



## Overviews of Data Mining in Hospital Information System.

**Pattanapruteep O, Ph.D. candidate<sup>1</sup>, Pongcharoensuk P, Ph.D.<sup>1</sup>, Kaojarern S, MD.<sup>2</sup>**

<sup>1</sup> *Department of Pharmacy, Faculty of Pharmacy, Mahidol University, Bangkok, Thailand*

<sup>2</sup> *Department of Medicine, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand*

### Abstract

---

Hospital information system captures huge data in each hospital database, but currently few knowledge is produced because of its complexity. Data mining has great potential for exploring the hidden patterns in complex data sets of the hospital information system. Introduction to data mining will provide hospital staff understand how data mining discovers and extracts useful patterns from this large data.

**Keywords:** Data mining, Hospital information system, Knowledge discovery in database

---

## Introduction

In recent years, the increasing volume of health data and growing of quality demand in hospital make it necessary to explore the potential of stored data efficiently, not only the clinical data, in order to improve diagnoses and treatments, but also on management, in order to minimize costs and improve the care given to the patients. Data mining (DM), which is an integral part of knowledge discovery in databases (KDD), can provide important benefits to the health sector, as a fundamental tool to analyse the data gathered by hospital information systems (HIS) and obtain models and patterns which can improve patient assistance and a better use of resources and pharmaceutical expenses.

DM or data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of dataset. It is the process of automatically discovering useful information in large data repositories.<sup>(1)</sup> The term is often used to apply to two separate tasks of description and prediction. The objective of descriptive task is to derive patterns - for example trends, clusters, anomalies - that summarize the underlying relationships among the data. While the objective of predictive task is to predict the value of a particular attribute based on the values of other attributes- such as decision trees and neural networks. Traditional data analysis use standard query tools to answer questions such as, "How many patients in hospital readmits within 24 hours?" DM is valuable for more complicated queries such as, "What are the important predictors of 24 hours readmission?"

The paper overviews DM concepts and techniques, in order to introduce its developmental history and how it extracts useful knowledge from large hospital database system. In addition, a number of applications of data mining in health care system are presented.

## Developmental history of DM

DM is a concept emerged in the late 1980s. But it soon attracted huge interests for research works with many new and remarkable techniques being discovered throughout the 1990s. DM has evolved from a number of different disciplines such as statistics, machine learning, artificial intelligence, and database technologies.<sup>(1)</sup> The concept of DM is difference from data warehousing which process is used to integrate and combine data from multiple sources and format into a single unified schema. Data warehouse provides the enterprise with a storage mechanism for its huge amount of data and develops an architectural model for the flow of data from operational systems to decision support environments. On the other hand, DM is the process of extracting interesting patterns and knowledge from huge amount of data. However, DM techniques can be applied on the data warehouse of an enterprise to discover useful patterns.

DM is also different from statistics although statistics is required in several DM techniques. Statistical inference is assumption driven in the sense that a hypothesis is formed and tested against data. DM, in contrast, is discovery driven. That is, the hypothesis is automatically extracted from the given data.

The use of DM is in a variety of purposes in both health and other sectors. Industries such as medicine, banking, insurance, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical community sometimes uses data mining to help predict the



effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together.<sup>(2,3)</sup>

### Process of DM

DM can be considered as one step in KDD. Other steps in the KDD process, in progressive order, include data integration, data selection, data preprocessing or cleaning, data transformation, data mining, pattern evaluation, and knowledge presentation.<sup>(4)</sup> Various steps that are involved in mining data are shown in figure 1.

1) Data integration. First of all, the data are collected and integrated from all different sources.

2) Data selection. Since all the collected data in the first step may not be needed, in this step, only

those data which are useful for data mining will be selected.

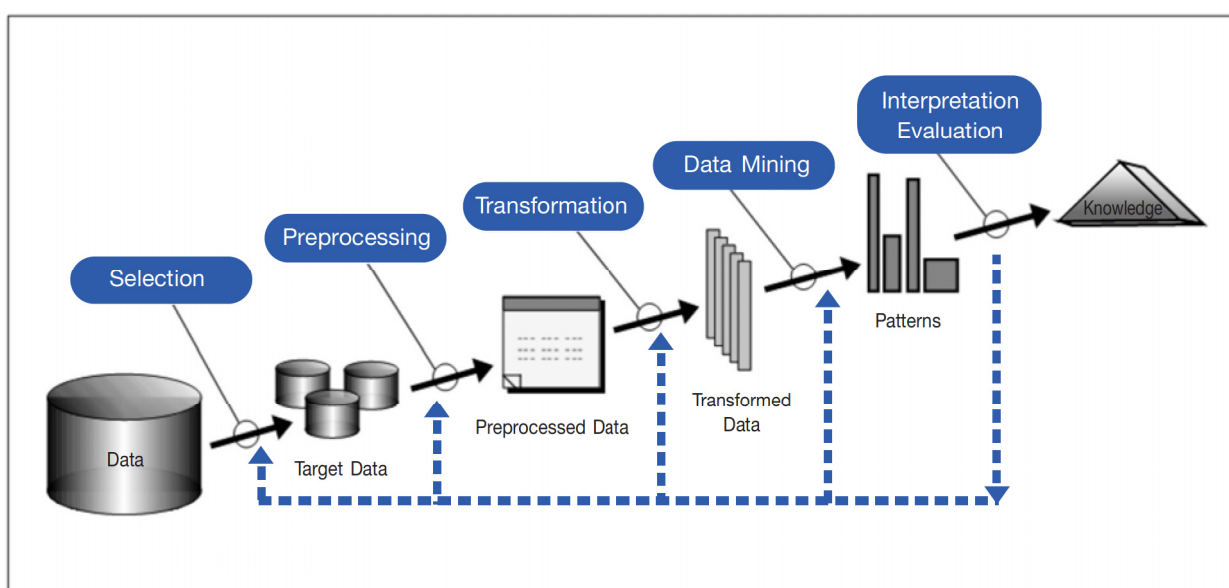
3) Data preprocessing or data cleaning. Unclean data may contain errors, missing values, noisy or inconsistent data; so different techniques are applied to get rid of such anomalies.

4) Data transformation. The data, even after cleaning are not ready for mining as may need to be transformed into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.

5) Data mining. The data are then ready to be applied data mining techniques, to discover the interesting patterns. Techniques like clustering and association analysis are among the many different techniques used for data mining.

6) Pattern evaluation. This step involves visualization, transformation, removing redundant patterns etc. from the generated patterns.

7) Knowledge presentation. Finally, this step helps user make use of the knowledge acquired to take better decisions.



**Fig.1** An Overview of the Steps That Compose the KDD Process.

## DM techniques

In the DM process, many techniques are used for either descriptive or predictive task. The four core techniques<sup>(1,5)</sup> are: associations, mining frequent patterns; anomaly detection; classification and prediction; and clustering.

### Associations, mining frequent patterns

These methods are to discover patterns that describe strongly associated features in the data. Useful applications of association analysis include finding groups of genes that have related functionality, market basket analysis, etc.

### Anomaly detection

Anomaly detection is the task of identifying observations whose characteristics are significantly different from the rest of data. Such observations are known as anomalies or outliers. Applications of anomaly detection include fraud detection, unusual patterns of disease, etc.

### Classification and prediction

The classification and prediction models are two data analysis techniques used to describe data classes and predict future data classes which may be either discrete or continuous variable. The goal is to learn a model that minimizes the error between the predicted and true values of data variables. Some techniques are decision trees and neural networks, for example, to judge whether a patient has a particular disease based on the results of medical tests.

### Clustering

Cluster analysis seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters. Clustering has been used to identify group of similarities, characterize patient groups based on pattern of disease, payer, demographic data, etc.

## Applications of DM in hospital information system (HIS)

Many health care organizations worldwide have applied DM to derive knowledge from HIS. Some of which are described below.

### Clinical data

The applications of DM on clinical data are, detecting adverse events, identifying high risk patients and improving their outcomes, quality improvement, and building nursing knowledge.<sup>(6)</sup> In developed countries, the Food and Drug Administration applies neural network - a useful DM techniques - to collect adverse drug event data in order to detect when too many adverse events occur or detect link between drug and adverse event.<sup>(5)</sup> Too many adverse events linked to a drug might indicate drug is too dangerous or health of patient is at risk.

DM is also implemented to identify high risk patients and consequently, improve their outcome. For example, in diabetic patients, classification technique on medical data can early detect patients' complication from their condition attributes.<sup>(7)</sup> American Healthways uses Enterprise miner, a predictive modeling technology from SAS, to identify high risk diabetic patients to improve their quality of health care and prevent future health problems.<sup>(8)</sup>

To obtain interesting patterns in infection control data, data mining surveillance system is implemented using association rules on culture and patient care data. Early recognition of outbreaks and emerging resistance required proactive surveillance then are detected.<sup>(9,10)</sup>

### Financial data

To control healthcare cost and improve use of resources for more efficiently, many health insurance companies and hospitals implement DM applications to detecting health care fraud/abuse and predicting hospital charge.<sup>(11)</sup> Fraud detection system can automatically execute overnight, assigning any fraud





probabilities to each medical claim.<sup>(12,13)</sup> A case study in Chile in 2003 revealed that approximately 75 fraudulent and abusive cases per month could be detected from 800 daily medical claims by a private health insurance company, making the detection 6.6 months earlier than without the system.

### Products of DM related software

Because of growing needs on DM applications, many companies have developed DM products to the market - for example, Enterprise Miner by SAS, Kate by Acknosoft, Knowledge Seeker by Angoss, Business Miner by Business Objects, Intelligent Miner

by IBM Corporation, Mineset by Silicon Graphics, DataMind DataCrucher by DataMind, and Drawin by Thinking Machines Corporation.

### Conclusion

Data mining is the process of extracting information from large datasets, such as in hospital information system, in order to find useful patterns that can convert to knowledge. An understanding of the knowledge hidden in the hospital information system is vital to hospital management team and staff to lowering health care cost while improving the quality of patient's care.

### References

1. Tan P, Steinbach M, Kumar V. Introduction to data mining. Pearson International edition. MA: Boston; 2006: 1-6.
2. Edelstein HA. Introduction to data mining and knowledge discovery. 3<sup>rd</sup> ed. Two Crows Corporation. MD: Potomac; 1999: 5.
3. Dillon PM. Data mining: transforming business data into competitive advantage and intellectual capital. The information management forum. GA: Atlanta; 1998:5-6.
4. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine 1996; Fall:37-54.
5. Ranjan J. Applications of data mining techniques in pharmaceutical industry. J Theoretical Appl Inform Technol 2007;3:61-7.
6. Goodwin L, VanDyne M, Lin S, Talbert S. Data mining issues and opportunities for building nursing knowledge. J Biomed Inform 2003;36:379-88.
7. Kaur H, Wasan SK. Empirical study on applications of data mining techniques in healthcare. J Comput Sci 2006;2:194-200.
8. Ridinger M. American healthways uses SAS to improve patient care. DM review 2002;12:139.
9. Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004;25:690-5.
10. Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc 1998;5:373-81.
11. Lee SM, Kang JO, Suh YM. Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs. decision tree models. J Korean Med Sci 2004;19:677-81.
12. Ortega PA, Figueroa CJ, Ruz GA. A medical claim fraud/abuse detection system based on data mining: a case study in Chile. Presented at the 2006 international Conference on Data Mining (DMIN); June 26-29, 2006; Las Vegas, Nevada, USA.
13. Yang WS, Hwang SY. A process mining framework for the detection of healthcare fraud and abuse. Expert Syst Applications 2006;31:56-68.



## Overviews of Data Mining in Hospital Information System.

อรลักษณ์ พัฒนาศรีกุล, พว. candidate<sup>1</sup>, เพชรรัตน์ พงษ์เจริญสุข, Ph.D.<sup>1</sup>,  
สมิง เก้าเจริญ, พว.<sup>2</sup>

<sup>1</sup> Department of Pharmacy, Faculty of Pharmacy, Mahidol University, Bangkok, Thailand

<sup>2</sup> Department of Medicine, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

### บทคัดย่อ

---

การจัดเก็บข้อมูลจำนวนมากในฐานข้อมูลของระบบสารสนเทศโรงพยาบาลที่มีความซับซ้อนในปัจจุบันทำให้การนำข้อมูลมาสร้างเป็นความรู้ทำได้ค่อนข้างยาก การทำเหมืองข้อมูลมีความสามารถในการดึงรูปแบบที่ซ่อนอยู่ในข้อมูลที่มีความซับซ้อนให้เกิดเป็นความรู้ได้อย่างเป็นระบบ ความรู้เบื้องต้นของการทำเหมืองข้อมูลจะส่งเสริมให้ผู้ปฏิบัติงานในโรงพยาบาลมีความเข้าใจในการค้นหาและสร้างรูปแบบที่เป็นประโยชน์จากฐานข้อมูลขนาดใหญ่ได้

---