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]

[*] Corresponding author

[1]CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra ACT 2601, Australia
Email: Firstname.Lastname@csiro.au

[2]Current address: Australian Taxation Office, Australia
Email: Graham.Williams@togaware.com

[3] National Centre for Epidemiology and Population Health, the Australian National University, Australia
Email: Chris.Kelman@anu.edu.au

[4]Department of Mathematics and Computing, University of South Queensland, Australia
Email: jiuyong@usq.edu.au

Abstract. An association classification algorithm has been used to explore adverse drug reactions in a large medical transaction data set with unbalanced classes. Rules discovered can be used to alert medical practitioners, when prescribing drugs to certain categories of patients, potential adverse effects. It is essential to present these rules to the medical practitioners in a form which is easy for them to understand and interpret. We assess the rules identified by the association classification algorithm using survival charts and propose two kinds of probability trees to present them. Both of them present the risk to the given adverse drug reaction of certain categories of patients in terms of risk ratios, which are familiar to medical practitioners. The first shows risk ratios when all rule conditions applied. The second presents the risk associated with a single risk factor with other parts of the rule identifying the cohort of the patient subpopulation. 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, the presentations can interpret clearly the risk factors of adverse drug reactions to medical practitioners.

1 Introduction

Data mining aims to discover previously unknown, potentially useful, understandable knowledge from large data sources. Data mining does not

provide us with benefits until we can understand how the method works and what the generated information means. Then we can evaluate this information and translate it into actionable solutions to problems. Since data mining usually involves extracting "hidden" knowledge (rules or patterns) from a database, understanding and evaluating the discovered patterns become more important and challenging [3, 6] , especially in health application [12].

Systematic monitoring of adverse drug reactions is important for both financial and social reasons. In general, the early detection of unexpected adverse drug reactions relies on a local spontaneous reporting system and collated statistics from overseas agencies. At present, adverse reactions resulting from new medications and their interactions with other medicines are detected only if they are either dramatic or common [9]. When a new drug is introduced, it is likely that unexpected side-effects will go unnoticed until a very substantial number of patients have been adversely affected. Spontaneous adverse event reporting databases are traditional data sources for most data mining work [4, 9, 5, 16], which focus on the generation of drug-event associations. However, the availability of a population-based prescribing data set, 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 addresses this data mining problem, where the objective here is to identify the factors, which increase the risk of the adverse drug reaction, directly from large linked health data rather than spontaneous adverse event reporting databases.

Prescribed drugs are recorded in PBS data using the WHO code, based on the Anatomical and Therapeutic Classification (ATC) system adopted by the World Health Organisation (WHO). Adverse reactions events are recorded in hospital data using ICD-9 code (International Classification of Diseases, Ninth Revision). Three case studies have been identified by our adviser, the Therapeutic Goods Administration, Australia. They are: 1) ACE inhibitors [1] usage associated with Angioedema. This case will serve as the main example to illustrate our method in this paper;

[1] ACE inhibitors are used to treat congestive heart failure (CHF) and high blood pressure (hypertension). Angioedema is a swelling (large welts or weals), that occurs beneath the skin rather than on the surface [11]. 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 [11].

2) Alendronate usage associated with Esophagitus; 3)Nefazodone usage associated with a number of Hepatitis conditions.

In our data, the distribution of classes with and without adverse events is highly unbalanced due to the intrinsic nature of adverse drug reactions (rare events in the administrative health data [2]). 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 [7] to discover rules which identify patient subgroups with a high proportion of patients with target events. The rules discovered by the association classification are assessed by using survival charts. We further propose two kinds of tree representation for mined rules to help them and potential users to gain understanding of the rules. To the best of our knowledge, there is no similar presentation of mined rules in the literature.

Organisation. The rest of the paper is organised as follows. Section 2 discusses our method to mine association classification rules from unbalanced health data. Section 3 describes the data set and features selected for the mining process. Section 4 reports mined rules validated by survival chart, and Section 4.2 presents two kinds of probability tree presentation of rules. Conclusion and a discussion complete the paper in Section 5.

2 Association Classification for Unbalanced Classes

In contrast to previous work [4, 5, 9, 16], the objective here is to discover rules which identify patient subgroups with a high proportion of patients with adverse drug reactions events, directly from large linked health data rather than spontaneous adverse event reporting databases. It is important to note that our linked data set has highly unbalanced classes as adverse drug reactions are rare events in the dataset (e.g., 116 patients with angioedema versus 131,884 without angioedema). Traditional 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) i.e. the class with adverse drug reaction events is given little attention by the approaches. In this paper, we modify the Optimal Class Association Rule Mining Algorithm [7] to identify higher risk patient groups of adverse drug reaction events.

To tackle this problem, we introduce local support and risk ratio (as defined by Equations 1 and 2) to discover rules that identify cohorts in which the risk of occurrence of the rare events is high. The support in minor class is called *local support* defined by Equation 1.

$$lsup(A \rightarrow c) = \frac{sup(A \rightarrow c)}{sup(c)} \tag{1}$$

Here $sup(c)$ and $sup(A \rightarrow c)$ represent the support (or proportion or relative frequency) of Class c in the whole population and the support of pattern A in Class c respectively. Local support is called minimum coverage in [1] and multiple minimum support in [8]. 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 *Risk Ratio* as interesting measure for patterns mining, which is represented by Equation 2.

$$RR(A \rightarrow c) = \frac{lsup(A \rightarrow c)sup(\overline{A})}{lsup(\overline{A} \rightarrow c)sup(A)} \tag{2}$$

The risk ratio defines the relative risk (belonging to Class 1) of the patients identified by rule A with respect to the majority patients [10, p. 35]. $\overline{A}$ denotes the absence of pattern A. For example, A defines patients in age group $[30, 40)$, then $\overline{A}$ defines patients outside the age group. Tan et al. [13] discuss the properties and their consistence of 21 existing objective interestingness measures in a framework of contingency table and statistical dependencies of two items. Risk ratio used here as the objective interestingness measure is not mentioned but it is common in health application and meaningful to medical practitioners.

3 Data Preparation and Feature Selection

We use the Queensland Linked Data Set(QLDS) [15] — a medical transaction data for the rule mining in this study. This data set has been made available under an agreement between Queensland Health and the Australian Department of Health and Ageing (DoHA). This data set links de-identified patient level hospital separation data (for the period between 1 July 1995 and 30 June 1999), Medicare Benefits Scheme (MBS) data, and Pharmaceutical Benefits Scheme (PBS) data (1 January 1995 to 31 December 1999) in Queensland. Each record in the hospital data corresponds to one in-patient episode. Each record in MBS corresponds

Table 1. List of variables used for association classification

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

to one MBS service for one patient. Similarly, each record in PBS corresponds to one prescription service for one patient. As a result, each patient may have more than one hospital, or MBS or PBS record. Each patient is assigned to a unique identifier, making it possible to link the records of each patient in three separate data sets.

For the implementation of the mining task, we need to extract profile data for all patients exposed to the drug of interest in a 180 day window, which is determined by 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 inhibitor) 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.

Table 2. Discretisation of continuous variables.

Variable	Groups	Description
Age	4	0-19, 20-39, 40-59, 60+
Bed days	3	$\leq$ 2, 3-14, $\geq$ 15
Total scripts	3	0, 1, $\geq$ 2

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 drug) were extracted. The "total number of 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 chooses some variables in the profiles in applying association classification algorithm. As the number of variables and possible values of the variables decrease, the run time of the algorithm will decrease. The algorithm requires all the variables take only a set of discrete values. There are many ways to discretise the continuous variables. For the sake of understandability and simplicity, we use cutoff values to discretise the variables. Table 2 lists the cutoff values used for continuous variables.

4 Representing Association Classification Rules

4.1 Case study - ACE inhibitor and angioedema

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. We could not present hundreds of rules to medical experts for the 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 data set 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 inhibitor 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

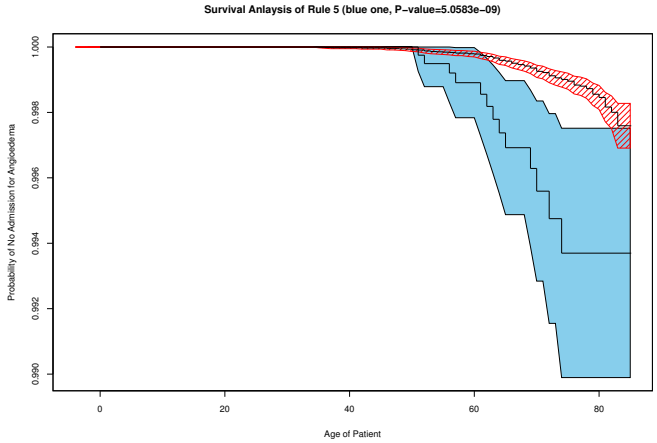

Fig. 1. Fleming-Harrington survival analysis of Rule 5 for the ACE inhibitor and angioedema combination.

  - Usage of drugs in category of "Nervous system" = No

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

where RR indicates the risk ratio.

For each rule discovered, we conduct further evaluation, e.g., the survival analysis and its significance test [10, pp. 159-169]. The survival analysis is concerned with the modeling of 'lifetime' data. We estimate the survivor function $S(t)$, simply the probability of surviving beyond time t, to distinguish the subgroup described by the rule from the others. In addition, we use log-rank test, a formal measure of the strength of evidence that two populations have different lifetimes. It is likely to detect a difference between groups when the survival curve is consistently higher for one group than another. A rule is statistically significant at the 0.01 level if its P-value is less than 0.01.

We exemplify the survival analysis on one rule. Figure 1 presents the estimated survivor functions of the subgroup described by Rule 5 (the one within the filled (blue) region) and the other patients (within the shaded (red) region). The filled (blue) region and the shaded (red) region indicates 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 is much lower than 0.01. It also 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 [14].

4.2 Tree presentations

The rules identified by the association classification algorithm provide useful knowledge to the medical practitioners, and can serve as a reference in their prescription of drugs to the patients. The patients' characteristics can be compared to the rules to evaluate their risk to the suspected adverse drug reaction. However, the rules presented above may not provide enough information to the medical practitioners. The further breakdown of the risks caused by individual risk factors provide 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 the risk ratio is presented on each node. The branch to the right of the node list the information for complementary population. The level down of each node gives another split of population using a new variable value pair. Figure 2 shows an example. The Rule 1 is presented as a tree with risk ratio (RR) as the main measure. The risk ratio is defined by Equation 2.

According to Figure 2, female users of ACE inhibitors are 1.5376 times more likely to have angioedema than the population average. For those female patients who have a circulatory disease, the likelihood increases to 1.8185. For those who are female, have a circulatory disease, and also have taken drugs falling in Various category(the 14th ATC level-1 drug category), the likelihood increases further to 3.9948. The tree presentation highlights how the risk ratio changes by 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 in Equation 3

$$RR(A \to C \mid U) = \frac{lsup(A \bigcap U \to C) sup(\overline{A} \bigcap U)}{lsup(\overline{A} \bigcap U \to C) sup(A \bigcap U)} \tag{3}$$

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 Figure 3. According to Figure 3, female users

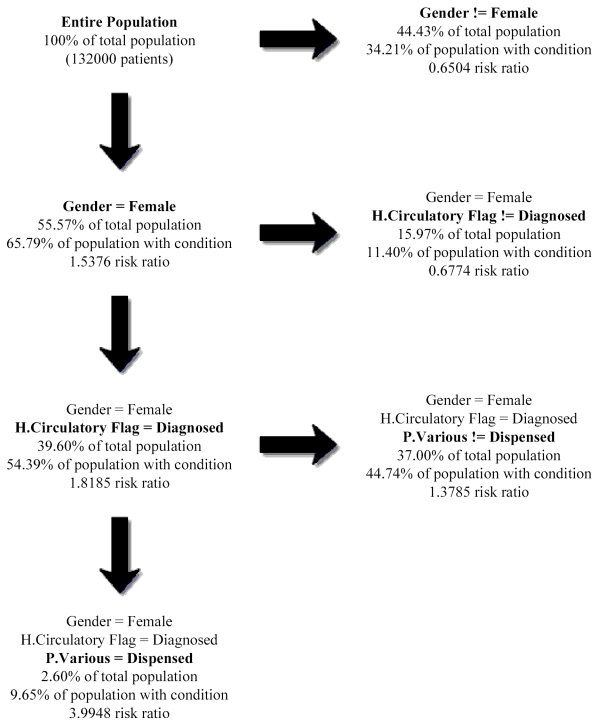

Fig. 2. The first tree presentation of Rule 1 for the ACE inhibitor and angioedema case study.

of ACE inhibitors are 1.5376 times more likely to have angioedema than the population average. For female patients, the patients who have a circulatory disease, are 1.923 times more likely to develop angioedema than other female patients. The female patients with a circulatory disease, and who have used drugs in Various category (the 14th ATC level-1 drug category) are 3.0647 time more likely than female patients with a circulatory disease but not taking drugs in that category.

Similar to Figure 2, Figure 4 represents Rule 1 (the rule with the highest risk ratio) for the alendronate and esophagitus case study. Clearly, users of alendronate aged 40-59 are 1.1026 times more likely to have esophagitis than the population average. For those patients aged 40-59 who have used drugs falling in the category of Alimentary tract metabolism, the likelihood increases to 1.6451. For those who are aged 40-59, have used drugs falling in the category of Alimentary tract metabolism, and also have taken drugs falling category of Cardiovascular systems, the likelihood increases further to 2.4607. This example illustrates that the pro-

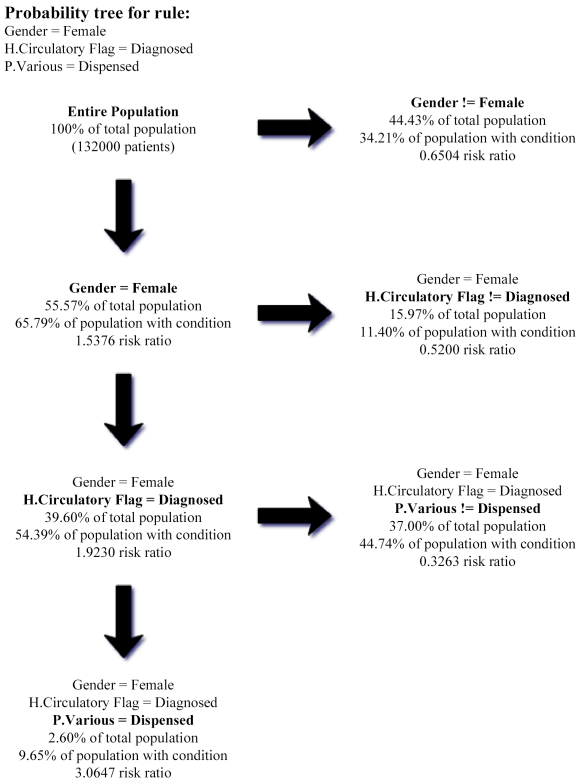

Fig. 3. The second tree presentation of Rule 1 for the ACE inhibitor and angioedema case study.

posed tree presentations are also suitable for the other two case studies, and more examples can be found in [14].

5 Discussion and Conclusions

In this paper, we have applied an modified association classification algorithm to health data to explore risk factors associated with adverse drug reactions. Due to the nature of the problem, our association classification mining method employs the local support and the risk ratio in order to identify risk groups of patients with unbalanced classes.

We assessed the discovered rules using survival charts and introduced two tree-type presentations to present risk factors in a comprehensible way, and demonstrated them on mining three adverse drug reactions. The first shows risk ratios when all rule conditions applied. The second

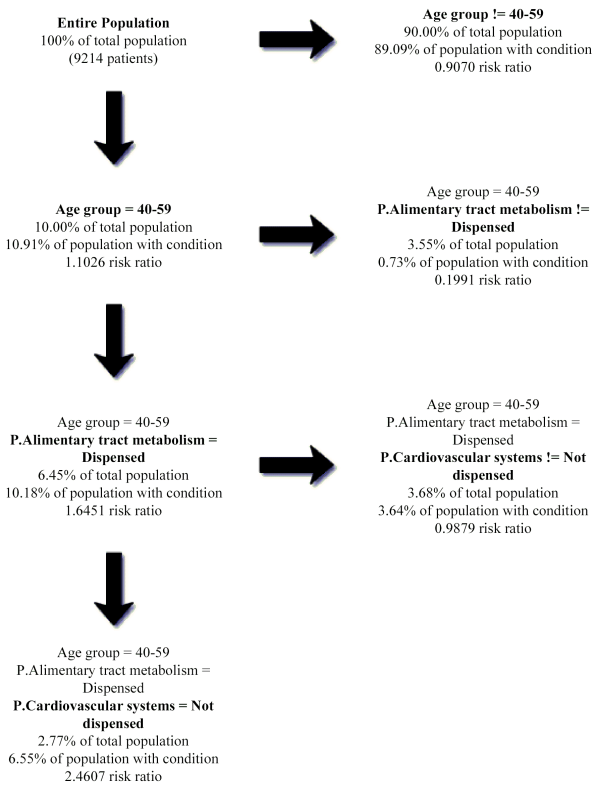

Fig. 4. The first tree presentation of Rule 1 for the alendronate and esophagitus case study.

presents the risk associated with a single risk factor with other parts of the rule identifying the cohort of the patient subpopulation. 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 of adverse drug reactions.

Acknowledgements

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

References

1. R. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense database. In *Proceedings of the 15th International Conference on Data Engineering*, pages 188–197, 1999.
2. J. Chen, H. He, G. Williams, and H. Jin. 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)*, pages 235–239, Sydney, Australia, May 2004.
3. U. Fayyad, G. Grinstein, and A. Wierse, editors. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufman, San Francisco, CA, USA, 2002.
4. D. M. Fram, J. S. Almenoff, and W. DuMouchel. Empirical bayesian data mining for discovering patterns in post-marketing drug safety. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 359–368. ACM Press, 2003.
5. J. Harvey, C. Turville, and S. Barty. Data mining of the australian adverse drug reactions database: a comparison of bayesian and other statistical indicators. *International Transactions in Operational Research*, 11(4):419–433, 2004.
6. H.-D. Jin, W. Shum, K.-S. Leung, and M.-L. Wong. Expanding self-organizing map for data visualization and cluster analysis. *Information Sciences*, 163:157–173, Jun. 2004.
7. J. Li, H. Shen, and R. Topor. Mining the optimal class association rule set. *Knowledge-Based Systems*, 15(7):399–405, 2002.
8. B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 337–341, N.Y., 1999. ACM Press.
9. H. J. Murff, V. L. Patel, G. Hripcsak, and D. W. Bates. Detecting adverse events for patient safety research: a review of current methodologies. *Journal of Biomedical Informatics*, 36(1/2):131–143, 2003.
10. S. C. Newman. *Biostatistical Methods in Epidemiology*. John Wiley & Sons, July 2001.
11. M. Reid, B. Euerle, and M. Bollinger. Angioedema, 2002. http://www.emedicine.com/med/topic135.htm.
12. J. Roddick, P. Fule, and W. Graco. Exploratory medical knowledge discovery : Experiences and issues. *SIGKDD Exploration*, 5(1):94–99, 2003.
13. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 32–41. ACM Press, 2002.
14. G. Williams, H. He, J. Chen, H. Jin, D. McAullay, R. Sparks, J. Cui, S. Hawkins, and C. Kelman. QLDS: Adverse drug reaction detection towards automation. Technical Report CMIS 04/91, CSIRO Mathematical and Information Sciences, Canberra, 2004.
15. G. Williams, D. Vickers, R. Baxter, S. Hawkins, C. Kelman, R. Solon, H. He, and L. Gu. The Queensland Linked Data Set. Technical Report CMIS 02/21, CSIRO, Canberra, 2002.
16. A. Wilson, L. Thabane, and A. Holbrook. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, 57(2):127–134, 2004.