Chapter 10
Exploring Disease Association from the NHANES Data: Data Mining, Pattern Summarization, and Visual Analytics

Zhengzheng Xing
Simon Fraser University, Canada

Jian Pei
Simon Fraser University, Canada

ABSTRACT
Finding associations among different diseases is an important task in medical data mining. The NHANES data is a valuable source in exploring disease associations. However, existing studies analyzing the NHANES data focus on using statistical techniques to test a small number of hypotheses. This NHANES data has not been systematically explored for mining disease association patterns. In this regard, this paper proposes a direct disease pattern mining method and an interactive disease pattern mining method to explore the NHANES data. The results on the latest NHANES data demonstrate that these methods can mine meaningful disease associations consistent with the existing knowledge and literatures. Furthermore, this study provides summarization of the data set via a disease influence graph and a disease hierarchical tree.

INTRODUCTION
The National Health and Nutrition Examination Survey (NHANES) is a nationwide survey conducted by the National Center for Health Statistics and some other health agencies since 1971 (CDC, n.d.). It aims at providing nationally representative information on the health and nutritional status of the population and tracking changes over time.
NHANES data has been used to evaluate the prevalence and risk factors of diseases in the population and to provide health guidelines. The prevalence of a disease is the percentage
Exploring Disease Association from the NHANES Data

of population having the disease. For example, in Beuther (2007) and Saydah et al. (2007), the NHANES data is used to study the prevalence of obesity and chronic kidney diseases over time and in different demographics groups (e.g., age, ethnicity and gender). A risk factor of a disease is a characteristic, condition or behavior that increases a person’s chance of developing the disease. The NHANES data has been used to verify the hypotheses of risk factors of chronic kidney (Saydah et al., 2007), obesity (Gangwisch et al., 2005), congestive heart failure (He et al., 2001) and some other diseases. The analysis results from the NHANES data have been used in the development of health related guidelines and public policies. For example, the early NHANES data revealed that the blood levels of lead among Americans were too high. The findings led to the federal regulations on reducing the amount of lead in gasoline, paint and soldered cans (Pirkle et al., 1998).

The NAHNES data contains a questionnaire component in which selected people are interviewed for their medical conditions and disease histories. It is a valuable data source for discovering disease associations among dozens of diseases. Disease associations can provide useful information in disease prevention, diagnosis and treatment.

There are some studies on evaluating correlated diseases by using statistical methods (He et al., 2001; Manjunath et al., 2003; Spence et al., 2003). The statistical methods focus on evaluating a number of pre-defined hypotheses of a set of risk factors or some associated diseases with respect to a particular disease. In contrast to the statistical methods, data mining methods aim at discovering the knowledge of associated diseases among a large number of diseases without any hypotheses. However, to the best of our knowledge, the NHANES data has not been systematically explored for mining associations among extensive diseases.

Is mining disease association patterns straightforward? One may think that association rule mining or association pattern mining (Agrawal et al., 2003) can provide an immediate solution. In an association rule about diseases \( A \Rightarrow B \), where \( A \) and \( B \) are two diseases, the probability that disease \( A \) appears in the population is called the support of the rule, and the probability that disease \( B \) appears in the condition of disease \( A \) appearing is called the confidence of the rule. Some other correlation measurements such as lift (Han et al., 2006), all-confidence (Omiecinski et al., 2003) and cosine (Han et al., 2006; Tan et al., 2002) are also proposed.

Since the number of people with diseases is usually much smaller than the number of healthy people, to mine association patterns of diseases, the support threshold often has to be set very low. Furthermore, diseases are very complex mechanisms. Different sub-types of a disease or people with different health conditions may have very different disease association patterns. Therefore, disease association patterns usually are not very strong. Consequently, a low confidence threshold has to be used in order to find many meaningful disease association patterns. Many other interestingness measures on association rules also meet some difficulties. For example, the lift for the patterns on high prevalence diseases is very different from the lift for patterns on low prevalence diseases.

Due to the complexity and diversity in disease association patterns, it is very difficult for a user to choose an appropriate threshold for a quality measure in the mining. If a user picks a low threshold in order to avoid missing some interesting patterns, the user may often be overwhelmed by a large number of rules and patterns which are hard to be analyzed and used.

To make disease association pattern mining practical and useful for health industry users, two problems need to be solved. First, to help users to understand the results, summarization of patterns and mining results should be provided. Second, a