Improving patient safety using data mining techniques and ICD-9 codes

Evi Guilliams
Research group Policy Management – Patient safety, Hasselt University, Diepenbeek, Belgium

Abstract
Purpose – The aim of this paper is to discuss how data mining techniques can be used to improve patient safety using administrative healthcare data.

Design/methodology/approach – By reviewing the existing literature and examining the National Hospital Discharge dataset, this research explains the opportunities and difficulties encountered with when applying data mining techniques on large healthcare datasets.

Findings – Data mining techniques could play an important role in medical decision support by enabling clinicians and managers to find valuable new patterns in data, such as detection of risk factors. This, in turn, can lead to a potential improvement of resource utilization and patient health. However, tangible results of empirical data mining research on administrative healthcare data are still rare.

Research limitations/implications – In this paper, we follow an exploratory approach by providing a theoretical framework for the use of data mining techniques on hospital data. Results of the actual implementation of these techniques on the Belgian Hospital Discharge dataset will be presented in a future research paper.

Originality/value – Many predictive data mining methods have been successfully applied to a variety of practical problems in medicine. However, very few papers have studied how data mining can be used to improve patient safety, or to explore patient safety indicators.

Keywords – patient safety, data mining, ICD-9 codes, adverse events

Paper type – Research paper
1. Introduction

In 1999, the Institute of Medicine (IOM) report ‘To Err is Human’ published disturbing figures about medical errors in US hospitals. More specifically, this report shows that each year between 48,000 and 98,000 people die from medical errors that occur in hospitals. Exact figures on the Belgian situation are not available, however it can be estimated from extrapolation that each year approximately 1,500 people die because of adverse events occurred in Belgian hospitals (Vleugels, 2003). These figures provide compelling evidence that medical errors pose daily risks throughout the healthcare system (Stelfox et al., 2006).

Patient safety research is concerned with the avoidance, prevention and amelioration of adverse outcomes or injuries stemming from healthcare systems instead of the underlying disease (Kohn et al., 1999). Central to this research is the notion that skilled and caring professionals can, and do make mistakes because, after all, to err is human. This is why it is vital that the improvement of the quality of healthcare is put at the top of national agendas and ways need to be sought to reduce these errors through the design of a safer health system (Kohn et al., 1999).

Healthcare organizations are generally characterized as information-dependent organizations. Knowledge-intensive technology is vital to these organizations as their environment becomes more and more complex (Chae, 2001). For this purpose, medical information systems are widely used and large amount of data on patients, clinical processes, and medical resource information are stored in digital form. This vast and growing collection of healthcare data and the willingness of clinicians to explore different technologies and methodologies to analyse this enormous amount of data have recently opened up new analytical possibilities for clinicians, operations researchers as well as information systems researchers (Kunene and Weistroffer, 2006). These results can be valuable to medical institutes because they may release important information, which may facilitate medical professionals to make better decisions (Lin et al., 1999).

In this context, over the last few years, the term data mining has been increasingly used in the medical literature (Belazzi and Zupan, 2006). In general, it can be described as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996). From a statistical perspective, it can be viewed as a computer automated exploratory data analysis of usually large datasets (Friedman, 1997). Data mining has already been applied with success to different fields of human endeavour, including marketing, banking, customer relationship management, engineering and various areas of science. However, its application to the analysis of medical data has until recently been relatively limited (Belazzi and Zupan, 2006).

In this paper, we will explore the possibilities to analyse the Belgian Hospital Discharge dataset in terms of patient safety indicators using data mining techniques. By reviewing the existing literature and examining the National Hospital Discharge dataset, we will highlight the opportunities and difficulties encountered with when applying data mining techniques on healthcare data. The scope of this paper will be limited to explorative research by providing a theoretical framework for the use of data mining techniques on hospital data. Results of the actual implementation of these techniques on the Belgian Hospital Discharge dataset will be presented in a future research paper.

2. Literature Review

2.1. Data mining

In general, the goal of data mining is to learn from data. More specifically, it is used to discover patterns and relationships in data, with an emphasis on large, observational databases (Friedman, 1997).
More specifically, data mining can be considered as a separate step of the ‘Knowledge Discovery in Databases’ (KDD) process (Fayyad et al., 1996)(see figure 1). This KDD process can be defined as the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data (Adriaans and Zantinge, 1996). It covers the overall process of extracting knowledge from very large databases starting from ascertaining the business goal to the eventual analysis of the results (Cabena et al., 1997). The additional steps in the KDD-process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge and proper interpretation of the results of data mining are essential to ensure that useful knowledge is derived from the data.

Furthermore, two broad categories of data mining strategies can be discerned: supervised and unsupervised learning (Matkovsky and Nauta, 1998). Supervised learning methods are deployed when values of variables (inputs) are used to make predictions about another variable with known values. Unsupervised learning methods can be used in similar situations, but are more frequently deployed on data for which a target with known values does not exist. An example of a supervised method would be healthcare organizations finding out through predictive modelling what attributes distinguish fraudulent claims. In supervised methods, the models and attributes are known and are applied to the data to predict and discover information. With unsupervised modelling, the attributes and models of fraud are not known, but the patterns and clusters of data uncovered by data mining can lead to new discoveries (Obenshain, 2004).

Table 1 presents an overview of the most important data mining tasks, which can be classified into two major categories, i.e. prediction and description. Indeed, classification and estimation tasks can be seen as prediction tasks: the user wants to predict the (discrete or continuous) value of an unknown attribute. On the other hand, affinity grouping and clustering can be seen as description tasks: it is the objective of the analyst to gain insight into the underlying relationships that exist between attributes or instances in the database. However, these two categories are not mutually exclusive in the sense that some techniques can be used for both purposes.

Table 1: Data mining techniques
The choice for one data mining technique or another is not always obvious. In fact this choice is largely dependent on the situation. Each technique has its strengths and weaknesses in terms of representation language, classification power, descriptive abilities and expert knowledge required (Geurts, 2006).

Finally, data mining attacks problems such as obtaining efficient summaries of large amounts of data, and using a set of previously observed data to construct predictors of future observations (Friedman, 1997). Statisticians have well established techniques for attacking all of these problems as well. However, some distinct differences can be identified. Not only can datasets be much larger than in statistics, data mining analysis is also well fitted to the analysis of a vast amount of data where traditional statistical analysis based on the ‘hypothesis and test’ paradigm becomes a time consuming process (Duhamel et al., 2001). Furthermore, data mining is typically secondary data analysis and techniques can be implemented retrospectively on massive data in an automated matter, whereas traditional statistical methods used in epidemiology require custom work by experts. Additionally, traditional statistical methods generally require a certain number of predefined variables, whereas data mining can include new variables and accommodate a greater number of variables (Obenshain, 2004).

In organizational applications, the use of knowledge discovery and data mining technologies is driven by the concept that knowing things that nobody else knows bring success; this is true for both business and healthcare organizations (Kunene and Weistrofferb, 2006).

### 2.2. Data mining and healthcare data

The ‘To Err is Human’ report published by the Institute of Medicine (IOM) in 1999 called for a national effort to make healthcare more safe. Consequently, the literature on patient safety has dramatically increased since the IOM report. Furthermore, the subjects of interest have moved from malpractice to fields of human factors engineering, psychology, and informatics (Stelfox et al., 2006).

Recently, the term data mining has also been increasingly used in the medical literature. Especially, many predictive data mining methods have been successfully applied to a variety of practical problems in medicine (Belazzi and Zupan, 2006). The goal of predictive data mining in clinical medicine is to derive models that can use patient specific information to predict the outcome of interest and to thereby support clinical decision-making.
As explained in the previous section, data mining can be defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from large amounts of data (Frawley et al., 1991). Accordingly, data mining and data warehousing are becoming more prevalent in the healthcare industry because of the large quantities of data stored in various systems at medical institutions and the number of business decisions made based on the data (Herr, 1996; Borok, 1997; Scheese, 1998; Silver et al., 2001). According to Peter Bath (2004), in health/medical care, data are routinely generated and stored as part of the care process, for administrative purposes, or for research (Coiera, 1997; Peiia-Reyes and Sipper, 2000; Shortliffe and Blois, 2001). A single healthcare episode or research study may yield hundreds of variables and generate large amounts of data. Even though individual data items may be of little value in their own right, valuable information may be contained among them that is not immediately apparent, but that may be extracted and utilized using data mining (Kuo et al., 2001). For an appropriately formulated scientific question, thousands of data-elements can be brought to bear on finding a solution. For an appropriately formulated medical question, finding an answer could mean extending a life, or giving comfort to an ill person (Cios and Moore, 2002).

Accordingly, the availability of health/medical data and information, coupled with the need to increase our knowledge and understanding of the biological, biochemical, pathological, psychosocial, and environmental processes by which health and disease are mediated, mean that medicine/health is particularly suitable for data mining (Shortliffe and Barnett, 2001; Shortliffe and Blois, 2001).

However, the process of data analysis in healthcare is becoming more and more complicated for a number of reasons. Most importantly, medicine and health deal with complex organisms (humans/patients) and with higher-level processes than other branches of science, such as physics and chemistry (Shortliffe and Blois, 2001). Therefore, data mining in medicine is distinct from that in other fields, because the data are heterogeneous. Furthermore, special ethical, legal, and social constraints apply to private medical information and analysis tools must address these heterogeneity and social issues. Finally, medicine itself has a special status in life. However, for all its perils, medical data mining can also be the most rewarding (Cios and Moore, 2002).

Traditional statistical methods, such as statistical process control on various underlying probability distribution functions, have also been successfully implemented in hospital infection control (Birnbaum, 1984; Kaminsky et al., 1992; Finison et al., 1993; Sellick, 1993; Ngo et al., 1996; Benneyan, 1998; Gustafson, 2000; Benneyan, 2001; Obenshain, 2004). Furthermore, in the context of Bayesian disease mapping, recent literature presents the development of Bayesian statistical methodologies that make it possible to study associations between health problems and risk factors (MacNab, 2006; MacNab and Gustafson, 2007). Also, retrospective analysis of hospital discharge data has been used with success to discover the rate of adverse outcomes in Belgian hospitals (Van den Heede et al., 2006). However, direct comparison of traditional statistical methods with datamining would require competitive results on the same data. Both methods can be used to analyse medical data, however, as explained in section 2.1, the problems and methods of data mining have some distinct features of their own.

Although a number of data mining methods have been successfully applied to healthcare data (e.g. predictive modelling in clinical medicine (Belazzi and Zupan, 2006); predicting health outcomes (Chae, 2001; Kunene and Weistrofferb, 2006); modelling clinical costs and predictions (Riano and Prado, 2000), very few papers have studied how data mining can be used to improve patient safety or search for patient safety indicators using patient discharge data. However, some research on data mining techniques that are used to identify patterns in medical data to improve healthcare quality and not to predict outcomes can be found (e.g. the detection of quality relevant differences in care patterns (Stühlinger et al., 2000), discovering time dependency patterns in clinical pathways (Lin et al., 1999), the use of text mining to list patient risk factors and complications (Cerrito and Cox, 2003). Furthermore, data mining has already proven to be an interesting technique for the identification of variables and patterns in large datasets in other
fields such as traffic safety (identifying accident circumstances that frequently occur together (Geurts, 2006)) or marketing (identifying consumers behaviour (Brijs, 2002)).

In this sense, data mining can contribute with important benefits to the health sector, as a fundamental tool to analyse the data gathered by hospital information systems and obtain models and patterns which can improve patient assistance and a better use of resources and pharmaceutical expenses (Riaño and Prado, 2000; Riaño and Prado, 2001; Goodwin et al., 2003; Orallo et al., 2004; Alapont et al., 2005). Furthermore, it has been reported in medical fields that a combination of machine learning methods with statistical models in particular is very useful. More specifically, machine learning methods are useful for extracting ‘hypotheses’ which may not be significant from the viewpoint of statistics. After deriving these hypotheses, statistical analysis is used for its validity. Accordingly, it is expected that a combination of these two methodologies will play an important role in medical decision support, such as intra-hospital infection control, detection of risk factors (Tsumoto and Hirano, 2004).

3. Data source

In Belgium, the collection of hospital discharge data has been compulsory since 1990 for all in-patients in all acute hospitals (Van den Heede et al., 2006). Furthermore, this Belgian Hospital Discharge dataset was commissioned by the Belgian Ministry of Public Health via the Royal Decree of 6 December 1994. This same decree directed the appointment of a commission to control the content and format of patient registration, the data collecting procedures, and the completeness, validity, and reliability of the collected data. More specifically, the quality of the data is audited by the Ministry of Public Health in two ways. Firstly, a software program checks the data for missing, illogical, and outlier values. Secondly, by regular hospital visits, a random selection of patient records is reviewed to ensure that data were recoded correctly (Vander Meersch, 2002; Van den Heede et al., 2006). The Belgian Hospital Discharge dataset contains information about the patient, the hospital stay, the specialism, diagnoses and procedures. These diagnoses and procedures are registered according to the international code; ICD-9. These ICD-9 or ‘International Classification of Diseases, 9th version’ contains a list of injuries, developed under sponsorship of the World Health Organization (eICD, 2007).

Because of the compulsory characteristic of the dataset, the amount of information on hospital stays is growing every day. The result of the latest feedback (based on statistical analyses of the hospitals stays supervised by the Belgium Federal Government of 2004, counts 1,526,872 registered hospital stays. This implicates a medical cost of 5,716,897,898.17 euro for the year 2004 (this figure includes nursing costs, medications, doctor consultations, clinical tests and other costs). In the year 2003, 1,532,567 hospitalized patients were registered and in 2002 1,525,454 hospital stays. Accordingly, an increase of approximately 1,525,000 records per year can be assumed (Technical Cel, 2007).

These patient discharge data contain a rich source of information on the different diagnoses and treatments during the patient’s hospital stay. In particular, the patient discharge data contain the information on:
- the main diagnoses (reason of the hospital stay; diagnoses ICD-9-CM, type of hospital stay, …),
- the medical treatment (procedure ICD-9-CM, type of discharge, …),
- possible side diagnoses (diagnoses ICD-9-CM, number of days between 1st & 2nd hospital stay,…), which indicates an adverse events (injuries caused by medical management rather than by underlying disease or condition of the patient) caused by the medical action and the needed extra care (Lammar, 2004).

More specifically, in the Belgian hospital data registration system, after medical examination, diagnoses are registered by specialism or pathology using he ICD-9-CM codes. This allows to define a primary diagnoses, which was indicated after research and which is the cause of the
prerecording of the patient in the hospital. Accordingly, the table ‘DIAGNOSE’ (table 2) contains at least one record ‘primary diagnoses’ per stay in the specific pathology.

Secondary diagnoses are defined as disorders which already exist at the moment of the prerecording or which are developed during the hospitalization of the patient (pre-recorded diagnoses, complications and adverse events). These diagnoses have an influence on the current care or treatment of the main disorder of the patient. If the patient has suffered an adverse event, it will be defined in the ‘Diagnose Table’ as one of these secondary diagnoses.

Taking into account the increase of approximately 1,525,000 records a year and the possibility of 14,000 different ICD-9 codes, this results in a large and complex database.

Table 2: Diagnose table (Lammar, 2004)

<table>
<thead>
<tr>
<th>#</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field 1</td>
<td>CIV number</td>
</tr>
<tr>
<td>Field 2</td>
<td>Registration year</td>
</tr>
<tr>
<td>Field 3</td>
<td>Registration semester</td>
</tr>
<tr>
<td>Field 4</td>
<td>Hospital stay number</td>
</tr>
<tr>
<td>Field 5</td>
<td>Record number specialism</td>
</tr>
<tr>
<td>Field 6</td>
<td>Diagnoses and info emergency</td>
</tr>
<tr>
<td>Field 7</td>
<td>Code primary-/secondary diagnoses and info emergency</td>
</tr>
<tr>
<td>Field 8</td>
<td>Degree of certainty</td>
</tr>
</tbody>
</table>

Finally, the Belgian MKG dataset contains patient hospital discharge data. This medical data has its own characteristics and limitations. First of all, patient discharge data is very heterogeneous, voluminous and is submitted to ethical, legal, and social constraints. Additionally, medicine has a special status in science, philosophy, and daily life. The outcomes of medical care are life-or-death, and they apply to everybody. Indeed, medicine is a necessity, not merely an optional luxury, pleasure, or convenience. Furthermore, medical care is sometimes risky, but when it fails, the desire for legal revenge is intense and punitive. Last but not least, medical information about the individual patient is considered highly private, and the general public is extremely fearful about disclosure. We all enjoy the benefits of medical research conducted on other patients, but we are very often reluctant to contribute or release our own information for such purposes. Accordingly, when medical data are published, it is expected that the researchers will maintain the dignity of the individual patient, and that the results will be used for socially beneficial purposes (Saul, 2000; Cios and Moore, 2002).

4. Implications and limitations

Based on the results from the literature review, we can conclude that data mining can contribute with important benefits to the healthcare sector, as a fundamental tool to analyse data gathered by hospital information systems and obtain models and patterns which can improve patient safety, efficient resources allocation and control of expenses.

Furthermore, close examination of the Belgian Hospital Discharge dataset shows that there is a large amount of information available on patients and the patient’s hospital stay. The datasets are increasing by approximately 1,525,000 records a year and contain information about the patient, the hospital stay, the specialism, diagnoses and procedures. Therefore, we believe that the use of data mining techniques on the Belgian Hospital Discharge dataset can lead to a potential improvement of resource utilization and patient healthcare quality, and a decrease of adverse events in Belgium hospitals.
More specifically, the data mining technique of generating association rules can be used to discover associations among variables in the database. These associations are often referred to as correlations between variables. In the context of patient safety analysis, the association algorithm is therefore able to identify all the adverse event circumstances that frequently occur together. However, the use of this technique is of an explorative character since this algorithm cannot give any explanation about the causality of these adverse event patterns. Therefore, its role is to give direction to more profound research on the causes of these patterns. Furthermore, the use of some additional techniques and expert knowledge will be essential to explain and interpret the interestingness of the results. Finally, when analysing the Belgian Hospital Discharge database we need to take into account the constraints and difficulties of mining privacy-sensitive, heterogeneous data of medicine.

5. Conclusions and further research

To conclude, future research will evolve around the improvement of methods for adverse event analysis, more specifically by the use of data mining techniques. However, data mining is still below its full potential in many areas. Healthcare, especially public healthcare, is one of these areas. Accordingly, in literature, only few research examples exist, using data mining to analyse patients discharge datasets. In this paper, we have analysed the opportunities and difficulties of data mining techniques on medical data, to improve the medical quality and to support hospital management.

The Belgian Hospital Discharge dataset is a voluminous dataset (and increasing every year) populated by all Belgian acute hospitals since 1990. However, these data are not actively used to improve patient safety. These data are only used to perform simple statistical analysis for example to give financial feedback per pathology per hospital, national financial overviews, reference financial figures, ... Furthermore, it is not possible to use traditional statically methods on such a large dataset, to find possible patterns, dependencies and relations between variables to define and therefore prevent adverse events. In this context, we have illustrated the promising possibilities of the use of data mining techniques, to analyse the large dataset of patients discharge data that is currently available.

Therefore, in our future research, we will examine the Belgian Hospital Discharge data using data mining techniques to identify injuries caused by adverse events, patient safety risks, and in the end to define proposals to improve healthcare quality. Accordingly, detailed analyses of the Belgian Hospital Discharge data datasets could give a complete picture of the Belgian Healthcare situation, landscape risk indicators, help to define patient safety precautions and ultimately lead to efficient hospital management. Therefore, the success of our future research using data mining on patient discharge data could be useful for many hospitals which cannot afford a complete data mining programme of their own. Finally, we believe that the combination of data mining and large medical datasets can lead to better and safer healthcare.

Ultimately, it is clear that the cross-fertilization between two fascinating research domains, i.e. data mining and patient safety is challenging but results indicate promising outcomes.
References

Geurts, K. (2006), "Ranking and profiling dangerous accident locations using data mining and statistical techniques", Hasselt University, Belgium.


