Representing Association Classification Rules Mined from Health Data*

Jie Chen[1], Hongxing He[1], Jiuyong Li[4], Huidong Jin[1], Damien McAullay[1], Graham Williams[1,2], Ross Sparks[1], and Chris Kelman[3]

[1] CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra ACT 2601, Australia
{Jie.Chen,Hongxing.Heinst,Huidong.Jin}@csiro.au
{Damien.McAullay,Graham.Williams,Ross.Sparks}@csiro.au
[2] Current address: Australian Taxation Office, Australia
Graham.Williams@togaware.com
[3] National Centre for Epidemiology and Population Health,
The Australian National University, Australia
Chris.Kelman@anu.edu.au
[4] Department of Mathematics and Computing,
University of South Queensland, Australia
jiuyong@usq.edu.au

Abstract. An association classification algorithm has been developed to explore adverse drug reactions in a large medical transaction dataset with unbalanced classes. Rules discovered can be used to alert medical practitioners when prescribing drugs, to certain categories of patients, to potential adverse effects. We assess the rules using survival charts and propose two kinds of probability trees to present them. Both of them represent the risk of given adverse drug reaction for certain categories of patients in terms of risk ratios, which are familiar to medical practitioners. The first approach shows risk ratios when all rule conditions apply. The second presents the risk associated with a single risk factor with other parts of the rule identifying the cohort of the patient subpopulation. Thus, the probability trees can present clearly the risk of specific adverse drug reactions to prescribers.

1 Introduction

Data mining usually involves extracting actionable knowledge from databases. Thus, understanding and evaluating the discovered patterns become increasingly important, especially in health applications. Systematic monitoring of adverse drug reactions is important for both financial and social reasons. At present, the early detection of unexpected adverse drug reactions relies on a national

* The authors acknowledge the Commonwealth Department of Health and Ageing, and the Queensland Department of Health for providing data and support for this research.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1225–1231, 2005.

spontaneous reporting system and collated statistics from overseas agencies [1, 2]. However, the recent availability of a population-based prescribing dataset, such as the Pharmaceutical Benefits Scheme (PBS) data in Australia, when linked to hospital admissions data, provides a unique opportunity to detect rare adverse drug reactions at a much earlier stage before many patients are affected. This paper focuses on identifying the factors, which increase the risk of the adverse drug reaction, directly from large linked health data rather than spontaneous reporting databases.

Prescribed drugs are recorded in PBS data based on the Anatomical and Therapeutic Classification (ATC) system. Adverse events are recorded in hospital data using ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) code. Three case studies have been identified by experts from the Therapeutic Goods Administration, Australia. ACE inhibitors[1] usage associated with Angioedema will serve as the main case study to illustrate our method in this paper. In our data, the distribution of classes with and without adverse events is highly unbalanced due to the intrinsic nature of adverse drug reactions. Moreover, rules identified may be used to alert medical practitioner in their prescription of drugs to certain categories of patients, who are vulnerable to some adverse drug effects. It is therefore essential to present the knowledge to medical practitioners in a form easy to understand and interpret. To address this health data mining problem, we first modify the Optimal Class Association Rule Mining Algorithm [4] to discover rules which identify patient subgroups with a high proportion of patients with target events. We further propose two kinds of tree representation for mined rules to help them and potential users to gain understanding of the rules.

2 Association Classification for Unbalanced Classes

Traditional association classification approaches search for the rules represented by patterns which have high global support and high confidence. Since the "normal" group comprises more than 99% of all cases in the dataset, the class of interest(Class 1 defined in Section 3) is given little attention by these approaches. In this paper, we modify the Optimal Class Association Rule Mining Algorithm [4] by introducing local support and risk ratio to identify higher risk patient groups of adverse drug reaction events. The support in minor class is called *local support* defined as $\frac{sup(A \rightarrow c)}{sup(c)}$. Here $sup(c)$ and $sup(A \rightarrow c)$ represent the support (or proportion) of Class c in the whole population and the support of pattern A in Class c respectively. Minimum local support can be used as a parameter of the algorithm to specify the minimum fraction of population of interest in each class of the unbalanced dataset. We propose to use the *Risk Ratio* as measure of interestingness for pattern mining, which is represented by $RR(A \rightarrow c) = \frac{lsup(A \rightarrow c) sup(\overline{A})}{lsup(\overline{A} \rightarrow c) sup(A)}$.

[1] Angioedema is a swelling that occurs beneath the skin rather than on the surface [3]. There are a number of case series in the literature demonstrating that ACE inhibitor-related angioedema is responsible for as many as 40% of angioedema episodes [3]

Table 1. List of variables used for association classification

Variable	Values	Variable	Values
Gender	m,f	Alimentary tract metabolism	0,1
Age group	1,2,3,4	Blood and blood forming organs	0,1
Indigenous	0,1	Cardiovascular systems	0,1
Sickness(bed days)	1,2,3	Dermatologicals	0,1
Hosp. Neoplasm Flag	0,1	Genito urinary system and sex hormones	0,1
Hosp. Diabetes Flag	0,1	Systematic hormonal preparations	0,1
Hosp. Mental Health Flag	0,1	General anti-infective for systematic use	0,1
Hosp. Circulatory Flag	0,1	Antineoplastic and immunimodulating agents	0,1
Hosp. Ischaemic Heart Disease Flag	0,1	Musculo-skeletal system	0,1
Hosp. Respiratory Flag	0,1	Nervous system	0,1
Hosp. Asthma Flag	0,1	Antiparasitic products insecticides and repellents	0,1
Hosp. Musculoskeletal Flag	0,1	Respiratory system	0,1
Total Scripts	0,1,2	Sensory organs	0,1
Class	0,1	Various	0,1

The risk ratio defines the relative risk (belonging to Class 1) of the patients identified by rule A with respect to the majority of patients [5, p. 35]. $\overline{A}$ denotes the absence of pattern A.

3 Data Preparation and Feature Selection

The Queensland Linked Data Set [6] links hospital admissions data from Queensland Health with the pharmaceutical prescription data from the Commonwealth Department of Health and Ageing, providing a de-identified dataset for analysis. For the implementation of the mining task, we chose to extract profile data for all patients exposed to the drug of interest in a 180 day window, which was selected using domain knowledge. The patients are further partitioned into two classes (Class 1 and 0). The patients in Class 1 are such patients that have taken the target drugs (e.g. ACE inhibitors) within the time window prior to the first adverse drug reaction event, and other patients are in Class 0. Features selected for the profile of each patient are described below.

From the hospital data, demographic variables such as age, gender, indigenous status, postcode, the total number of bed days and the eight hospital diagnosis flags are extracted. The hospital diagnosis and the total number of bed days can be used to infer the health status of an individual. From the PBS data, another 15 variables (including such variables as the total number of scripts of the specified drug and the 14 ATC level-1 drugs) were extracted. The "Total scripts" is used to indicate how long an individual has been exposed to the drug (because each script usually provides medication for one month). The 14 ATC level-1 drug categories may be useful in measuring adverse drug reactions caused by interactions between the specified drug and other drugs.

Table 1 lists the variables representing the profiles of patients. We chose some variables in the profiles in applying the association classification algorithm. "Age", "Bed days" and "Total scripts" are discretised because the algorithm requires all the variables take only a set of discrete values. Since the aim of the algorithm is to identify the group of patients who are more likely to have an adverse drug reaction than the general population, we choose these variables, which are most commonly considered as important for their health and wellbeing. We consulted medical practitioners to incorporate their knowledge in our study.

There are limitations in selecting best variables as our dataset is not from survey data, e.g. some desirable variables such as life style information can not be obtained.

4 Representing Association Classification Rules

Usually when the modified optimal class association rule mining algorithm is applied to identify the high risk groups, a large number of rules with risk ratio greater than 2.0 are generated. The exceptional rules (risk ratio is less than 1.0) could be interesting in identifying lower than general risk groups. However, they are not primary objectives of the current study and therefore ignored. We could not present hundreds of rules to medical experts for inspection. Furthermore, most of them are correlated and provide similar information. We can select rules by an effective method. Let all generated rules match all records in the dataset and only keep the rule with the highest risk ratio for each record. This will reduce the number of rules significantly.

The five rules with highest risk ratio for the ACE inhibitors and angioedema case study are listed below:

Rule 1: RR = 3.9948
- Gender = Female
- Hospital Circulatory Flag = Yes
- Usage of Drugs in category "Various" = Yes

Rule 2: RR = 3.8189
- Age > 60
- Usage of drugs in category of "Genito urinary system and sex hormones" = Yes
- Usage of drugs in category of "Systematic hormonal preparations" = Yes

Rule 3: RR = 3.4122
- Usage of drugs in category of "Genito urinary system and sex hormones" = Yes
- Usage of drugs in category of "General anti-infective for systematic use" = Yes
- Usage of drugs in category of "Nervous system" = No

Rule 4: RR = 3.3269
- Gender = Female
- Age group in [40, 59]
- Total bed days ≥ 15

Rule 5: RR = 3.2605
- Usage of drugs in category of "Alimentary tract metabolism" = No
- Usage of drugs in category of "Genito urinary system and sex hormones" = Yes
- Usage of drugs in category of "General anti-infectives for systematic use" = Yes

For each rule discovered, we conduct further evaluation, e.g., the survival analysis and its significance test [5, pp. 159-169]. In addition, we use the log-rank test, a formal measure of the strength of evidence that two populations have different lifetimes. Fig. 1 presents the estimated survival functions of the subgroup described by Rule 5 (the one within the filled region) and the other patients (within the shaded region). The filled region and the shaded region indicate their confidence intervals, respectively. Clearly, for the age range from 60 to about 80, the subgroup indicated by Rule 5 has significantly higher probability of hospital admission for angioedema than the other patients. The P-value of the log-rank test is 5.0583e-09, which suggests that the sub-group described by Rule 5 is overwhelmingly different from the other patients. Similar interesting results are also found in other rules [6].

The rules identified by the association classification algorithm provide useful knowledge to medical practitioners, and can serve as a reference in their prescription of drugs to the patients. The patients' characteristics can be compared

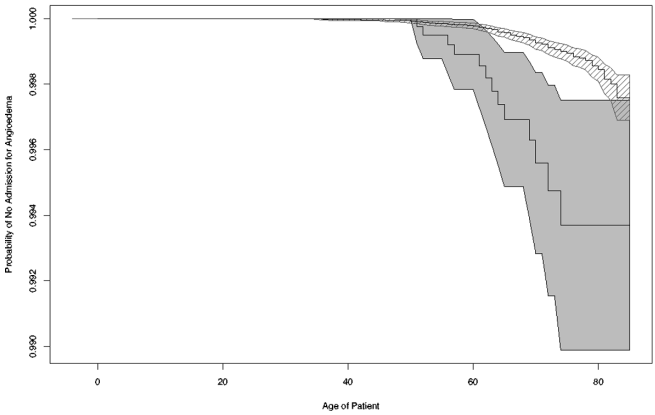

Fig. 1. Fleming-Harrington survival analysis of Rule 5

to the rules to evaluate their risk to the suspected adverse drug reaction. However, the rules presented above may not provide enough information for clinical use. The further breakdown of the risks caused by individual factors provides important information in their assessment of the risk. Therefore we employ a tree structure to visualise the rules mined. A variable value pair is presented at each node of the tree. The information on the support of the population, its percentage and risk ratio is presented on each node. The branch to the right of the node lists the information for complementary population. The level down of each node gives another split of population using a new variable value pair. As an example, Rule 1 is presented as a tree in Fig. 2. Note that most commonly used multiple logistic regression models can be used and similar tree structures could be obtained accordingly. However, the presentation method can rank rules according to their relative risks automatically to avoid time consuming model analysis work.

According to Fig. 2, female users of ACE inhibitors are 1.54 times more likely to have angioedema than the population average. For those female patients who have a circulatory disease, the likelihood increases to 1.82. For those who are female, have a circulatory disease, and also have taken drugs falling in the "Various" category (the 14th ATC level-1 drug category), the likelihood increases further to 4.0. The tree presentation highlights how the risk ratio changes with each individual component. Further stratifications may help to make rules more adaptable in clinical decisions. Alternatively, we can define the risk ratio at each node to be relative to the population of its parent node. Accordingly the risk ratio at each node is expressed by $RR(A \rightarrow C \mid U) = \frac{lsup(A \bigcap U \rightarrow C) sup(\overline{A} \bigcap U)}{lsup(\overline{A} \bigcap U \rightarrow C) sup(A \bigcap U)}$, where U is the rule on the parent node.

The tree presentation of the same rule using the alternative definition of risk ratio is presented in Fig. 3. According to Fig. 3, female users of ACE inhibitors are 1.54 times more likely to have angioedema than the population average. For female patients, the patients who have a circulatory disease, are 1.92 times more

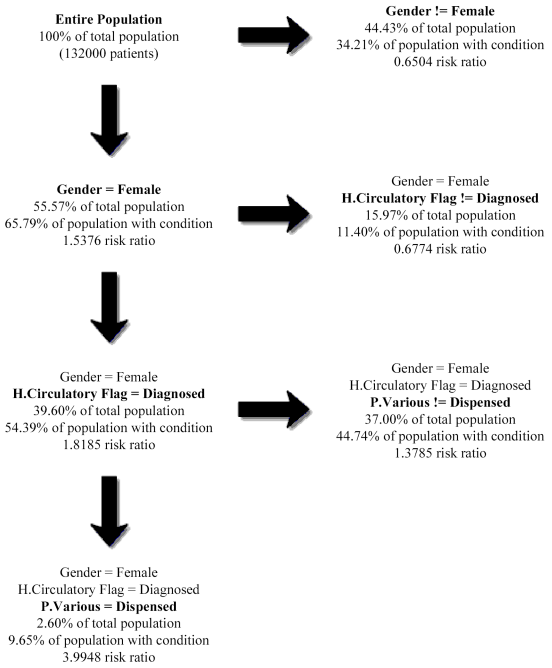

Fig. 2. The first tree presentation of Rule 1

likely to develop angioedema than other female patients. The female patients with a circulatory disease, and who have used drugs in the "Various" category are 3.06 times more likely than female patients with a circulatory disease but not taking drugs in that category. However, we need to keep in mind that the rule presentation can help doctors to be alert in prescribing medicine to patients with certain characteristics. The indication becomes more complex when patients have multiple diseases such as asthma and diabetes etc.

5 Conclusion

In this paper, we have applied a modified association classification algorithm to health data to explore risk factors associated with adverse drug reactions. We assessed the discovered rules using survival charts and introduced two tree-type presentations to present risk factors in a comprehensible way. The tree presentations are able to demonstrate the heightened risks due to a combination of risk factors as well as due to a single risk factor. Thus, they provide an effective way for medical practitioners to interpret clearly the risk factors for prescribing certain drugs to specific patient sub-groups. The consequence of this should be more effective use of medicines and reduced morbidities or costs from adverse drug events. Such knowledge could be readily implemented in electronic prescribing systems.

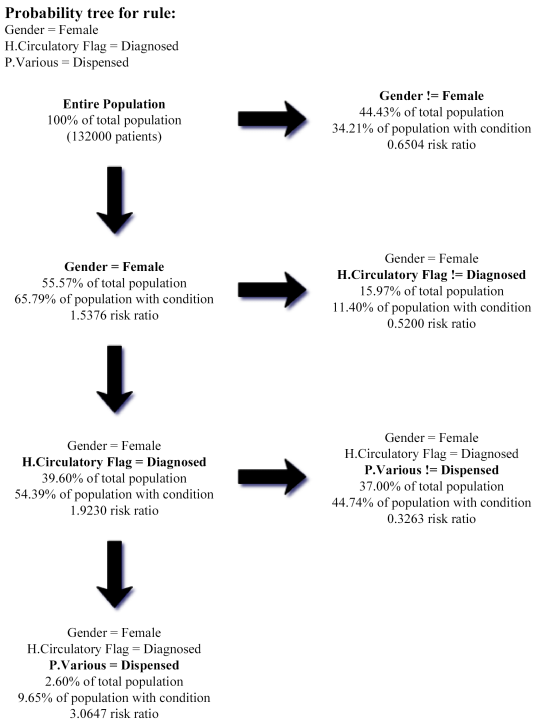

Fig. 3. The second tree presentation of Rule 1

References

1. David M. Fram, June S. Almenoff, and William DuMouchel. Empirical bayesian data mining for discovering patterns in post-marketing drug safety. In *Proceedings of KDD 2003*, pages 359–368, 2003.
2. Harvey J. Murff, Vimla L. Patel, George Hripcsak, and David W. Bates. Detecting adverse events for patient safety research: a review of current methodologies. *Journal of Biomedical Informatics*, 36(1/2):131–143, 2003.
3. M. Reid, B. Euerle, and M. Bollinger. Angioedema, 2002. http://www.emedicine.com/med/topic135.htm.
4. J. Li, H. Shen, and R. Topor. Mining the optimal class association rule set. *Knowledge-Based Systems*, 15(7):399–405, 2002.
5. Stephen C. Newman. *Biostatistical Methods in Epidemiology.* John Wiley & Sons, July 2001.
6. Graham Williams, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay, Ross Sparks, Jisheng Cui, Simon Hawkins, and Chris Kelman. QLDS: Adverse drug reaction detection towards automation. Technical Report CMIS 04/91, CSIRO Mathematical and Information Sciences, Canberra, 2004.