Chapter XVI
Mining Tuberculosis Data

Marisa A. Sánchez
Universidad Nacional del Sur, Argentina

Sonia Uremovich
Universidad Nacional del Sur, Argentina

Pablo Acrogliano
Hospital Interzonal Dr. José Penna, Argentina

ABSTRACT

This chapter reviews the current policies of tuberculosis control programs for the diagnosis of tuberculosis. The international standard for tuberculosis control is the World Health Organization's DOT (Direct Observation of Therapy) strategy that aims to reduce the transmission of the infection through prompt diagnosis and effective treatment of symptomatic tuberculosis patients who present at health care facilities. Physicians are concerned about the poor specificity of diagnostic methods and the increase in the notification of relapse cases. This works describes a data-mining project that uses DOT´s data to analyze the relationship among different variables and the tuberculosis diagnostic category registered for each patient.

INTRODUCTION

Tuberculosis has been a major killer disease for several years. It is estimated that around 1.6 million people die each year from tuberculosis; and in 2005 figures indicate that approximately 8.8 million people developed the disease (World Health Organization, 2007b). The international standard for tuberculosis control is the World Health Organization's DOT (Direct Observation of Therapy) strategy that aims to reduce the transmission of the infection through prompt diagnosis and effective treatment of symptomatic tuberculosis patients who present at health care facilities. The treatment is based on the strict supervision of medicines intake. The supervision is possible thanks to the availability of an information

system that records the individual patient data. These data can be used at the facility level to monitor treatment outcomes, at the district level to identify local problems as they arise, at provincial or national level to ensure consistently high-quality tuberculosis control across geographical areas (World Health Organization, 2007b). In Argentina, the Health Ministry gathers DOTS data since 1996.

Identification of individuals latently infected and effective treatment are important parts of tuberculosis control. The DOTS strategy recommends identification of infectious tuberculosis cases by microscopic examination of sputum smears. However, this function requires a strong laboratory network and high-quality sputum smear microscopy. In children, the diagnosis of pulmonary tuberculosis is difficult because collection of sufficient sputum for smear microscopy and culture is difficult. The HIV epidemic has led to huge rises in incidence of tuberculosis in the worst affected countries, with disproportionate increases in smear-negative pulmonary tuberculosis in children and adults (Getahun, 2007). Additionally, the use of chest radiography for diagnosis of pulmonary tuberculosis can be compromised by poor film quality, low specificity, and difficulties with interpretation (World Health Organization, 2004).

Physicians are concerned about the poor specificity of current methods. In particular, they want to analyze diagnosis of childhood tuberculosis. In addition, the notification rate of relapsed cases is slightly increasing so they are interested in finding patterns that can explain this trend. Thus, the purpose of this study is to review current policies of local tuberculosis control programmes for the diagnosis of tuberculosis.

Data analysis is vital for answering these questions. The availability of DOTS records gives an opportunity to use Data Mining techniques such as demographic clustering, or decision trees. In particular, decision trees have an immense prediction power and provide an explanation of the results. Decision trees are widely used in medicine (Jianxin, 2007; Prakash, 2006; Šprogar, 2002; Cios, 2002).

The chapter is further structured as follows: next section provides definitions of data mining concepts and tuberculosis terms. Then, we present our data mining project and highlight our key findings and the main issues and problems that have arisen during the project. Finally, we summarize the contributions of the chapter.

BACKGROUND

Technology evolution has promoted the increase in the volume and variety of data. The amount of data increases exponentially with time. As a consequence, the manual analysis of this data is complex and prone to errors. When the amount of data to be analyzed exploded in the mid-1990s, knowledge discovery emerged as an important analytical tool. The process of extracting useful knowledge from volumes of data is known as knowledge discovery in databases (Fayyad, 1996). Knowledge discovery's major objective is to identify valid, novel, potentially useful, and understandable patterns of data. Knowledge discovery is supported by three technologies: massive data collection, powerful multiprocessor computers, and data mining (Turban, 2005).

Data mining derives its name from the similarities between searching for valuable business information in a large database, and mining a mountain for a vein of valuable ore. Data mining can generate new business opportunities by providing automated prediction of trends and behaviors, and discovery of previously unknown patterns.

A data mining project compromises a multi-step, iterative process. CRISP-DM (CRoss-Industry Standard Process for Data Mining) was conceived in late 1996 by DaimlerChrysler, SPSS and NCR as