American Journal of Applied Sciences 6 (2): 345-351, 2009
ISSN 1546-9239

Generating Treatment Plan in Medicine: A Data Mining Approach

Ahmad Mahir Razali and Shahriyah Ali
School of Mathematical Sciences, Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor D.E., Malaysia

Abstract: This study reports on a research effort on generating treatment plan to handle the error and complexity of treatment process for healthcare providers. Focus has been given for outpatient and was based on data collected from various health centers throughout Malaysia. These clinical data were recorded using SOAP (Subjective, Objective, Assessment and Plan) format approach as being practiced in medicine and were recorded electronically via Percuro Clinical Information System (Percuro). Cross-Industry Standard Process for Data Mining (CRISP-DM) model has been utilized for the entire research. We used data mining analysis through decision trees technique with C5 algorithm. The scopes that have been set are patient's complaint, gender, age, race, type of plan and detailed item given to patient. Acute upper respiratory infection disease or identified as J06.9 in International Classification of Diseases 10 by World Health Organization has been selected as it was the most common problem encountered. The model created for J06.9 disease is that type of plan recommended through giving drug to patients without the need to consider patient's complaint, gender, age and race, with accuracy obtained for the model is 94.73%. Inspite of that, we also identified detailed items that have been given to J06.9 patients and the occurancy of them. This can be as a guideline for future treatment with item recommendation is less than 0.078% compared to item inventory in Percuro database. The research is expected to aid healthcare provider as well as to minimize error during treatment process while benefited from technology information to increase the health care delivery.

Key words: SOAP format, percuro clinical information system, cross-industry standard process for data mining (CRISP-DM), international classification of disease and acute upper respiratory infection

INTRODUCTION

Accurate and error-free of diagnosis and treatment given to patients has been a major issue highlighted in medical service nowadays. Traditionally, healthcare providers provide services based on their knowledge and experiences whether individually or collectively depending on cases. Research done proved that hospitals do not all provide the same quality of service even though they provide the same type of service[1]. To achieve service excellence, hospitals must strive for zero defections[2] that require continuous effort to improve the quality of the service delivery system[3].

Treatment plan on the other hand, refers to management on any interventions consists of treatment and/or examination which will be initiated for each problem based on patient's history, physical examination, provisional diagnosis and differential diagnosis[4]. Treatment plan can be generated to provide useful evidence as a basis for future medical practice, by utilizing previous treatment patterns from clinical records database. The amount of collected and stored data in databases has increases dramatically due to advancements in software capabilities and hardware tools, along with decreasing trend of hardware and software cost.

In spite of that, data mining techniques which are part of knowledge discovery in databases (KDD), have become popular research tools for medical researchers who seek to identify and exploit patterns and relationships among large number of variables and be able to predict the outcome of a disease using the historical cases stored within datasets[5,6]. Applications of data mining have already been proven to provide benefits to many areas of medicine, including diagnosis, prognosis and treatment[7]. Data mining techniques have been applied in various medical fields, amongst with are health administration[8,9], adverse drug reactions[10-12], drug safety[13,14], predicting breast cancer survivability[15], predicting survival time for kidney

Corresponding Author: Ahmad Mahir Razali, School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor D.E., Malaysia

dialysis patients[16], knowledge discovery in hypoplastic left heart syndrome[17] and predicting protein function[18].

Because of these issues, we think there is a need of aid for health practitioners during treatment process, as well as consideration of patient's well being. Taking advantage of massive clinical data gathered from information technology and the importance of data mining nowadays in decision making, generating treatment plan seems to encounter the above mentioned problems.

MATERIALS AND METHODS

This research is conducted based on outpatient clinical data gathered from various health centers throughout Malaysia. These data were stored electronically via Percuro Clinical Information System (Percuro), which was provided by RareSpecies Corporation Sdn. Bhd., a medical software development company. Figure 1 shows Percuro main module for consultation and treatment session[19].

Percuro applies SOAP (Subjective, Objective, Assessment and Plan) format in recording all medical information as being practiced in medicine. Further details on SOAP can be seen in Table 1.

All data related to patients are recorded electronically direct into patients' record and are stored in the Percuro database. Demographic information was recorded by staffs in charge in registration counter during registration while clinical information was recorded by health practitioners during treatment process.

Throughout generating treatment plan, Cross-Industry Standard Process for Data Mining (CRISP-DM) has been used as foundation (Fig. 2). CRISP-DM has a life cycle consisting of six phases. Each phase is followed until research objective is achieved.

Business understanding was well defined and data understanding was thoroughly observed, before data preparation can be made. After understanding the whole set of data and what can be extracted from it, the objective of the study was determined.

As much as 88,355 clinical data have been gathered for the duration period of 18 months. Acute upper respiratory infection (J06.9 as identified in International Classification of Disease 10 by World Health Organization) was set to be the disease for generating treatment plan as it was the most common problem encountered. Acute upper respiratory infection is a severe adenovirus infection of the respiratory tract characterized by fever, sore throat and cough.

Table 1: SOAP format

SOAP format		Details
S	Subjective data	History: Information requested from patients on principal symptoms, history of present illness, past history, social history, family history and systems review
O	Objective data	Physical examination, provisional diagnosis and differential diagnosis: Records from physical and laboratory findings that relevant to patient's complaint
A	Assessment	Interpretation of any relevant findings for each problem.
P	Plan	Any interventions that will be initiated for each problem consists of treatment and/or examination. Treatment: drug, procedure Examination: laboratory, imaging

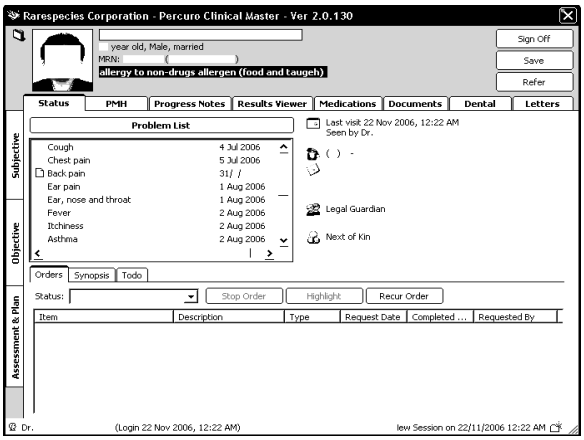

Fig. 1: Percuro main module for consultation and treatment session

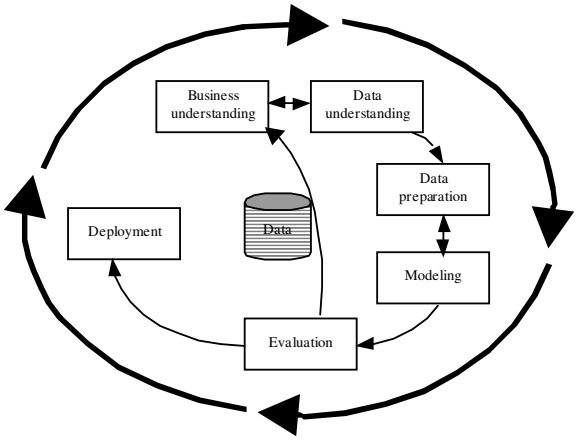

Fig. 2: CRISP-DM model (Adopted from http://www.crisp-dm.org)

This study did not consider all medical parameters that have been listed in Table 1. But however this study

put into account patient's complaint, gender, age, race and also type of plan given to patient. Figure 3 shows our approach in detail.

Data preparation is the most time consuming and labor intensive phase. The records from the last six months of the Percuro database were used for this project, while the records from the first 12 months have been discarded. The main reason for doing this was because Percuro system was under maintenance during the first 12 months and in addition, the number of records stored in Percuro within the period was not consistent. At the end of this phase, only 664 records were used for treatment plan modeling purposes from the total amount of 88,355 clinical data that have been gathered by Percuro. Only records that complied to study scope were considered.

In spite of that, data preparation was also done to obtain age variable from dataset. This is because the data only provides patient's birth date. Therefore, modification on the data was made to get age variable by differencing birth date and consultation date.

Preparation for type of treatment plan variable is also needed. Treatment plan given to patient consisted of one or combinations of drug, procedure, laboratory and imaging (as shown in Table 1). Percuro on the other hand, kept these clinical data in patient's consultation record individually, for every type of treatment plan. In order to capture the complete treatment given to patient and representation data was made. The treatment given to patient were coded in one single record per consultation for all 14 types of treatment plan possibilities. A different table of data set has been created to store the new coded variable (type of plan given to patient). The table was later joined together with age variable as well as other variables to get final data set for modeling phase.

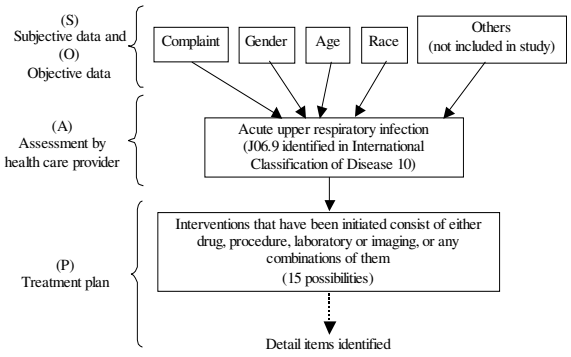

Fig. 3: A sketch of our approach to generate treatment plan and the association with SOAP format

RESULTS AND DISCUSSION

Modeling, evaluation and deployment phases of CRISP-DM model are covered in this section. In order to get the treatment plan model for J06.9 disease, data mining analysis was conducted. As mentioned earlier, final data set from Data Preparation Phase inclusive of 664 records was used in this phase. Table 2 displays the frequency for each classification in variables that contain values in this study.

Since clinical data used has many categorical variables, data mining analysis through Decision Trees technique is recommended[20]. Many researchers have also applied this technique[18,21] and have proved that it gave the best result among other techniques in data

Table 2: Frequency for classification in variables that contain values in study

Atribute	Classification	Frequency
Age	< = 12 years (children)	7
	> 12 years (adult)	657
Gender	Female	454
	Male	210
Race	Malay	464
	Missing value	47
	Iban	3
	Kedayan	1
	Iban/sea dayak	1
	Kayan	1
	Bidayuh	1
	Dusun	2
	Myanmar	1
	Indian	1
Complaint	Fever	270
	Cough	177
	Sore throat	168
	Running nose	22
	Others	6
	Headache	5
	Rhinitis	3
	Asthma	2
	Swelling	2
	Allergy	1
	Back pain	1
	Chest pain	1
	Diarrhea	1
	Ear, nose and throat	1
	Hypertension	1
	Lump	1
	PR bleeding	1
	Skin rash	1
Plan	Drug	629
	Drug and lab	31
	Imaging	1
	Drug and Imaging	1
	Drug, procedure and imaging	1
	Lab	1

Table 3: Summary of output generated by Mode 1 and Mode 2

Modeling phase	Mode 1	Mode 2
Training:		
Pruned tree	Drug	Drug
Number of leaves	1	1
Size of the tree	1	1
Testing:		
Correctly classified instances	219 (96.90%)	629 (94.73%)
Incorrectly classified instances	7 (3.10%)	35 (5.27%)
Total number of instances used	226	664

mining[15], especially in overall prediction accuracy[16]. Thus we adopted the same technique, with C5 algorithm and tested two test modes; Mode 1) split data to 66% for training and 34% for testing and Mode 2) 10-fold cross-validation. The summary of output generated by these modes is shown in Table 3.

There are not any differences between the two modes in training the model. The pruned tree was set to be drug for both modes, thus giving the number of leaves and size of the tree as 1. However, differences occur for the model testing. Mode 2 gives more correctly classified instances and also total number of instances used compared to Mode 1. Mode 1 used only 226 records for model testing, that is 34% out of total records. Whereas for Mode 2 used 1/10 of total records for testing the model created and kept on repeating the process for the second portion of 1/10 and so forth until 10 times. This gave total records used for model testing as 664.

Both modes gave the same results for attributes selection, model structure, number of leaves and size of tree. After evaluating these results, we decided to choose drug classification from plan attribute to be the model for our study as can be seen in Fig. 4.

The accuracy for the model is set to be 94.73% and not 96.90% because of higher number of records considered. The gained result left other studied attributes behind and did not include them in the treatment plan model. In short, the model created for J06.9 disease is that type of plan recommended through giving drug to patients without the need to consider patient's complaint, gender, age and race, with accuracy obtained for the model is 94.73%.

The result seems not to be surprising. It is true that most of outpatient cases are treated through drug giving. We tried to prove by considering Percuro database on interventions given to patients. Drug covered as much as 92.79%, while other types of interventions beared a very little percentage. Refer Table 4 for details.

Next, we tried to deploy the created model by identifiying types of item given to treated patients with J06.9 disease for the duration period of 6 months. The findings are shown in Fig. 5. There have been

Table 4: Interventions given to patients from Percuro database

Intervention	Frequency	Percentage (%)
Drug	100,597	92.79
Laboratory	6,251	5.77
Procedure	1,203	1.11
Imaging	358	0.33
Total	108,409	100.00

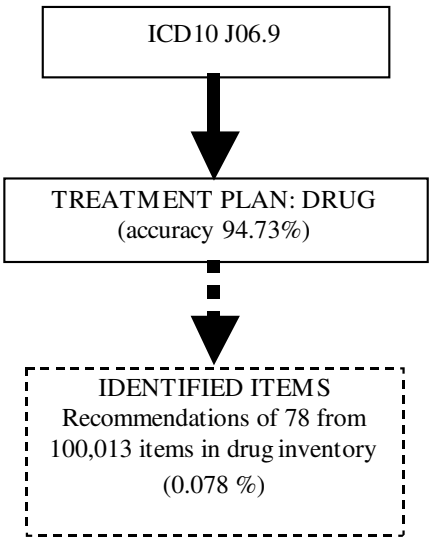

Fig. 4: Treatment plan model for acute upper respiratory infection (J06.9)

75 types of drug items found during this period, but only top 13 are listed here. Eumentol, Paracetamol 500mg and Ascorbic Acid 500mg are among the most common items which covered more than 50% of total drug items given to patients. The 65 remaining items are represented by others and each item is less than 1% each. With the information, the risk of making mistakes will be very much lower as drug suggestion for J06.9 disease is only 0.078%, that is 78 possibilities out of 100,013 types of drug item found in Percuro drug inventory.

It is very much important to get to know the whole overview of a subject and this applies to medicine field especially. Since model is not 100% accurate, we tried to study the remaining treatment pattern for an addition to the above results. There are another 35 cases out of 664 that differ from drug treatment. The second highest intervention came from drug-laboratory which consisted of 31 cases. We executed detailed items given to patients from Percuro database as shown in Figure 6. As much as 29 drug items being given out to patients along with 3 items from laboratory. Decision that have been made for drug choices was 0.029% from 100,013 items in drug inventory. It is noticeable that the first

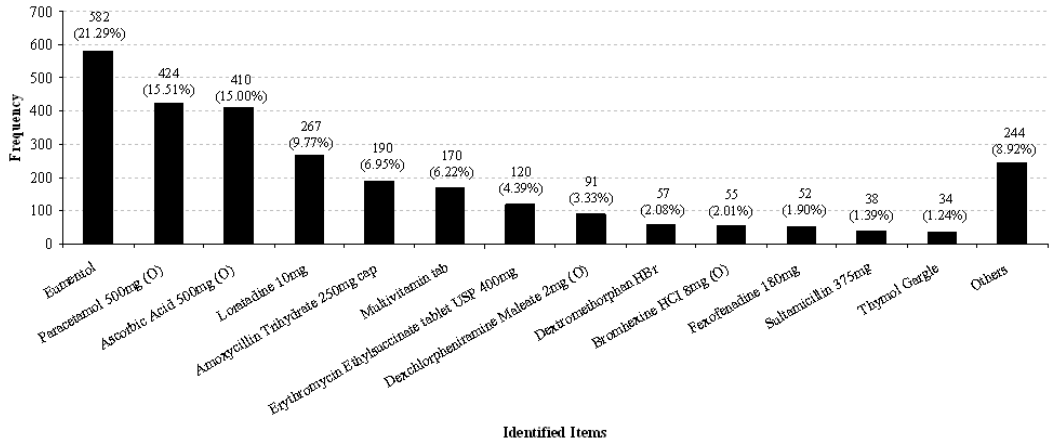

Fig. 5: Drug items been given to J06.9 patients for drug treatment

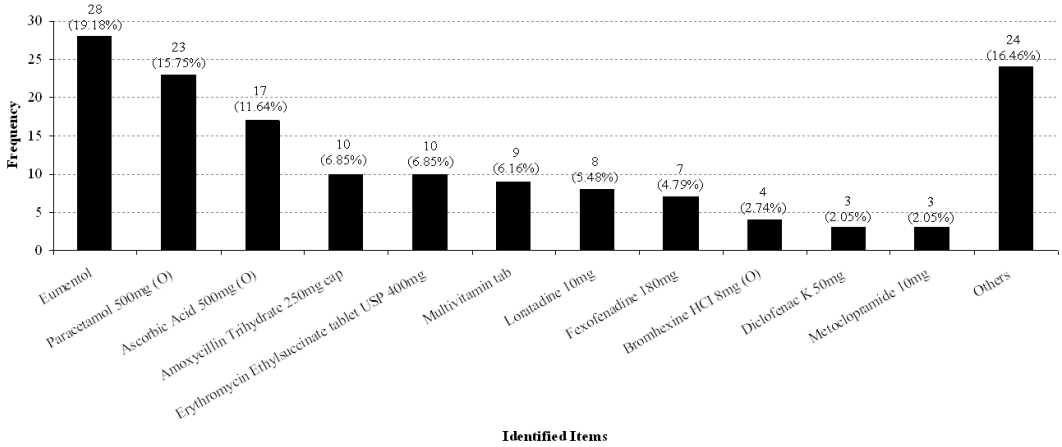

Fig. 6: Drug items been given to J06.9 patients for Drug-Laboratory interventions

three items for drug were the same as in Figure 5. Other items had different order, but types of items were almost the same for both treatment. Whereas for laboratory, choices was 9.68% from 31 classification in laboratory. Full blood count represented the highest occurance as much as 91.89% (34 cases), while urine FEME and G6PD represented 5.41% (2 cases) and 2.70% (1 case) respectively. Four more cases are from imaging, drug-imaging, drug-procedure-imaging and laboratory interventions, which represented one case each. Summarized findings are shown in Table 5.

Data mining can be an important tool in medical field. The treatment plan model included herein is just one example of the value of data mining. This method can be applied to other diseases in order to generate treatment plan. However it is advisable to include all medical parameters in the analysis. Parameters such as symptom, result investigation, laboratory investigation and physical investigation are important and are part of treatment plan decision making. Utilization of these parameters can ensure a better model be build for acute upper respiratory infection treatment plan, as well as can provide closest result as health practitioners. It is essential to do so if patients are to feel safe and completely reliant on the service offered.

Comparison with other techniques that use different algorithms such as rough-set, neural network, regression modeling and clustering can also be done. The methodology presented here can be adopted except for the modeling phase that should accommodate with the algorithm chose.

The data set used herein was rather small after being cleansed out during data preparation. A larger set is recommended because it could provide more meaningful results and useful evidence as a basis for future medical practice. By identifying patterns within

Table 5: Intervention summary for J06.9 disease from Percuro database

	Treatment type detailed items			
Treatment plan	Drug	Procedure	Imaging	Laboratory
Drug	As in Fig. 5	-	-	-
Drug-laboratory	As in Fig. 6. All item type were found to be the same as Figure 5 but addition for items: 1. Dimenhydrinate 50 mg 2. Doxycycline HCl 100 mg 3. Methyl salicylate 12.7%, Menthol 5.8% 4. Metoclopramide 10 mg 5. Metronidazole BP 200 mg (O) 6. Sangobion	-	-	full blood count
Imaging	-	-	chest-posterior anterior (PA)	-
Drug-imaging	Eumentol, ascorbic acid 500 mg (O) and Dextromethorphan HBr (items belong to first ten items as in Fig. 5)	-	chest-posterior anterior (PA)	-
Drug-procedure-imaging	Combination of drug candesartan cilexetil 16 mg (the 60th item from drug treatment-Fig. 5) dan Hydrochlorothiazide 12.5mg	Echocardiography	chest-posterior anterior (PA)	-
Laboratory	-	-	-	full blood count

the large sums of data, data mining can and should, be used to gain more insight into the diseases, generate knowledge that can potentially fuel lead to further research in many areas of medicine[6].

In order to increase the efficient and practical use of the generated treatment plan, it is best to integrate it with existing clinical information system. Fast result can be generated and be applied in decision making process to determine appropriate treatment plan for patients.

CONCLUSION

The model created for acute upper respiratory infection is that type of plan recommended through giving drug to patients without the need to consider patient's complaint, gender, age and race. It is also shown that most outpatient treatment is through drug giving, not only for acute upper respiratory infection but for other diseases as well.

However, it is adviseable to remember that the findings does not reflect to all treatments all the time. The result reacted as a guidance to monitor future undertakings in order to control them from being far away from the appropriate action. In certain cases where changes occur maybe due to environmental changes, new medical findings and so forth, healthcare providers' service are still needed. Data mining can provide assistance in making the diagnosis or prescribing the treatment, but it still cannot replace the physician's intuition and interpretive skills[6].

Proposed items to be given to J06.9 patients seem to narrow down the probability of making mistakes and reduce the complexity of treatment process. The research is expected to aid healthcare provider during treatment while benefited from technology information to increase our health care delivery.

ACKNOWLEDGEMENT

Our special thanks to RareSpecies Corporation Sdn. Bhd. for the clinical data sets and Dr. Shamsul Amri Ramli for medical significance related inputs.

REFERENCES

1. Youssef, F.N.,D. Nel and T. Bovaird, 1996. Health care quality in nhs hospitals. Int. J. Health Care Qual. Assur., 9: 15-28.
2. Reichheld, F.F. and W.E. Sasser, 1990. Zero Defections: Quality comes to services. Harvard Business Rev., 68: 105-111.
3. Lim, P.C. and N.K.H. Tang, 2000. A study of patients' expectations and satisfaction in singapore hospitals. Int. J. Health Care Qual. Assur., 13: 290-299.
4. Talley, N. and S. O'Connor, 1992. Clinical Examination. 2nd Edn. New South Wales: MacLennan and Petty Pty Limited.
5. Lavrac, N., 1999. Selected techniques for data mining in medicine. Artificial Intel. Med., 16: 3-23.
6. Richards, G., V.J. Rayward-Smith, P.H. Sonksen, S. Carey and C. Weng, 2001. Data mining for indicators of early mortality in a database of clinical records. Artificial Intel. Med., 22: 215-231.

7. Whiting-O'Keefe, Q.E., D.W. Simborg, W.V. Epstein and A. Warger, 1985. A computerized summary medical record system can provide more information than the standard medical record. J. Am. Med. Assoc., 254: 1185-1192.
8. Kum, H.C., D. Duncan, K. Flair and W. Wang, 2003. Social welfare program administration and evaluation and policy analysis using knowledge discovery and data mining (KDD) on administrative data. In: Proceedings of the NSF National Conference on Digital Government Research (DG.O), pp: 39-44.
9. Rao, R.B., R.S. Sandilya, R.S. Niculescu, C. Germond and H. Rao, 2003. Clinical and financial outcomes analysis with existing hospital patient records. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 416-425.
10. Murff, H.J., V.L. Patel, G. Hripcsak and D.W. Bates, 2003. Detecting adverse events for patient safety research: A review of current methodologies. J. Biomed. Inform., 36: 131-143.
11. Harvey, J., C. Turville and S. Barty, 2004. Data mining of the australian adverse drug reactions database: A comparison of bayesian and other statistical indicators. Int. Trans. Operat. Res., 11: 419-433.
12. Chen, J., H. He, G. Williams and H. Jin, 2004. Temporal sequence associations for rare events. In: Proceedings of 8th Pacific-Asia International Conference on Knowledge Discovery and Data Mining (PAKDD), Lecture Notes in Computer Science (LNAI 3056), Sydney, pp: 235-239.
13. Almenoff, J.S., W. DuMouchel, L.A. Kindman, X. Yang and D. Fram, 2003. Disproportionality analysis using empirical bayes data mining: A tool for the evaluation of drug interactions in the post-marketing setting. Pharmacoepidemiol. Drug Saf., 12: 517-521.
14. Wilson, A., L. Thabane and A. Holbrook, 2004. Application of data mining techniques in pharmacovigilance. Br. J. Clin. Pharmacol., 57: 127-134.
15. Delen, D., G. Walker and A. Kadam, 2005. Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intel. Med., 34: 113-127.
16. Kusiak, A., B. Dixon and S. Shah, 2005. Predicting survival time for kidney dialysis patients: A data mining approach. Comput. Biol. Med., 35: 311-327.
17. Kusiak, A., C.A. Caldarone, M.D. Kelleher, F.S. Lamb, T.J. Persoon and A. Burns, 2006. Hypoplastic Left heart syndrome: Knowledge discovery with a data mining approach. Comput. Biol. Med., 36: 21-40.
18. King, R.D., P.H. Wise and A. Clare, 2004. Confirmation of data mining based predictions of protein function. Bioinformatics, 20: 1110-1118.
19. RareSpecies Corporation Sdn. Bhd, 2007. Percuro Clinical Information System. http://www.rarespecies.com.my/index.php [14 September 2007].
20. Berry, M.A.J. and G.S. Linoff, 2004. Data Mining Techniques: For Marketing, Sales and Customer Relationship Management. 2nd Edn. John Wiley and Sons, New York.
21. Calderon, T.G., J.J. Cheh and I.W. Kim, 2005. How large corporations use data mining to create value. Accountants Today, September: 24-31.