A Delivery Framework for Health Data Mining and Analytics

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

[*] Corresponding author
Email: Damien.McAullay@csiro.au

[1]CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra ACT 2601, Australia
Email: Damien.McAullay, Jie.Chen, Warren.Jin, Hongxing.He, Ross.Sparks@csiro.au

[2]Current address: Australian Taxation Office
Email: gjw@togaware.com

[3]Department of Health and Ageing (DoHA)
Email: Chris.Kelman@anu.edu.au

Abstract

The *i*Health Explorer tool, developed by CSIRO and DoHA, delivers web services type data mining and analytic facilities over a web interface, providing desktop access to sophisticated analyses over very large data collections. The tool allows users to access large transactional datasets to create profiles of selected patients. The patients' profiles, together with windowed event sequences data, can then be analyzed using a user chosen data mining tool. The results of the analysis can then be visualized using various forms of knowledge representation methods. Although the initial implementation of the tool is focused on its application in adverse drug reaction exploration, this tool and the embedded data mining tools can be used in broad areas of health data analysis.

Keywords: web services, health data, adverse drug reaction, data mining, record linkage, association, classification

1 Introduction

As storage capacity increases exponentially, more and more transactional data is being collected by various industries. The health industry is no exception. For example, there are 5.7 million hospital admissions, 210 million doctor's visits, and a similar number of prescribed medicines dispensed in Australia annually. Records of all of the above listed transactions are captured electronically. There are trillions of medical records stored world wide every year. Unfortunately, such valuable available data is not being appropriately utilized to provide useful evidence as a basis for future medical practice. Easy to use and readily available analysis tools are urgently needed (Roddick, Fule & Graco 2003, Cios & Moore 2002). Data mining techniques have been applied for health administration (Kum, Duncan, Flair & Wang 2003, Rao, Sandilya, Niculescu, Germond & Rao 2003), adverse drug reactions (Murff, Patel, Hripcsak & Bates 2003, Harvey, Turville & Barty 2004, Chen, He, Williams & Jin 2004) and drug safety (Almenoff, DuMouchel, Kindman, Yang & Fram 2003, Wilson, Thabane & Holbrook 2004), and there are also concerns about network architecture for health research data analysis (Taylor, O'Keefe, Colton, Baxter, Sparks, Srinivasan, Cameron & Lefort 2004).

In this paper, our attempt towards a web based health data analysis tool with an intuitive and efficient graphical user interface is described. The initial implementation is focused on adverse drug reaction exploration. This will be extended to allow users to explore health data for other objectives.

An important part of any system or tool is the delivery framework; the way in which the user comes about receiving and is presented with the content. This is especially true in the area of analytics, where it is not uncommon for the user to be confronted with summaries and analysis of large amounts of data.

We present a tool that was developed by CSIRO and DoHA called *i*Health Explorer as an example of a delivery framework for analytics. In this particular case, the tool is designed to perform summary and analysis operations on large amounts of health data, and to display the results to the user in a comprehensible manner.

*i*Health Explorer is a data exploration tool designed for use with health data. The tool provides the user with a web based interface to a suite of summarisation and analysis methods, including data mining methods. The tool allows the user to generate subsets of the main dataset, and perform further analyses on these subsets. Various visualisation methods are available to better convey the results to the user.

The web-based approach to content delivery has several merits.

- Distribution: Web based content is easily distributed across the Internet, and therefore around the world.
- Security: Modern Internet architecture allows easy security implementation for isolated, intranet, extranet, or Internet access.

The user interface has been designed with the assistance of medical professionals that are representative of the end-users. The graphs and other visual aids have been designed to be most understandable by these users.

At present, the tool is primarily designed for the discovery of adverse drug reactions. Adverse drug re-

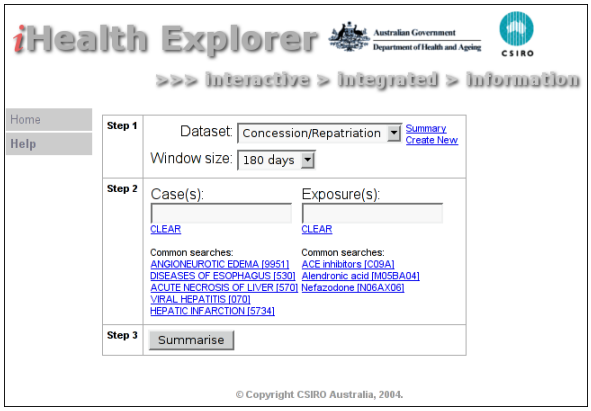

Figure 1: Tool front-end

actions are the unintended medical consequences of treatment with pharmaceuticals. They can vary in severity between minor nuisance and death (Murff et al. 2003). These reactions may or may not be previously known. *i*Health Explorer can be used to discover and rank the risk of such unintended outcomes. The software can thus make a risk assessment of selected adverse reactions, identify associated risk factors and generate hypotheses for further clinical testing.

2 *i*Health Explorer

*i*Health Explorer is a tool that provides researchers, and regulators with a unified interface for viewing and analysing large amounts of electronic health data. *i*Health Explorer provides a suite of tools that allow the user to summarise the data for easier interpretation, and to perform complex data mining analyses to discover, for example, adverse reactions to particular drugs. The web-based delivery platform allows for easy distribution across large organisations and government departments. Data mining methods such as association and classification analyses have been employed. It is envisaged that this system will be further developed to perform a wide range of data mining analyses on many different sources of health data.

2.1 Tool Front-end

This first page provides the user with an optimised interface for performing different types of analyses. The system will determine the appropriate analyses from the user's choices.

- Choosing general system options.
  - Dataset: allows the user to choose the dataset to use as the source of the data.
  - Window size: allows the user to choose the number of days to use as the data window.
- Case and/or exposure selection

  The choice made here will determine what type of analyses are performed. E.g. if both a case and an exposure is chosen, a cross-analysis on the case and exposure will be performed. Cases are indicated by an identifier followed by a code. The identifier indicates what type of code will follow. For example, in "D:9951" the "D" indicates that the code will be an ICD9 code, and "9951" is the code for Angioneurotic Edema (Angioedema) (Reid, Euerle & Bollinger 2002). Exposures are indicated in the same way. For example, in "W:C09A" the "W" indicates that the code is a WHO code and the "C09A" is the code for ACE Inhibitors.

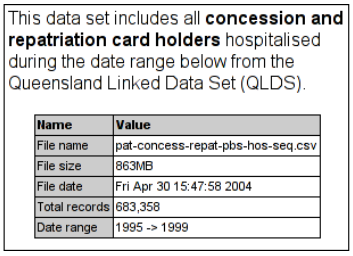

Figure 2: Dataset summary

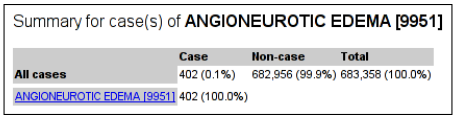

Figure 3: Case analysis

2.1.1 Dataset Summary

Summary information about each dataset is available. This feature displays information about the current dataset, such as file name, date, size, and number of records, as shown in Figure 2.

2.1.2 Create New Dataset

This feature will be used to create new datasets, defined by a set of constraints defined by the user. Constraints will include options such as which features to include, date range of records to include, and records matching particular criteria.

2.2 Case Analysis

When the user selects only a case to be analysed, the case analysis summary page will be displayed as shown in Figure 3. The table presented here contains a summary of the selected cases within the dataset. The first row in the table shows the number of cases and non-cases, that is, the patients who have been admitted for the selected case, and the patients who have not. If a high level code (3 digit) was chosen, sub-codes will be included. For example, if the user chose "530" as the case, "5301", "5302", etc. will be included.

The remaining rows in the table show the breakdown of individual sub-codes that were found in the dataset.

Short descriptions are available for all types of cases.

Further analyses are listed in the menu bar.

2.2.1 Association Analysis

Association analysis uses a tool called Magnum Opus (Webb 2000) to discover exposures that cause a high risk of being admitted into hospital for the chosen case.

All of the patterns are listed for particular age and gender groups, as shown in Figure 4.

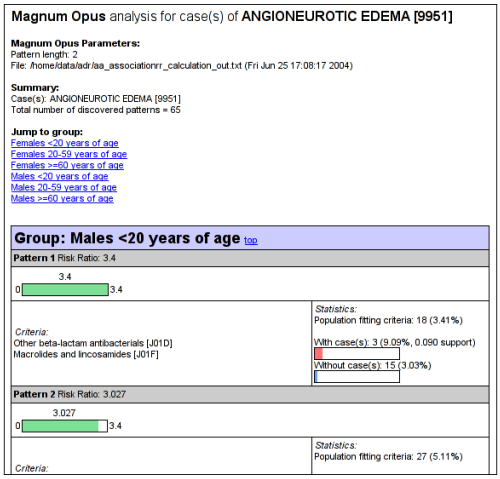

Figure 4: Association analysis

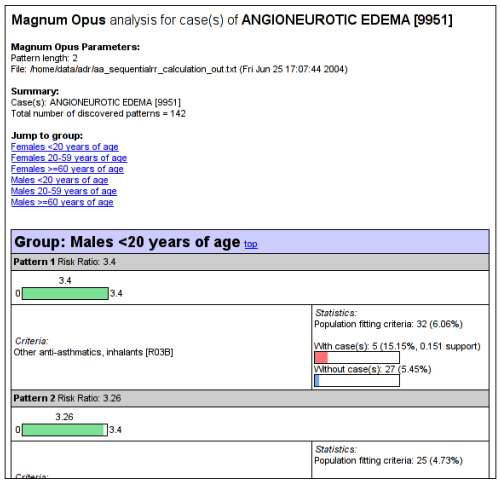

Figure 5: Sequence analysis

Patterns are sorted by their risk factor (called Odds Ratio, which is defined in Equation 3). The higher odds ratio means that patients fitting the pattern are more likely to be admitted to hospital for the chosen case.

2.2.2 Sequence Analysis

Sequence analysis uses the tool called SLP-Miner (Seno & Karypis 2002) to discover sequences of exposures that cause a high risk of being admitted to hospital for the chosen case.

All of the patterns are listed for particular age and gender groups, as shown in Figure 5.

Patterns are sorted by their risk factor (Odds Ratio). A higher odds ratio means that patients fitting the pattern are more likely to be admitted to hospital for the chosen case.

2.3 Exposure Analysis

When the user selects only an exposure to be analysed, the exposure analysis summary page will be displayed as shown in Figure 6. The table presented on this page contains a summary of the selected exposure within the dataset. The first row in the table shows the number of exposures and non-exposures, that is, the patients who have experienced the selected exposure, and the patients who have not. If a high level exposure was chosen, sub-codes will be included. For example, if the user chose "C09A" as the exposure, "C09AA01", "C09AA02", etc. will be included.

The remaining rows in the table show the breakdown of individual sub-exposures that were found in the dataset.

Short descriptions are available for all types of exposures.

2.4 Case and Exposure Analysis

When the user selects a case and an exposure to analyse, the case and exposure analysis page will be displayed as shown in Figure 7.

The table shows the number of patients that have been exposed to the chosen exposure, admitted for the chosen case, exposed and admitted, and neither. The table also shows the row and column totals (total exposures, non-exposures, cases, non-cases, and population).

Comments on the chosen case and exposure combination by other system users may be displayed if they are available.

Graphs are displayed to summarise the population. Age, gender, total number of scripts, and temporal information are represented.

For example, if ace-inhibitor is chosen as exposure and angioedema is chosen as case, the age distribution can be displayed as in Figure 8.

Figure 8 shows that the adverse reaction happened to the patients of age ranged 20+ with median around 65. This is in excellent agreement with the cases reported to TGA by 1996. Further analyses are listed in the menu bar.

2.4.1 Classification Analysis

The objective of the classification analysis is to identify the factors which increase the risk of some adverse drug reaction. We use the association classification algorithm (Li, Shen & Topor 2002, Gu, Li, He, Williams, Hawkins & Kelman 2003) to find rules (or patterns) which identify patient subgroups with a high proportion of patients with target events. This approach can be used to determine higher risk patient groups. The output is in the form of rules, that consist of patient attributes that lead to a high risk of being admitted to hospital for the chosen cases.

As shown in Figure 9, rules are sorted by their risk factor (called heuristic risk ratio, see, e.g., Equation 4 in Section 3.2). The higher heuristic risk ratio means that patients fitting the rule are more likely to be admitted to hospital for the chosen cases when they have been exposed to the chosen exposure (Newman 2001). Each rule has a survival graph and probability tree displayed with it.

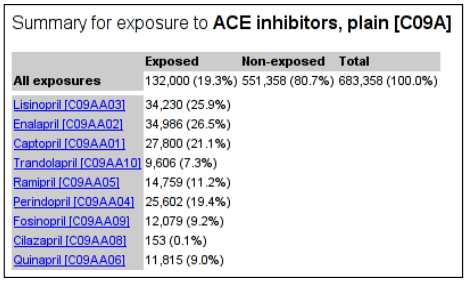

Figure 6: Exposure analysis

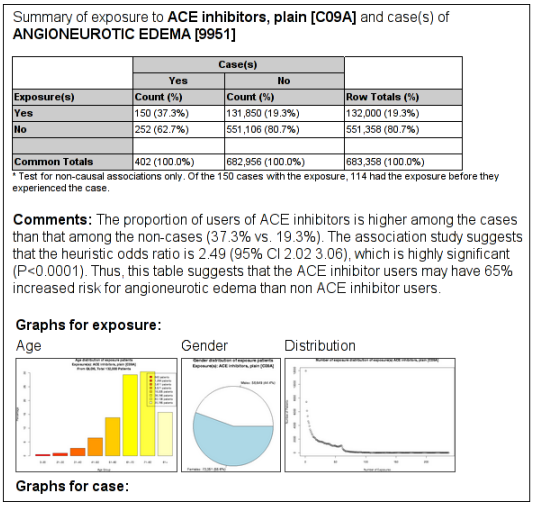

Figure 7: Case and exposure analysis

We take one rule discovered for the ACE inhibitor and Angioedema combination as an example to illustrate the functionality of *i*Health Explorer.

Rule 6: Risk Ratio = 3.3

- Alimentary tract metabolism = No
- Genito urinary system and sex hormones = Yes
- General anti-infectives for systematic use = Yes,

where Risk Ratio indicates the heuristic risk ratio.

This rule describes a sub-group of patients who have not claimed any Pharmaceutical Benefit Scheme (PBS) items (such as drugs) falling in the category of "Alimentary tract metabolism", have claimed PBS items falling in categories of "General anti-infective for systematic use and Genito urinary system and sex hormones" during the period of study. There are 13 cases among 114 that satisfy Rule 6, and 5,000 non-cases among 131,886 satisfied Rule 6. The risk ratio is 3.3.

Figure 10 presents the estimated survivor functions of the subgroup described by Rule 6 (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

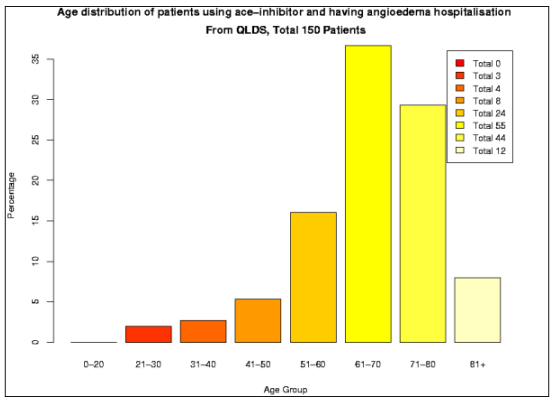

Figure 8: Age distribution of patients with angioedema and exposed to ace-inhibitor

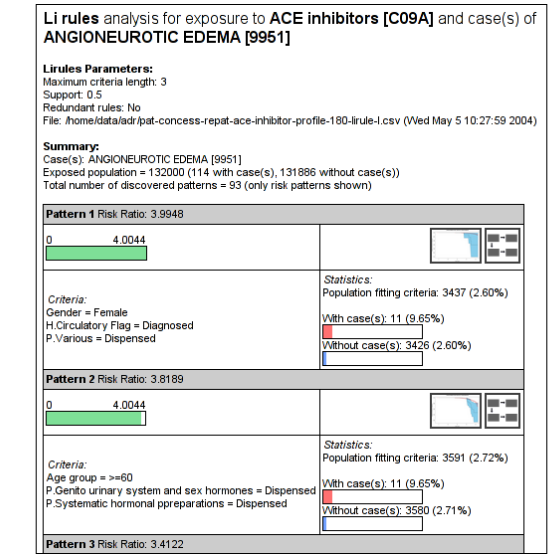

Figure 9: Classification analysis

Rule 6 has significantly higher probability of hospital admission for Angioedema than the other patients. Moreover, we conduct the log-rank test which is likely to detect a difference between groups when the survival curve is consistently higher for one group than another. 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 6 is overwhelmingly different from the other patients.

3 Data Mining Methods

3.1 Association

The aim of associations analysis is to discover the associations between some drug usages and the specified diseases without prior knowledge. The associations discovery in the context of Adverse Drug Reactions can be described as a two phases approach. The first one is focused on pattern generation. The second one is focused on pattern evaluation. For pattern generation, the windowed PBS sequences of case patients are mined to yield association and sequential patterns of drugs. Then case control matched populations are exploited for pattern evaluation (Williams, He, Chen, Jin, McAullay, Sparks, Cui, Hawkins & Kelman 2004).

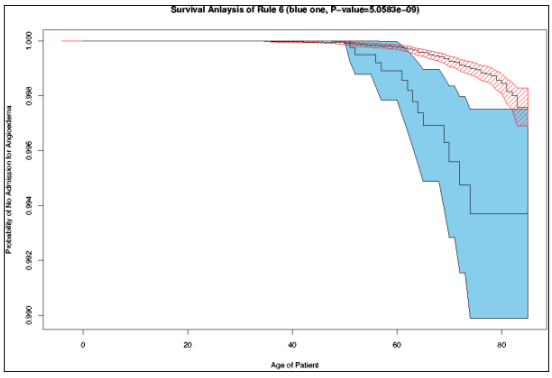

Figure 10: Fleming-Harrington survival analysis of Rule 6 for the ACE inhibitor and Angioedema combination.

We define cases as patients with the target hospitalisation event, and non-cases as patients without (but with any other hospitalisation event). For each case, we find N matching "controls", where the gender, age group, and month of admission of the first hospitalisation event of the control matches the gender, age group, and month of admission of the first target hospitalisation event of a case. Cases and non-cases that have a first month of admission before 1 July 1995 and after 30 June 1999 are excluded (these are the limits of the data used). We gathered 15 controls per case for Angioedema and Hepatitis, and 4 controls per case for Esophagitis due to the large number of case patients.

After the match process by gender, age and hospitalisation month, both the case and control populations are partitioned into six gender and age groups i.e. males < 20 years of age, males $20 - 59$ years of age, males ≥ 60 years of age, females < 20 years of age, females $20 - 59$ years of age and females ≥ 60 years of age.

Note that there are 2, 137 and 12 patients for the three studies respectively dropped because they have the hospitalisation events outside of the regular four year window.

Note that hospitalisation month is ignored in this partition, because the partitioned populations are quite small for both Angioedema and Hepatitis cases, and it makes frequent patterns mining over such partitioned populations meaningless. As a result, we mine frequent patterns directly from the windowed sequences of case population for a gender and age group, it means that we put the windowed sequences of case patients with different hospitalisation months together and ignore the differences among hospitalisation months.

For each association or sequential pattern mined from the case population for a gender and age group, a list of contingency tables illustrated by Table 1 are derived for the gender and age group, where k is the index of hospitalisation month. This 2×2 table denotes the frequencies with or without a pattern in both case and control population in the kth hospitalisation month for the gender and age group. Then Cochran-Mantel-Haenszel Chi-Squared Test is exploited to calculate odds ratios (Williams et al. 2004). The estimation of a common odds ratio is calculated by the equation given in (Agresti 1990, 230-235), where checking the common odds ratio assumption is important.

	Disease(Yes)	**Disease(No)**
Pattern(Yes)	n_{11k}	n_{12k}
Pattern(No)	n_{21k}	n_{22k}

Table 1: Contingency table

For the three case studies, both association and sequence patterns are mined by using Magnum Opus and SLPMiner respectively. The association and sequences analysis on the three case studies are performed without any prior knowledge about the association of drugs and hospitalisation.

3.2 Classification

Adverse drug reactions are rare events. In the classification task, the patients with the target event are classified as class 1 and other patients are class 0 patients. As class 1 is a very small class, commonly applied classification methods, where the purpose is to achieve high classification accuracy, do not work well. The objective here is to identify the factors which increase the risk of the adverse drug reaction. We use the Association Classification Algorithm (Li et al. 2002) to find rules which identify patient subgroups with a high proportion of patients with target events. This approach can be used to determine higher risk patient groups. The algorithm developed by Li *et al* (Li et al. 2002) generates the optimal class association rule set. Their recent work shows that the optimal class rule set achieves a very high classification accuracy.

However, our data set has very unbalanced classes (e.g., 116 patients with angioedema versus 131,884 without angioedema, in the case of the AA data set). Our main interest is in finding rules (or cohorts) which lead to higher occurrences of the target patients than the average occurrence. 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 adverse drug reaction data, the class of interest is given little attention by the algorithm. A new approach has been developed to handle this situation, typical of adverse drug events, where the key events of interest occur infrequently in the data. The new approach uses interestingness and local support measures (as defined by Equations 1 and 2) to overcome this problem. It aims to identify rules that identify cohorts in which the occurrence of the rare event is high. As a result, the new approach increases classification accuracy of Class 1 patients. For the sake of simplicity, the new algorithm is referred to as LiRule hereafter. "Local support" is 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.

Another criterion used in our rule discovery is "interestingness" as defined by Equation 2.

$$Interestingness(A \rightarrow C) = \frac{lsup(\rightarrow C)}{\sum_i (lsup(A \rightarrow C_i)} \tag{2}$$

The algorithm will output rules which give high "risk ratio" or "odds ratio" values for Class 1. When the rules or patterns are identified, a contingency table is formed as shown in Table 2.

	Class C	**Not class C**	**Total**
Pattern A	a	b	a+b
Not Pattern A	c	d	c+d

Table 2: Contingency table

The odds ratio is then defined in Equation 3.

$$OR(A \rightarrow C) = \frac{ad}{bc} \tag{3}$$

Similarly the risk ratio is defined in Equation 4.

$$RR(A \rightarrow C) = \frac{a(c+d)}{c(a+b)} \tag{4}$$

As in our study the population of class C is very small (unbalanced classes), therefore $a << b$ and $c << d$, the RR is very close to OR.

The developed new algorithm has applied to the case studies assigned by our domain experts, and part of the output rules have been evaluated by using their domain knowledge and the survival analysis as well.

4 Health Data (QLDS)

The Queensland Linked Data Set has been used as the data set in the initial test of concept phase. 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 inpatient episode. Each record in MBS corresponds 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 the three separate data sets (or tables). The data were linked using Medicare numbers with particular attention to privacy protection and used encryption of identifying keys.

The QLDS contains records for 1,176,294 individuals who were hospitalised in Queensland between 1995 and 1999. The QLDS therefore roughly covers 35 percent of the Queensland population. The QLDS contains 3,087,454 hospital records, but due to missing Medicare numbers in some hospital records, this does not represent every record for that period of time. The QLDS also contains the Medicare and pharmaceutical records for these 1,176,294 individuals as listed in Table 3 .

Hospital	
Records	*Individuals*
3,087,454	1,176,294
Medicare	
Records	*Individuals*
100,738,822	1,169,471
Pharmaceutical	
Records	*Individuals*
61,431,235	733,335

Table 3: QLDS Summary

5 Conclusion and Future Work

We have demonstrated an early implementation of a health data analysis tool with a user-friendly interface. The tool enables health researchers, regulators or other interested users to access health data and data mining tools through a world wide web interface. The initial implementation on adverse drug reaction exploration allows users to identify risk factors and high risk groups for adverse drug reactions. It also allows a search for possible unknown drugs responsible for adverse reactions without prior knowledge.

Further work is planned on *i*Health Explorer, including enhancements to existing features, adding more features, e.g., automatic ADR detection (Jin, Williams, He, Chen & McAullay 2004), outlier detection (Yamanishi, ichi Takeuchi, Williams & Milne 2004), cluster analysis (Jin, Leung & Wong 2002, Jin, Wong & Leung 2003) and its visualisation (Jin, Shum, Leung & Wong 2004), and increasing the capability of the system. The underlying data mining and statistical methods would be developed further, to improve efficiency and accuracy.

Acknowledgements

The authors wish to acknowledge the Commonwealth Department of Health and Ageing and Queensland Health for providing data and funding the project. Authors wish to acknowledge Dr Jisheng Cui for his useful comments.

References

Agresti, A. (1990), *Categorical Data Analysis*, Wiley, New York.

Almenoff, J. S., DuMouchel, W., Kindman, L. A., Yang, X. & Fram, D. (2003), 'Disproportionality analysis using empirical bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting', *Pharmacoepidemiology and Drug Safety* **12**(6), 517–521.

Chen, J., He, H., Williams, G. & Jin, H. (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, Australia, pp. 235–239.

Cios, K. J. & Moore, G. W. (2002), 'Uniqueness of medical data mining', *Artificial Intelligence in Medicine* **26**(1-2), 1–24.

Gu, L., Li, J., He, H., Williams, G., Hawkins, S. & Kelman, C. (2003), Association rule discovery with unbalanced class, *in* 'Proceedings of the 16th Australian Joint Conference on Artificial Intelligence (AI03), Lecture Notes in Artificial Intelligence', Perth, Western Australia.

Harvey, J., Turville, C. & Barty, S. (2004), '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.

Jin, H.-D., Leung, K.-S. & Wong, M.-L. (2002), Scaling-up model-based clustering algorithm by working on clustering features, *in* H. Yin, N. Allinson, R. Freeman, J. Keane & S. Hubbard, eds, 'Proceeding of Third International Conference on Intelligent Data Engineering and Automated Learning, IDEAL-2002', Springer, Manchester, UK, pp. 569–575.

Jin, H.-D., Shum, W., Leung, K.-S. & Wong, M.-L. (2004), 'Expanding self-organizing map for data visualization and cluster analysis', *Information Sciences* **163**, 157–173.

Jin, H.-D., Wong, M.-L. & Leung, K.-S. (2003), Scalable model-based clustering by working on data summaries, *in* 'Proceedings of Third IEEE International Conference on Data Mining (ICDM 2003)', Melbourne, Florida, USA, pp. 91–98.

Jin, H., Williams, G., He, H., Chen, J. & McAullay, D. (2004), Recent data mining research work using linked health data: Extended abstract, *in* 'PAKDD 2004 Workshop Notes: Current Research and Future Directions', Sydney, Australia, pp. 47–51.

Kum, H.-C., Duncan, D., Flair, K. & Wang, W. (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.

Li, J., Shen, H. & Topor, R. (2002), 'Mining the optimal class association rule set', *Knowledge-based Systems* **15**(7), 399–405.

Murff, H. J., Patel, V. L., Hripcsak, G. & Bates, D. W. (2003), 'Detecting adverse events for patient safety research: a review of current methodologies', *Journal of Biomedical Informatics* **36**(1/2), 131–143.

Newman, S. C. (2001), *Biostatistical Methods in Epidemiology*, John Wiley & Sons.

Rao, R. B., Sandilya, S., Niculescu, R. S., Germond, C. & Rao, H. (2003), Clinical and financial outcomes analysis with existing hospital patient records, *in* 'Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 416 – 425.

Reid, M., Euerle, B. & Bollinger, M. (2002), 'Angioedema'. http://www.emedicine.com/med/topic135.htm.

Roddick, J., Fule, P. & Graco, W. (2003), 'Exploratory medical knowledge discovery : Experiences and issues', *SIGKDD Exploration* **5**(1), 94–99.

Seno, M. & Karypis, G. (2002), SLPMiner: An algorithm for finding frequent sequential patterns using length decreasing support constraint, *in* 'Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM)', IEEE, Maebashi City, Japan, pp. 418–425.

Taylor, K., O'Keefe, C., Colton, J., Baxter, R., Sparks, R., Srinivasan, U., Cameron, M. & Lefort, L. (2004), A service oriented architecture for a health research data network, *in* 'Proceedings of the 19th International Conference on Scientific and Statistical Database Management(SSDBM'04)', Greece.

Webb, G. I. (2000), Efficient search for association rules, *in* 'Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 99–107.

Williams, G., He, H., Chen, J., Jin, H., McAullay, D., Sparks, R., Cui, J., Hawkins, S. & Kelman, C. (2004), QLDS: Adverse drug reaction detection towards automation, Technical Report CMIS 04/91, CSIRO Mathematical and Information Sciences, Canberra.

Wilson, A., Thabane, L. & Holbrook, A. (2004), 'Application of data mining techniques in pharmacovigilance', *British Journal of Clinical Pharmacology* **57**(2), 127–134.

Yamanishi, K., ichi Takeuchi, J., Williams, G. & Milne, P. (2004), 'On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms', *International Journal of Data Mining and Knowledge Discovery* **8**(3), 275–300.