This involves stability studies and the monitoring of TB main features, trying to track disease evolvement in time and geographically.

We expect that the accomplishments of this project bring social benefits, allow a better integration of the information technology in the diagnosis domain, and provide an infrastructure to enable an efficient communication and information exchange among tuberculosis experts.

ACKNOWLEDGEMENTS

The authors thank the Tuberculosis Research Unit, Faculty of Medicine, Federal University of Rio de Janeiro, for making available the data used in this work and CAPES, CNPq, and FAPERJ for financially supporting this project.

REFERENCES

Castelo A., Kritski A.L., Werneck A., Lemos A.C., Ruffino Netto A., et al., 2004. Brazilian Directives for Tuberculosis. *J Brás Pneumo*, 30 (supl 1). 1- 86. In Portuguese.

Chapman, A., 2005. Principles of Data Quality, Report, Global Biodiversity Information Facility.

Cook, J., 2000. XML Sets Stage for Efficient Knowledge Management, *IT professional*, v.2, n.3, 55-57.

El-Solh, A.A., Hsiao, C.-B., Goodnough, S., Serghani, J., Grant, B.J.B., 1999. Predicting Active Pulmonary Tuberculosis using an Artificial Neural Network. *Chest*, 116, 968–973.

Hendriks, P., Vriens, D. 1999. Knowledge-Based Systems and Knowledge Management: Friends or Foes?. *Information & Management*, v.35, n.2 (Feb), 113-125.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In International Joint Conference on Artificial Intelligence.

Mello, F.C.Q., 2001. Smear Negative Pulmonary Tuberculosis Predicting Models, Ph.D. Thesis, Medicine Faculty, Federal University of Rio de Janeiro, Brazil. In Portuguese.

Perkins, M.D.., Kritski, A.L., 2002. Perspectives. Diagnostic Testing in the Control of Tuberculosis. In: *Bull WHO*, 80 (6), 512-513.

Probst, G., Raub, S., Romhardt, K. 1999. *Managing Knowledge: Building Blocks for Success*, 368 pp, ISBN: 0-471-99768-4.

Rabarijaona, A., Dieng, R., Olivier, C., Quaddari, R. 2000. Building and Searching an XML-Based Corporate Memory, *IEEE Intelligent Systems*, v.15, n.3 (May), 56-63.

Sarmiento, O., Weigle, K., Alexander, J., Weber, D.J., Miller, W., 2003. Assessment by Meta-Analysis of PCR for Diagnosis of Smearnegative Pulmonary Tuberculosis, *Journal of Clinical Microbiology*, 41, 3233-3240.

Santos, A.M. 2003. Neural Networks and Classification Trees Applied to Smear Negative Pulmonary Tuberculosis Diagnosis, Ph.D. Thesis, COPPE/ UFRJ, Rio de Janeiro, Brazil. In Portuguese.

Santos, A.M., Pereira, B.B., Seixas, J.M., Mello, F.C.Q., Kristski, A.L., 2006. Neural Networks: an Application for Predicting Smear Negative Pulmonary Tuberculosis. In: Balakrishnan, N.; Auget, J.L.; Mesbah, M.; Molenberghs, G. (org.). In: *Advances in Statistical Methods for The Health Sciences*. 279-292.

Seixas, J.M., Calôba, L.P., Delpino, I., 1996. Relevance Criteria for Variable Selection in Classifier Design. In: International Conference on Engineering Applications of Neural Networks, 451-454.

Vassali, M.R., Seixas, J.M., Calôba, L.P., 2002. A Neural Particle Discriminator Based on a Modified Art Architecture. In: IEEE International Symposium on Circuits and Systems, v. II., 121-124.

World Health Organization (WHO), 2002. Stop TB annual report 2001.