

Data Mining Applications in the Context of Casemix

H C Koh,*^{PhD, CPA}, S K Leong,**^{MBBS}

Abstract

In October 1999, the Singapore Government introduced casemix-based funding to public hospitals. The casemix approach to health care funding is expected to yield significant benefits, including equity and rationality in financing health care, the use of comparative casemix data for quality improvement activities, and the provision of information that enables hospitals to understand their cost behaviour and reinforces the drive for more cost-efficient services. However, there is some concern about the “quicker and sicker” syndrome (that is, the rapid discharge of patients with little regard for the quality of outcome). As it is likely that consequences of premature discharges will be reflected in the readmission data, an analysis of possible systematic patterns in readmission data can provide useful insight into the “quicker and sicker” syndrome. This paper explores potential data mining applications in the context of casemix by using readmission data as an illustration. In particular, it illustrates how data mining can be used to better understand readmission data and to detect systematic patterns, if any. From a technical perspective, data mining (which is capable of analysing complex non-linear and interaction relationships) supplements and complements traditional statistical methods in data analysis. From an applications perspective, data mining provides the technology and methodology to analyse mass volume of data to detect hidden patterns in data. Using readmission data as an illustrative data mining application, this paper explores potential data mining applications in the general casemix context.

Ann Acad Med Singapore 2001; 30(Suppl):41-9

Key words: Casemix-based funding, Enterprise Miner, Premature discharges, Predictive modelling, Readmission data

Introduction

By classifying similar episodes of care into useful and meaningful groups, casemix classification provides certain advantages. Firstly, casemix classification possesses clinical meaning because patients in the same classification have similar clinical conditions. Secondly, following from above, casemix classification indicates similar resource use as patients in the same classification generally cost about the same amount to treat. The primary desired consequence of casemix-based funding is that hospitals will have the incentive to ensure efficient and effective use of the available resources. Casemix-based funding also helps ensure that available funds are distributed more fairly and rationally.

In addition, casemix information can help hospitals understand better the cost of providing hospital care. With better understanding, care delivery can be made more efficient and effective. For example, casemix information on patient care such as the average length of stay can help identify variations and lead to clinical guidelines and best practices. Casemix information can also be used for

comparative purposes—for example, to compare among hospitals and among doctors. This can facilitate the setting of benchmarks.

Although casemix-based funding has several important advantages, it has also led to some concerns. One concern is that DRGs (Diagnosis Related Groups) may not adequately consider differences in therapeutic philosophies, patient needs and technological developments. Another concern is that patients may be denied access to medical care or that they may be discharged prematurely. There is anecdotal evidence of premature hospital discharges,¹ also referred to as the “quicker and sicker” syndrome (that is, the rapid discharge of patients with little regard for the quality of outcome). Empirical evidence has also been reported that casemix accounting systems have affected resource allocation and changed working practices, including closing operating theatres, cutting staffing levels and reducing patients’ length of stay.²

As a strong relationship between premature discharges and readmissions can be expected, this paper focuses on

* Associate Professor and Vice-dean
Nanyang Business School (NTU)

** Deputy Director (Casemix)
Tan Tock Seng Hospital

Address for Reprints: Associate Professor Koh Hian Chye, Vice-dean, Nanyang Business School (01a-32), Nanyang Technological University, Nanyang Avenue, Singapore 639798.
E-mail: ahckoh@ntu.edu.sg

the analysis of readmission data. In particular, it explores potential data mining applications for readmission data in particular and casemix data in general. The objective of this paper is to explore data mining applications in the context of casemix and not to report the results of a data mining exercise. It should also be noted that while casemix implementation has given rise to the importance of evaluating readmission data, in any output based or managed care funded environment, providers and regulators would be interested in the issue.

The remainder of the paper is organised into the following three sections. The first section discusses the data mining methodology and tools, and the data mining literature. The second section illustrates how data mining can be applied to analyse readmission data. It also suggests other potential data mining applications using casemix data. Finally, the concluding section highlights the limitations of data mining.

Overview of Data Mining

Data mining can be defined as the process of selecting, exploring and modelling large amounts of data to uncover previously unknown patterns of data.³ The SAS data mining methodology comprises the following five stages: Sample, Explore, Modify, Model and Assess (i.e., SEMMA). Nazem and Shin⁴ described the SEMMA methodology as an effective search methodology and solution framework to guide data mining. Data mining tools can be broadly classified based on what they can do, namely:

- 1) description and visualisation;
- 2) association and clustering; and
- 3) classification and estimation (i.e., prediction).

Description and Visualisation

Description and visualisation can contribute greatly towards understanding a data set (especially a large one) and detecting hidden patterns in data (especially complicated data containing complex and non-linear interactions). They are usually performed before modelling is attempted and represents exploration in the SEMMA methodology. Standard description tools include summary statistics (e.g., measures of central tendency and measures of dispersion) and graphical representations (e.g., distributions and plots). Visualisation can be considered an enhanced graphical approach that allows user input and interaction. An example is a rotating multidimensional plot that permits the user to define the multiple dimensions (multiple variables) in the plot as well as the direction and angle of rotation to facilitate viewing complex relationships. Colours can also enhance visualisation tools. In the data mining context, description and visualisation tools can be used to understand people, products and process and study the relationships among variables. The results from such analyses are seldom

an end in themselves but are usually used as a means to construct data mining models (e.g., to predict certain target variables).

Association and Clustering

In association, the objective is to determine which variables go together and in what order. For example, market basket analysis refers to a technique that generates probabilistic statements such as: if a patient has Complication A, there is a 0.35 probability that he/she will also have Complication B. Market basket analysis can be applied not only to events occurring concurrently but also to events occurring sequentially. Another data mining tool, link analysis, can also be considered an association technique. It looks at connection relationships—that is, how people, places and things are connected (e.g., call patterns in telecommunications).

In clustering, the objective is to group objects in such a way that objects belonging to the same cluster are similar and objects belonging to different clusters are dissimilar. The two most common data mining tools for clustering are cluster analysis and self-organising map (or Kohonen network). As an application, clustering can be used to segment patients into homogeneous groups.

Classification and Estimation (Prediction)

The most common and important applications in data mining probably involve prediction (sometimes referred to as modelling). Classification refers to the prediction of a target variable that is categorical in nature (e.g., predicting readmission versus non-readmission). Estimation, on the other hand, refers to the prediction of a target variable that is continuous in nature (e.g., predicting the length of stay).

To construct predictive models, at least one of the following data mining tools are usually used: multiple regression (including logistic regression), neural network and decision tree. While multiple regression is a traditional statistical method, neural networks are derived from work in the artificial intelligence area. They are useful for recognising patterns in data and are modelled after the human brain, which can be perceived as a highly connected network of neurons (called nodes in neural network terminology).

The objective of decision trees is estimation and/or classification by dividing observations into mutually exclusive and exhaustive subgroups. The end product can be graphically represented by a tree-like structure. In applying the decision tree methodology, each observation is eventually assigned to a node (also called leaf) that has a predicted value or classification. Of the three predictive models, decision trees are the most

interpretable in that they can be translated into decision rules. Further, as in the case of neural networks, decision trees can be used to model complex non-linear and interaction relationships.

The Data Mining Literature

The late 1990s have seen an explosion of data mining literature. The professional literature is quite liberal in reporting success stories of data mining applications and recommending data mining for business advantage.^{5,6} For example, the Farmers Insurance Group data-mines customer information to develop competitive rates, Foote Cone & Belding analyses data mined from operational and transactional systems to refine clients' direct mailings and advertising campaigns to improve catalogue sales, and Axios Data Analysis Systems use data mining to help identify what could be fraudulent health insurance claims for one of its clients. Other successful users of data mining include Fingerhut, American Century Investments, Charles Schwab & Company, Chase Manhattan Bank, Bank of America, US West, Bell Atlantic, Alltel, Wal-Mart and Boots PLC.⁷⁻¹⁰

Despite its obvious commercial value, data mining can also be viewed from a decision support system perspective beyond mere statistical analysis and information technology.¹¹ As a decision support system, data mining has potential application in almost all areas, including the context of casemix.

An Illustrative Data Mining Application in the Context of Casemix: Readmission

To illustrate a possible data mining application in casemix, assume that a local public hospital is interested to analyse readmission data. Assume further that the purpose of such an analysis is to understand the readmission data and detect patterns (if any) to assess the "quicker and sicker" syndrome (i.e., the possibility of premature discharges). For this data mining illustration, readmission data from a local public hospital for the period 1 October 2000 to 31 March 2001 were used. The readmission data set comprised 795 cases or observations.

For simplicity, only the following data items were used:

- 1) readmission clinical department [a_orgfa],
- 2) readmission admit source [fadmsrce],
- 3) previous discharge status [pdstat],
- 4) period of readmission (or length of lapse between discharge and readmission) [diff],
- 5) chronic conditions [chron1],
- 6) gender of patient [sex],
- 7) age of patient [age], and
- 8) race of patient [race].

SAS (Statistical Analysis Software) Enterprise Miner

Version 4.0 (a data mining software) was used in this illustration.¹² The data mining diagram associated with the illustration is given in Figure 1.

Description and Visualisation

As mentioned earlier, description and visualisation can contribute greatly towards understanding a data set and detecting patterns. For description, summary statistics (e.g., frequency distributions and means) of the readmission data can be generated. For example, among the readmission patients, the average age was 62.16 years, 66.16% were males and 77.36% were Chinese ... etc. Also, for visualisation of patterns, a 3-dimensional rotating plot can be generated. Figure 2 shows the readmission clinical department (a_orgfa), age of patient (age) and period of readmission (diff) on the axes. The actual screen is 3-dimensional and in colour, which allow the relationships (or patterns), if any, among the three variables to be analysed. Multiple 3-dimensional rotating plots involving different variable can also be generated to visualise the readmission data.

Clustering

To illustrate clustering, assume that to understand the readmission data better, it is desirable to classify the readmission patients into different homogeneous groups. The clustering node in Enterprise Miner (EM) can be used for this purpose (Fig. 1). Figure 3 shows that the readmission patients can be grouped into six relatively homogeneous groups. EM can also provide information on the characteristics of all the readmission patients (i.e., give an overall picture) based on the variables used for clustering, and the characteristics of each of the clusters or groups (in isolation and in comparison to the overall). This is shown in the right panel of Figure 3 and Figure 4. Detailed analysis of the clusters can also be performed (e.g., by using the Insight node in Enterprise Miner as shown in Figure 1). An illustrative screen is given in Figure 5 for the previous discharge status (pdstat) and chronic conditions (chron1). By studying the box/mosaic plots, the profile of each cluster of readmission patients can be developed. With the different features, clustering results are useful for exploring the readmission data

Predictive Modelling

To illustrate predictive modelling using readmission data, suppose that the local public hospital is interested to study the factors affecting the length of lapse between discharge and readmission (i.e., the "diff" variable discussed earlier). In other words, "diff" or the length of lapse is the target variable in constructing the predictive model. The input variables are:

- 1) readmission clinical department [a_orgfa],

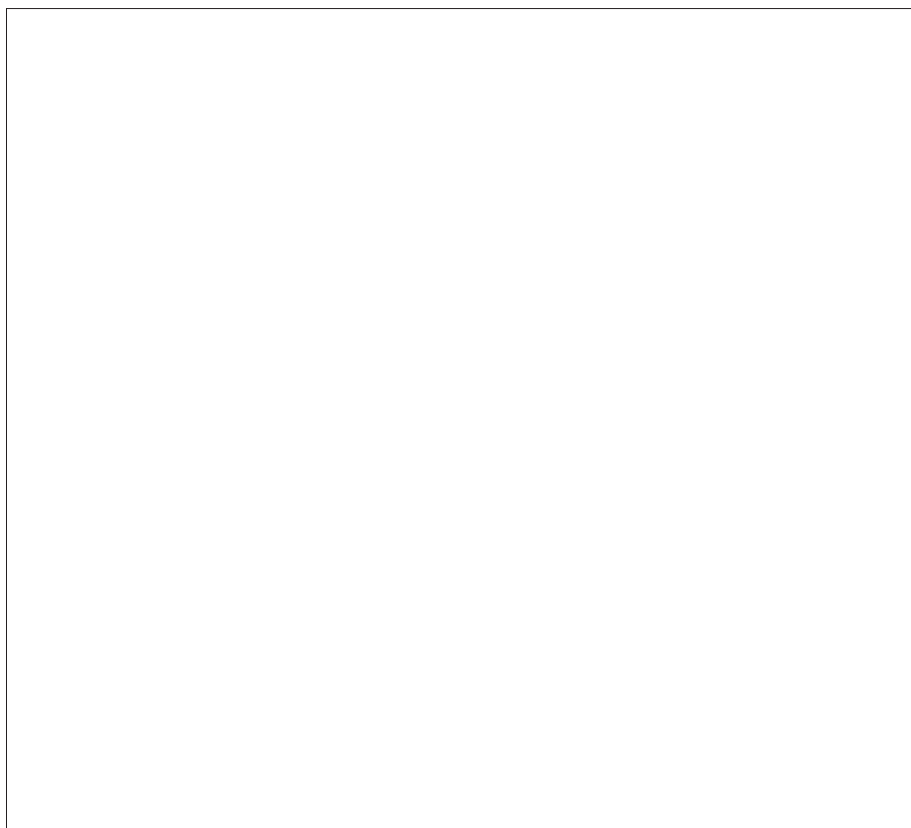


Fig. 1. Data mining diagram.

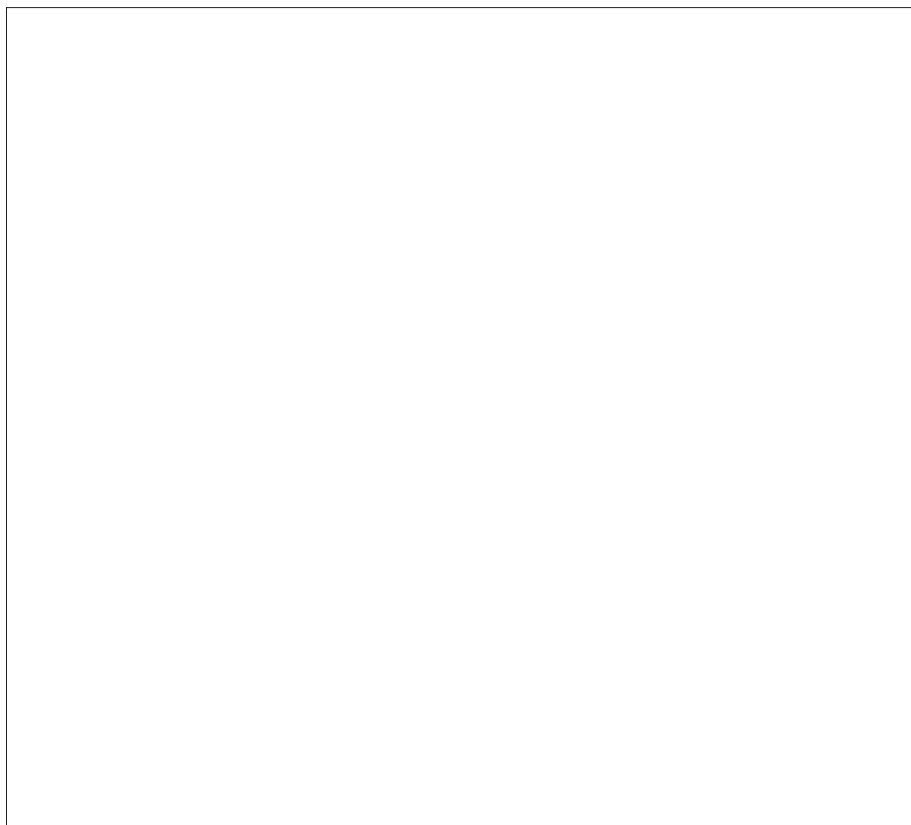


Fig. 2. Three-dimensional rotating plot.



Fig. 3. Clustering results.

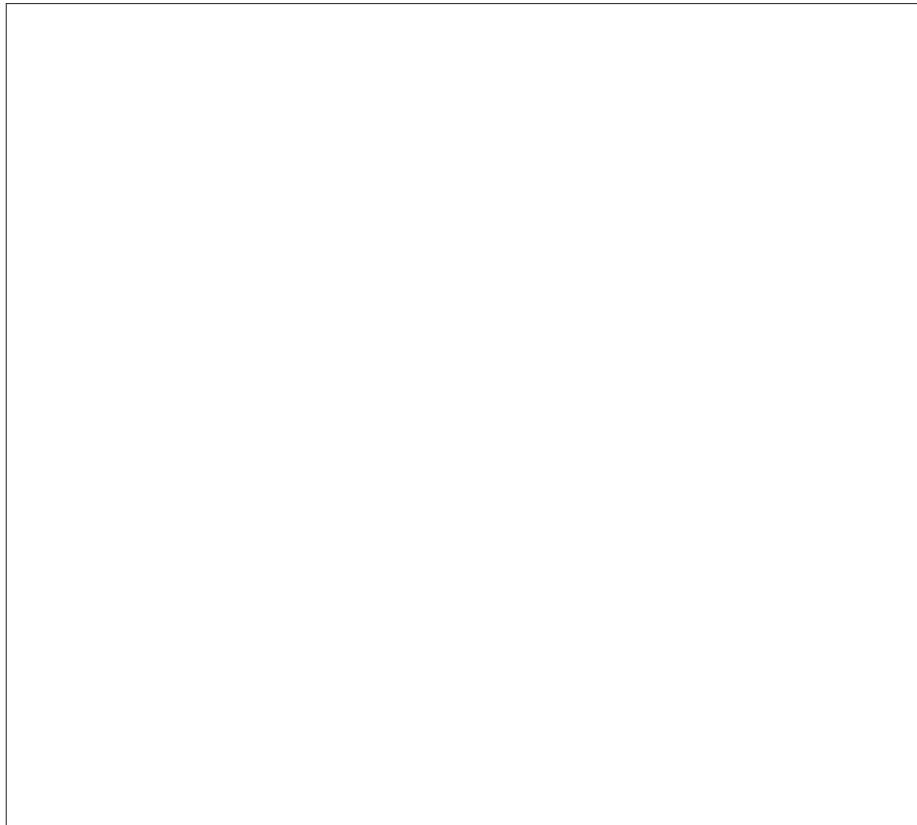


Fig. 4. Profile of clusters.

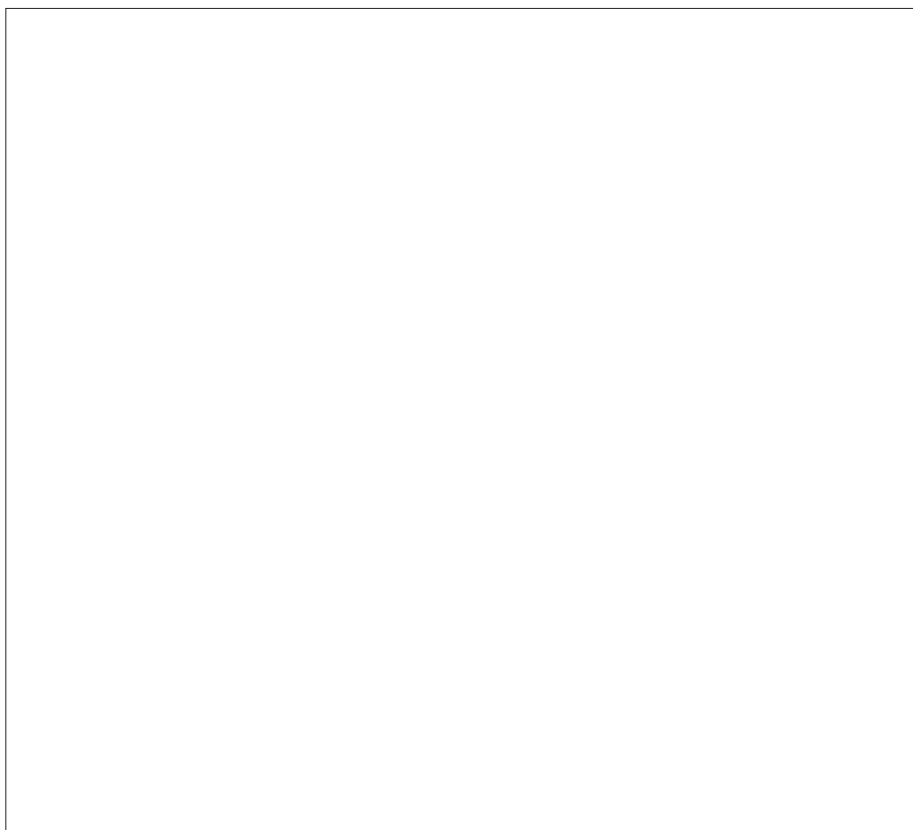


Fig. 5. Detailed analysis of clusters.

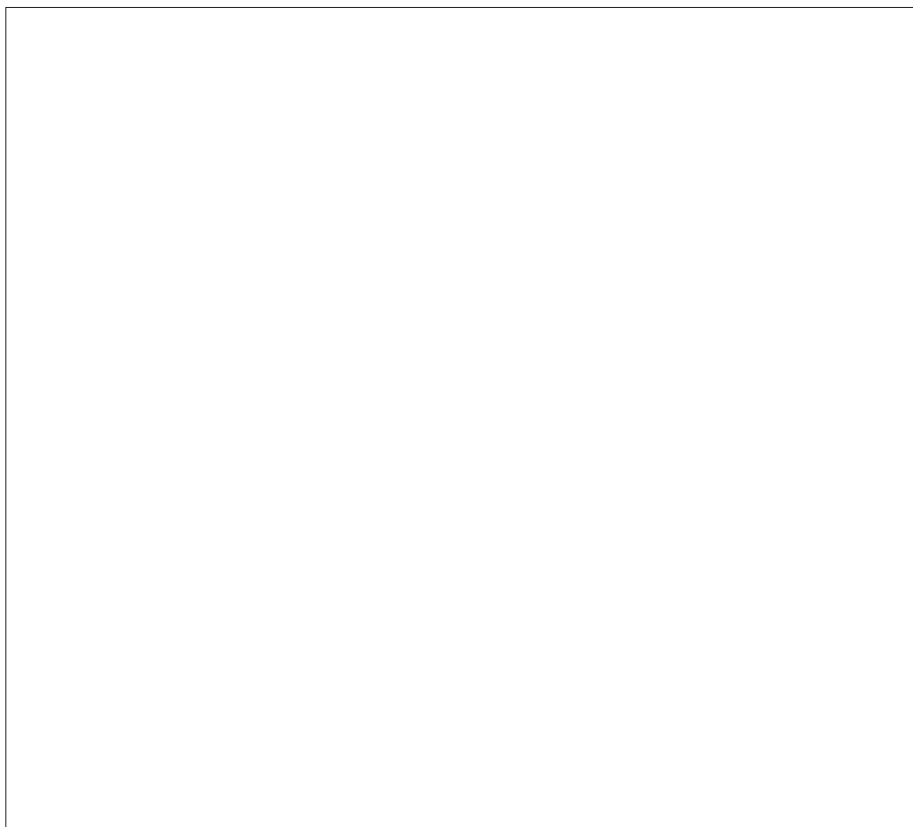
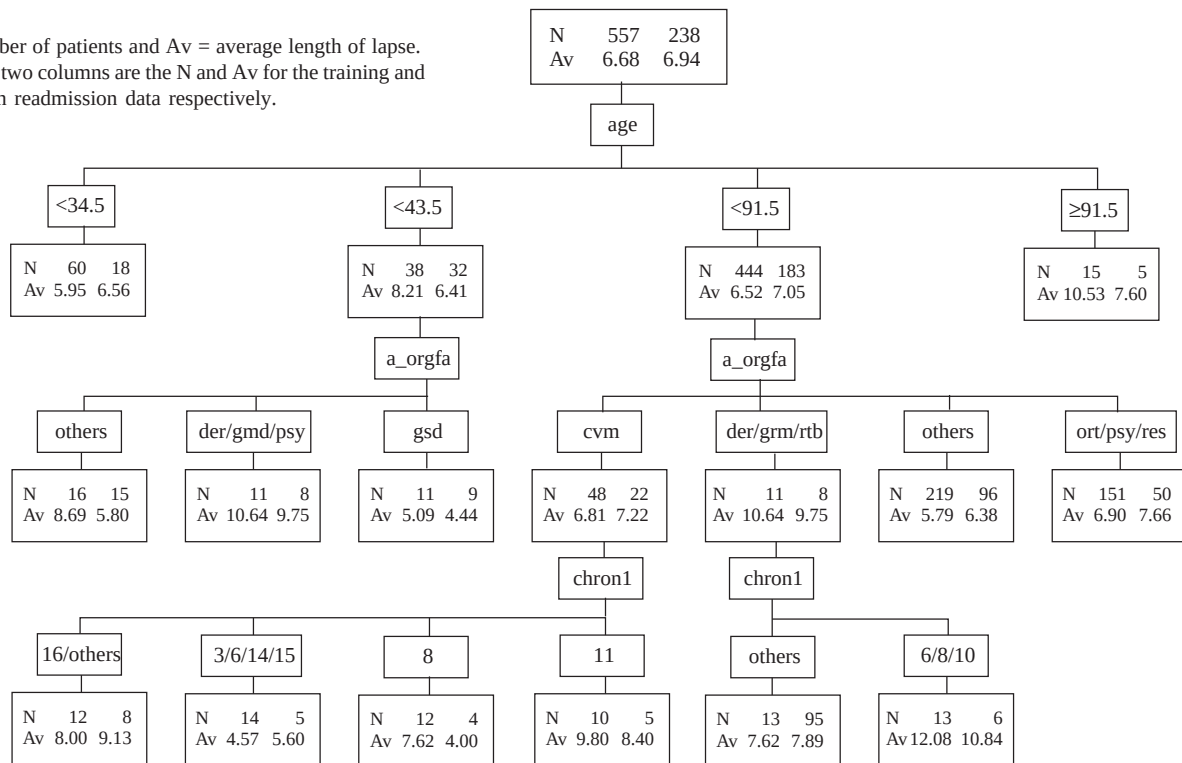


Fig. 6. Multiple regression results.

Note:

N = number of patients and Av = average length of lapse.
The next two columns are the N and Av for the training and validation readmission data respectively.

**Abbreviations:**

Clinical departments: cvm = cardiology
der = dermatology
ent = otorhinolaryngology
eye = ophthalmology
gmd = general medicine
grm = geriatric medicine
gsd = general surgery
ids = infectious diseases
nld = neurology medicine
nsd = neurosurgery
ort = orthopaedic surgery
psy = psychological medicine
rai = rheumatology and immunisation
reh, rca, rcb, rcc, rcd, rce = rehabilitation medicine
res = respiratory medicine
rtb = respiratory medicine/tuberculosis

Chronic conditions:

- 1 = infected gangrene wound
- 2 = bedsores
- 3 = recurrent pain
- 4 = recurrent low back pain
- 5 = advance cancers–recurrent symptoms
- 6 = chest infection
- 7 = urine infection
- 8 = others
- 9 = asthma exacerbation
- 10 = chronic obstructive pulmonary disease (COPD) exacerbation
- 11 = congestive heart failure
- 12 = uncontrolled hypertension
- 13 = uncontrolled diabetic mellitis
- 14 = panic syndrome–atypical chest pain
- 15 = severity of illness e.g., end stage renal failure (ESRF)/end stage heart failure (ESHF) etc.
- 16 = unstable angina

Fig. 7. Decision tree.

- 2) readmission admit source [fadmsrce],
- 3) previous discharge status [pdstat],
- 4) chronic conditions [chron1],
- 5) gender of patient [sex],
- 6) age of patient [age], and
- 7) race of patient [race].

For predictive modelling, three data mining tools are usually appropriate; namely, multiple regression, neural network and decision tree. Given that interpretability is an important requirement in this application, neural networks are not appropriate because their results are often difficult to interpret. Hence, only multiple regression and decision

tree will be used in this illustrative data mining application.

Figure 1 shows the Regression and Decision Tree nodes. Further, the Data Partitioning node is also incorporated to partition the readmission data into a training set to construct the model (70% of the readmission data selected randomly) and a validation set to validate/test the constructed model (30% of the readmission data selected randomly).

Multiple Regression Results

The multiple regression results are captured in Figure 6. As shown, the regression model is significant with a *P* value of 0.0248 and R-square of 13.25%, which is

deemed acceptable for the purpose of this illustration. However, an inspection of the individual variables shows that none of them is significant at a 0.05 level of significance. The results imply that the length of lapse between discharge and readmission (i.e., diff) for readmission patients cannot be explained by differences in the variables:

- 1) readmission clinical department [a_orgfa],
- 2) readmission admit source [fadmsrce],
- 3) previous discharge status [pdstat],
- 4) chronic conditions [chron1],
- 5) gender of patient [sex],
- 6) age of patient [age], and
- 7) race of patient [race].

Decision Tree Results

The length of lapse (diff) can also be modelled using the decision tree approach. The results are summarised in Figure 7. As shown, for the readmission cases, the length of lapse between discharge and readmission is associated with the age of the patients, the admission clinical department and the chronic conditions. Conversely, the results also mean that the length of lapse is not associated with the other variables such as sex, race and the previous discharge status. These findings can aid the understanding of the readmission data. It is noted that the decision tree results and multiple regression results do not lead to identical conclusions. One possible explanation is that decision trees can incorporate non-linear and complex interaction effects better than multiple regression. Another advantage of decision trees is that the results are rule-based and hence more interpretable.

Other Potential Data Mining Applications in the Context of Casemix

Although the illustrative data mining application above focuses only on readmission data, data mining can also be applied on casemix data in general. For example, it can be used to identify the factors affecting the average length of stay. This can facilitate casemix-based funding by looking at how differences in the demographic factors and medical conditions of patients affect the average length of stay. The same can be done for the consumption of hospital resources and services. Such findings can help identify a more equitable and rational basis of funding. They can also aid better understanding of health care cost behaviour.

Another potential area for data mining applications is the comparison of doctors, departments and/or hospitals. Data mining can detect systematic patterns relating the average length of stay and consumption of hospital resources/services to particular doctors, departments or hospitals. These findings can help understand variations in medical costs as well as aid the formulation of clinical guidelines, best practices or other forms of benchmarks.

Back to the readmission data illustration, further data mining can also be done to investigate the “quicker and sicker” syndrome of premature discharges. In particular, with both admission data and control data (of non-readmission patients), predictive modelling can be performed to examine the differences between readmission and non-readmission patients. The findings will be useful in understanding the possible reasons for readmission cases (e.g., premature discharges, chronic conditions, panic behaviour ... etc.) and factors related to readmission (e.g., age, availability of family care, clinical department ... etc.).

Conclusion

In recent years, data mining has gained widespread attention and increasing popularity in the commercial world. Successful data mining applications in the financial, retailing and manufacturing sectors have been reported. It is thus not surprising that recent surveys found that data mining had grown in usage and effectiveness.^{13,14} Professional bodies^{15,16} have also identified data mining as an important technology for the twenty-first century. However, the usefulness of data mining is not limited to the commercial world. More importantly, data mining is a methodology, process and technology that is beyond mere data/statistical analysis and information technology. It can be considered a decision support system that has application in all sectors. This includes the health care sector and casemix data.

This paper looks at the potential usefulness of data mining in the context of casemix-based funding. The data mining methodology and its tools are also discussed and the data mining literature summarised. An illustrative data mining application using readmission data is included. It is appropriate in this concluding section to highlight some of the limitations of data mining.

Limitations of Data Mining

Firstly, a sufficiently exhaustive mining of data will certainly throw up patterns of some kind that are a product of random fluctuations.¹⁷ This is especially so for large data sets with many variables. Hence, many interesting and/or significant patterns and relationships found in data mining may not be useful.

Secondly, from a statistical perspective, while data mining is well developed for modelling (i.e., prediction), it is not as well developed for effect assessment. Murray¹⁸ and Hand¹⁹ have warned against using data mining for data dredging or fishing (i.e., trawling through data in the hope of identifying patterns) because of the statistical problems involved.

Thirdly, successful application of data mining requires

the user to be knowledgeable in the domain area of application as well as in the data mining methodology and tools. Without a sufficient knowledge of data mining, the user may not be aware of or be able to avoid the pitfalls of data mining.²⁰

Finally, businesses developing data mining applications also need to make a substantial investment of their resources (i.e., money, time and effort) in data mining. It should be borne in mind that data mining projects can fail for a variety of reasons (e.g., lack of management support, unrealistic user expectations, poor project management, inadequate data mining expertise ... etc.).

Despite the limitations highlighted above, it is expected that data mining can play a very important role in the context of casemix. In particular, it is envisaged that data mining could be used to facilitate analysis of hospital-based patient data for better management decision making.

REFERENCES

1. Cheah J, Chee Y C. Casemix—for better or for worse? *Singapore Med J* 1999; 40:9-12.
2. Lowe A. Accounting in health care: some evidence on the impact of casemix systems. *British Accounting Review* 2000; 32:189-211.
3. SAS Institute. From data to business advantages: data mining, the SEMMA methodology and SAS software. North Carolina: SAS Institute, 1998.
4. Nazem S M, Shin B. Data mining: new arsenal for strategic decision-making. *Database Management* 1999; 10:39-42.
5. Vanecko J J, Russo A W. Data mining and modelling as a marketing activity. *Direct Marketing* 1999; 62:52-5.
6. Zorn P, Emanoil M, Marshall L, Panek M. Mining meets the web. *Online* 1999; 23:16-28.
7. Lach J. Data mining digs in. *American Demographics* 1999; 21:38-45.
8. Scholber D. Data detectives. *Telephony* 1999; 237:20-4.
9. Stedman C. Data mining despite the dangers. *Computerworld* 1998; 32:61-2.
10. Brabazon T. Data mining: A new source of competitive advantage? *Accountancy Ireland* 1997; 29:30-1.
11. Dhar J, Stein R. Seven methods for transforming corporate data into business intelligence. New Jersey: Prentice Hall, 1997.
12. de Ville B. SAS Enterprise Miner—first looks. *Marketing Research* 1998; 10:37-40.
13. Davis B. Data mining transformed. *Informationweek* 1999; 751:86-8.
14. Kuykendall L. The data-mining toolbox. *Credit Card Management* 1999; 12:30-40.
15. Freedman J. IIA announces 1997 research priorities. *Management Accounting* 1997; 78:65-6.
16. American Institute of Certified Public Accountants (AICPA). Top 10 technologies – plus 5 for tomorrow. *Accountancy* 1999; 187: 16-7.
17. Hand D J. Data mining: statistics and more? *Am Stat* 1998; 52:112-8.
18. Murray L R. Lies, damned lies and more statistics: the neglected issue of multiplicity in accounting research. *Accounting and Business Research* 1997; 27:243-58.
19. Hand D J. Data mining: statistics and more? *Am Stat* 1998; 52:112-8.
20. McQueen G, Thorley S. Mining fool's gold. *Financial Analysts* 1999; 55:61-72.