An Application of Classification Analysis for Skewed Class Distribution in Therapeutic Drug Monitoring – The Case of Vancomycin

Jian-Xun Chen
Institute of Information Management
National Sun Yat-sen University
Jamesc02@ms9.hinet.net

Tsang-Hsiang Cheng
Department of Business Administration
Southern Taiwan University of Technology
cts@mail.stut.edu.tw

Agnes L.F. Chan*,
Hue-Yu Wang
Department of Pharmacy
Chimei medical center
{cmh5500,cmh9005}
@mail.chimei.org.tw

Abstract

Vancomycin can induce potent adverse side effects if drug concentration is not controlled within a narrow safety range. Therefore, therapeutic drug monitoring (TDM) is followed to adjust dose and help monitor treatment effects. Because TDM are not helpful in patients taking vancomycin for the first time, use it to ensure medication safety has a limitation. This study aimed at using decision tree induction to predict outcomes of vancomycin. Research results demonstrate that the asymmetric distribution among classes in the TDM data would result in prediction deviation. An ideal model with good prediction efficacy could be established by adjusting the ratio among outcome classes through "over-sampling for expanding minority data". The prediction model would be helpful in controlling the positive and negative effects of vancomycin treatment, improving care at the patient level and improving costs at the social level. Some interesting decision rules derived from the decision tree were analyzed its clinical meanings. Precious prescription knowledge is thus extracted and accumulated.

Keywords：vancomycin, therapeutic drug monitoring (TDM), C4.5 decision tree analysis, prediction deviation

1. Introduction

Since vancomycin was first marketed in 1958, it has become one of the most commonly prescribed antibiotics in medical institutions due to its powerful bacteriostatic effect on Gram-positive bacteria and its effective treatment to Methicillin-resistant *Staphylococcus aureus* (MRSA). However, vancomycin has a very narrow range of effective concentrations (peak and trough blood concentrations 20 ~ 40 and 5~10 μg/mL, respectively). Extremely high vancomycin concentrations can induce toxic responses, whereas low concentrations can induce bacterial resistance to antibiotics, influencing immediate therapeutic effect and probability of future success with antibiotic treatment. Therefore, therapeutic drug monitoring (TDM) is usually employed to adjust dose and help monitor treatment effects. TDM is a procedure that encompasses the measurement of serum drug concentrations in blood or other fluids (e.g. saliva) and the application of clinical pharmacokinetics to optimize drug therapy in individual patients [4]. The initial dose of vancomycin can be estimated based on the physician's personal experience or a simple nomogram dosing method based on information about gender, age, SCR, and BUN [10, 11, 14], but the inexactitude of the methodology leaves open the possibility of insufficient or over dosage.

Data mining can extract potential valuable knowledge submerged in a large body of existing data to establish a model describing the correlation between characteristics of data and results of events, which is helpful in decision-making. It has been widely applied in medical research including the prediction of optimal respirator pressure in intensive care units [5], breast cancer diagnosis [13], prediction of cardiac ischemia [7], search for appropriate clinical pathway [9], and so on. These applications have achieved good results in their setting. The present study aimed at using decision (or classification) tree technique, which can generate easy and interpretable rules, to construct a classification model for predicting the outcome of vancomycin treatment from a TDM data. A successful model would be helpful in controlling the positive and negative effects of vancomycin treatment, improving care at the patient level and improving costs at the social level.

The decision tree generates rules for the classification of a dataset. It is focused on use of training

* corresponding author

data (historical data) with known classes and attributes (the characteristics of data that may affect the class of training data) to establish a tree-like structure that represent the correlation between attributes and classes. This will help users to classify a new (unclassified) dataset.

The key concept of decision trees is using an attribute selection function to choose the most appropriate attribute from all candidate attributes as the root nodes and internal nodes that can used to classify training data. Basic algorithm of decision trees is shown as following:

1. First of all, construct the root node of the decision tree. Set the root node as the current node C, and assign all the training examples to node C.
2. If all the examples of node C belong to the same class, then this class should be identified as decision class of node C, and the analysis can be stopped. Otherwise, proceed to step 3.
3. According to all examples of node C, with use of attribute selection function, choose the attribute Ac that has the best differential capacity from all candidate attributes Ai. Attribute Ac should not been used on the pathway from the root node to the current node C.
4. Suppose the attribute value of the chosen attribute Ac has m values; establish child node C1, C2,…., to Cm under the node C according to the m value of Ac, and assign all examples of node C into appropriate child nodes based on the attribute value.
5. Set each child node Ci as the current node C, goes to step 2, and proceed to construct decision tree.

Most common decision tree methods include ID3、C4.5 [12]、CN2 and AQ15 [1]. The major difference between them is attribute selection they used. The C4.5 is an extension of the ID3 algorithm and it is capable of dealing with numerical (continuous) attributes. It uses Gain-Ratio Measure to select attributes so that it can prevent from the preference of choosing multi-value attribute. Based on these considerations, C4.5 is chosen in this study.

2. Methods

The dataset

The dataset in the present study was obtained from the pharmaceutical records of 1075 patients who were prescribed vancomycin for more than five days and received TDM monitoring at a southern medical center from January 1990 to December 2003. Patients receiving dialysis were excluded because of impaired renal function and the differential hemodialytic capacity, which would complicate the factors affecting circulatory drug concentration. Therefore, the TDM records of 955 eligible patients formed the dataset. When records with missing body weight information were excluded, 669 complete TDM records were identified. We categorize the TDM records as adequate and inadequate treatment outcomes in consultation with professional pharmacists. The definition of adequate treatment outcome for TDM data is documentation of peak and trough of circulatory drug concentrations of 20 ~ 40 μg/mL and below 10 μg/mL, respectively, after more than five days of vancomycin treatment. Among the 669 records, 122 patients and 547 patients were classified as adequate and inadequate treatment outcomes, respectively.

The present study used seven TDM data attributes, including gender, age, and body weight, renal function tests BUN and SCR, as well as vancomycin dose and dosing intervals to construct the decision tree for predicting the adequacy of vancomycin treatment. The attributes and statistical description of vancomycin TDM data are listed in Table 1.

Table 1. Attributes and descriptive statistics of vancomycin TDM data

Attributes	Attribute values	descriptive statistics
Gender	M for male, F for female	462 male and 207 female
Age (years)	Minimum 17, maximum 98	Mean 61.21, standard deviation 15.5
Body weight (kg)	Minimum 30, maximum 121	Mean 63.77, standard deviation 14.48
BUN	Minimum 1.14, maximum 162.2	Mean 29.7, standard deviation 24.1
SCR	Minimum 0.22, maximum 5.91	Mean 1.27, standard deviation 0.73
Dose (mg)	Minimum 50, maximum 2000	Mean 690.4, standard deviation 241.58

Dosing Intervals	Minimum 6hr, maximum 48hr	6hr:126 persons 8hr:131 persons 12hr:271 persons 24hr:120 persons 48hr:21 persons

Construction of prediction model

The model for predicting vancomycin treatment outcome was initially established by using C4.5 decision tree methods to analyze the TDM data. The prediction efficacy of the model is usually good if the classification model is established by using data that is evenly distributed among various classes. However, the model prediction results would deviate to the class with the majority of data if a significant difference exists in data quantity among various classes. This was the situation in our TDM data because there was a significant difference in the distribution of data class Y and N (Y, adequate: N, inadequate = 122 : 547).

Under-sampling for reducing majority data [6, 15] and over-sampling for expanding minority data [3, 8] are the two most common methods in the literature to resolve a putative prediction deviation. Under-sampling for reducing majority data can resolve prediction deviation through exclusion of some data from majority class to lessen the distribution difference among classes. However, the constructed decision tree would suffer from insufficient training due to the possibility of an insufficient training data.

On the contrary, over-sampling for expanding minority data is done to lessen the distribution difference among various classes by expanding minority data. DeRouin [3] proposed decreasing the difference in data quantity among classes by replicating minority data using a neural network. Lewis and Catlett [8] indicated that direct replication of data from minority classes is the simplest and most effective method to resolve prediction deviation problem, and this is the method we used in this study.

Evaluation criteria

Two performance indexes, including True Positive rate (TP rate) and False Positive rate (FP rate) for each class[16] were used to evaluate prediction efficacy. Taking the confusion matrix in Table 2 as an example, the TP rate for Y class is the ratio $[\frac{A}{(A+B)}]$, which is the number of the records correctly predicted as Y class over the number of the records belonging to the real Y class; the FP rate is the ratio $[\frac{C}{(C+D)}]$, which is the number of the records that incorrectly predicted as Y over the number of records belonging to the real N class. If we focused on the N class, the TP rate for N class would be $\frac{D}{(C+D)}$ and FP rate would be $\frac{B}{(A+B)}$. The overall accuracy formula $\frac{(A+D)}{(A+B+C+D)}$ evaluates the overall efficacy of the prediction by divided the number of correctly predicted as Y or N with the number of all training records.

The outcome prediction model constructed in this study requires robust prediction efficacy for each class so that it could help medical professionals. Therefore, the constructed model needs to have good TP rate with a low FP rate for each class.

Table 2. Confusion matrix

(Unit: number)	Predict adequate outcome (Y)	Predict inadequate outcome (N)
Adequate outcome(Y)	*A*	*B*
Inadequate outcome (N)	*C*	*D*

The above efficacy was evaluated by using 10-fold cross-validation. The technique of 10-fold cross-validation randomly divides the training data into 10 groups with equal number of data according to the original ratio of data classes, then uses 9 groups to construct the model and use the rest group to test its efficacy, respectively. The above steps are performed 10 times for each group to be the test data in turn; then efficacy of the prediction is the average of the 10 experimental results.

3. Results

We used data mining software Weka [16] for construction and performance evaluation of our prediction model. The prediction efficacy of the decision tree is shown in Table 3.

Table 3. Evaluation of efficacy of the decision tree

	TP rate	FP rate
Y class	0.180	0.011
N class	0.989	0.820
Accuracy: 84.16 %		

Table 3 shows that the prediction results of the decision tree have an apparent deviation toward N class

with majority data, whereas the prediction efficacy for the Y class with minority data was bad. The model prediction deviation possibly originated from the asymmetric distribution of data among classes (Y:N = 122:547) Therefore, over-sampling for expanding minority data as proposed by Lewis and Catlett [8] was implemented to adjust the classes distribution of data and, hopefully, to solve the problem of prediction deviation.

Over-sampling for expanding minority data to solve prediction deviation

To determine the optimal replication number for promoting efficacy of the prediction model, we expanded the data quantity of Y class to various status, such as 244, 366, 488, and 610; the effect on the efficacy of the prediction model was investigated by 10-fold cross-validation again; results are shown in Table 4.

Table 4. The effect of over-sampling for expanding minority data on model categorical efficacy

Ratio between classes (Y:N)	Y:TP rate	N:TP rate	Y:FP rate	N:FP rate	Accuracy
122:570	0.180	0.989	0.011	0.820	84.16 %
244:570	0.533	0.870	0.130	0.807	76.61 %
366:570	0.716	0.823	0.177	0.284	77.98 %
488:570	**0.828**	**0.768**	**0.232**	**0.172**	**79.61%**
610:570	0.884	0.755	0.245	0.116	82.28 %

Over-sampling for expanding minority data efficiently promoted the accuracy and TP rate for Y class. The TP rate of N class was slightly decreased due to alteration of class distribution. Considering the requirement of good prediction efficacy to each outcome class to meet the application requirement, we concluded that replicating data of Y class four times would significantly improve the efficacy of prediction model

Decision rules

The interpretation of the decision tree requires intensive discussions between information technologists and clinical pharmacists. Knowledge of treatment efficacy can provide clinicians with important information for making quantitative therapeutic decisions.

According to the decision tree has been constructed, the order of attributes the decision tree used for classification are: dosing interval、weight or SCr、age and dose. Some important rules derived from the decision tree are listed below.

R1: If (dosing interval is 24 hours) and (weight greater 30kg) and (SCr less than 0.99mg/dL) and (age greater than 50), then given dose between 500mg and 1000mg usually will obtain expected treatment outcome.

R2: If (dosing interval hour less then 8 hours) and (weight between 50kg and 59 kg) and (SCr less than 1.99mg/dL), then given dose 500mg usually will not have the expected treatment outcome.

R3: If (dosing interval equals to 12 hours or 24 hours or 48 hours) and (weight greater then 50kg) and (SCr between 2.01mg/dL and 2.99mg/dL) and (age greater then 40), then given dose between 500mg and 1000mg usually will not have the expected treatment outcome.

4. Conclusions

Clinical effect and medical safety are important factors when considering treatment with vancomycin. Although several methods in the past few years have been proposed to adjust doses for various patient populations, unexpected outcome still exists. The present study established a prediction model for evaluating outcome of vancomycin treatment by using decision tree analysis; use of this model would help medical professionals to make full use of the experience submerged in the vancomycin TDM records in order to control the effect of vancomycin treatment and reduce possible adverse side effects.

Another aim of this study is to facilitate the discovery of concise and interpretable information from large amounts of data. The decision tree can be transformed into easily understandable rules. Precious prescription knowledge can be accumulated in medical institutes if clinical pharmacists can analyze and interpret the clinical meanings of decision rules by using their own clinical experiences and pharmaceutical theory. In addition, the model can be used to teach medical professionals in training how to provide timely, effective, and safe treatment that can achieve the purposes of shortening treatment course and decreasing medical costs.

Although the present study constructed the prediction model for treatment outcome only on the bases of vancomycin-specific TDM data, the establishment

procedure can be applied to various medications for which TDM monitoring is used. Future research can integrate other data mining techniques and pharmacokinetic theory to develop a prediction model that can correctly predict medication dose. This would allow medical professionals to have the ability to predict the dose and use of medication with both more convenience and precision, improving the quality of medical care.

Acknowledgment

This study was supported by the Chimei medical center of R.O.C. under contract CMFHR9334.

References

[1] P. Clark, and T. Niblett, "The CN2 Induction Algorithm", *Machine Learning*, Vol. 3, 1989, pp.261-283.

[2] B. W. Corrigan, P. R. Mayo and F. Jamali, "Application of a neural network for gentamicin concentration prediction in a general hospital population", *The Drug Monitoring*, Vol. 19, No. 1, 1997, pp.25-28.

[3] E. DeRouin, J. Brown, H. Beck, L. Fausett and M. Schneider, "Neural Network Training on Unequally Represented Classes", *Intelligent Engineering Systems Through Artificial Neural Networks*, ASME Press, New York, 1991, pp.135-145.

[4] S. Dhillon and J. Cape, Clinical Pharmacy Practice Guide - Therapeutic Drug Monitoring, United Kingdom Clinical Pharmacy Association, 1987.

[5] S. Ganzert and J. Guttmann, "Analysis of respiratory pressure-volume curves in intensive care medicine using inductive machine learning," *Artificial Intelligence in Medicine*, Vol. 26, No. 1-2, 2002, pp.69-86.

[6] P. E. Hart, "The Condensed Nearest Neighbor Rule", *IEEE Transactions on Information Theory*, IT-14, 1968, pp.515-516.

[7] M. Kukar, I. Kononenko and C. Groselj, "Analysing and Improving the Diagnosis of Ischaemic Heart disease with Machining Learning", *Artificial Intelligence in Medicine*, Vol. 16, No. 1, 1999, pp.25-50.

[8] D. Lewis and J. Catlett, "Heterogeneous Uncertainty Sampling for Supervised Learning", *Proceedings of the 11th International Conference on Machine Learning*, 1994, pp.144-156.

[9] F. R. Lin, S. P. Chou and Y. Chen, "Mining Time Dependency Patterns in Clinical pathways", *International Journal of Medical Informatics*, Vol. 62, No. 1, 2001, pp.11-25.

[10] G. R. Matzke, R. W. Mcgory and C. E. Halstenson, "Pharmacokinetics of vancomycin in patients with various degree of renal function", *Antimicrob Agents Chemother*, Vol. 25, No. 4, 1984, pp.4337.

[11] R. C. Moellering, D. J. Krogstadf and D. J. Greenblatt, "Vancomycin therapy in patients with impaired renal function: a nomogram for dosage", *Annals of Internal Medicine*, Vol. 94, 1981, pp.343-346.

[12] J. R. Quinlan, "Induction of Decision Tree", *Machine Learning*, Vol. 1, 1986, pp.81-106.

[13] A. L. Ronco, "Use of Artificial Neural Networks in Modeling Association of discriminant Factors: Towards An Intelligent Selective Breast Cancer Screening", *Artificial Intelligence in Medicine*, Vol. 16, No. 3, 1999, pp.299-309.

[14] J. C. Rotschafer, K. Crossley and D. E. Zaske, "Pharmacokinetics of vancomycin: observation in 28 patients and dosage recommendations", *Antimicrob Agents Chemother*, Vol. 22, 1982, pp.391-394.

[15] D. Skalak, "Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms", *Proceedings of 11th Machine Learning Conference*, 1994, pp.293-301.

[16] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools with Java implementations, Morgan Kaufmann, San Francisco, 2000.

[17] B. K. Wong, T. A. Bonovich and Y. Selvi, "Neural Network Applications in Business: A Review and Analysis of the Literature", *Decision Support Systems*, Vol. 19, 1997, pp.301-320.