electronic Journal of Health Informatics
http://www.ejhi.net
2006; Vol 1 (1): e4

Frequency-based Rare Events Mining in Administrative Health Data

Jie Chen[1], Huidong Jin[1], Hongxing He[1], Christine M. O'Keefe[1], Ross Sparks[1], Graham Williams[1,2], Damien McAullay[1], Chris Kelman[3]

[1]CSIRO Mathematical and Information Sciences, Canberra ACT, Australia
[2]Department of Computer Science, Australian National University, Canberra ACT, Australia
[3]National Centre for Epidemiology and Population Health, Australian National University, Canberra ACT, Australia

Abstract

The low occurrence rate of adverse drug reactions makes it difficult to identify risk factors from a straightforward application of association pattern discovery in large databases. In this paper, we are interested in developing a data mining approach that can use the information about rare events in sequence data in order to measure the multiple occurrences of patterns in the whole period of target and non-target data. To address this, we define an interestingness measure which exploits the difference between the frequency of patterns in target and non-target sequence data. The proposed approach guarantees the easy generation of candidate patterns from the target sequence data by applying existing association mining algorithms. These patterns can then be evaluated by comparing their frequency in the target and non-target data. We also propose a ranking algorithm that takes into account both the rank of the patterns as determined by the interestingness measure and their supports in the target population. This algorithm can prune the patterns greatly and highlight more interesting results. Experimental results of a case study on Angioedema show the usefulness of the proposed approach.

Keywords: Adverse Drug Reaction, Temporal Pattern Mining, Administrative Health Data, Angioedema

1 Introduction

Informally, an Adverse Drug Reaction (ADR) indicates an undesirable response associated with the use of a medicine (The Adverse Drug Reactions Advisory Committee (ADRAC), 2005). It was estimated that in 1994 over 2 million patients hospitalised in USA had serious ADRs, which makes ADRs between the fourth and sixth leading cause of death in USA (Lazarou et al., 1998). Studies also show that 30% to 60% of ADRs are preventable or avoidable by careful prescribing and monitoring (Bates *et al.*, 2003; Lazarou et al., 1998). ADRs occur infrequently but may lead to serious or life threatening conditions requiring hospitalisation. At present, ADRs resulting from new medications and their interactions with other medicines are often detected only if there exist either dramatic or widespread reactions. When a new medicine is introduced, it is likely that unexpected side-effects will go unnoticed until a very substantial number of patients have been adversely affected. Thus, systematic monitoring of health data to more quickly identify possible ADRs is of financial and social importance.

At present, the early detection of unexpected adverse reactions relies on a national voluntary reporting system and collated statistics from overseas agencies. Spontaneous adverse event reporting databases are traditional data sources for most data mining work (Fram et al., 2003; Murff et al., 2003; Harvey et al., 2004; Wilson et al.,

The *electronic Journal of Health Informatics* is an international journal committed to scholarly excellence and dedicated to the advancement of Health Informatics and information technology in healthcare. ISSN: 1446-4381

2004), which focus on generating drug-event associations. The use of a population-based prescribing data, such as the Pharmaceutical Benefits Scheme (PBS) data in Australia, linked to hospital admissions data, would provide an opportunity to detect common and rare adverse reactions at a much earlier stage. Compared with spontaneous adverse event reporting data, these data sets are not lack of denominator data, inexpensive to use, and more complete in information e.g. population-based (Strom & Velo, 1992). From a data mining prospective, the low occurrence rate of ADRs in large administrative health databases often make it difficult to identify the risk factors from a straightforward application of an association pattern discovery algorithm. The problem domain has the following characteristics: (1) Primary interest lies in rare events amongst large datasets; (2) Factors leading to rare adverse drug reactions include temporal medicine exposure; (3) Rare events are associated with a small proportion of patients yet all data for all patients are required to assess the risk.

In this area, we can usually not identify, in advance, appropriate hypotheses. For example, for adverse drug reactions we usually have little prior knowledge about which medicine or medicine combinations might lead to unexpected outcomes (while the expected outcomes have often already been studied). Our aim is to discover temporal patterns associated with rare events that are then further assessed for their possible relationship with adverse outcomes. In our previous work (Chen et al., 2004), the information in the time window before the first target event was considered for the mining of temporal associations. In this paper, we intend to develop a data mining strategy that can use the information around rare events in sequence data. The main contributions of this paper are as follows.

- A new interestingness measure based on frequency of patterns is defined.
- Candidate patterns are generated from case sequences.
- Finally, a collaborative ranking algorithm that can prune the patterns greatly is proposed to highlight more interesting results.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 presents formal definitions. Section 4 outlines the proposed algorithm. Section 5 describes the dataset used in our experiments and reports on some encouraging results. Section 6 and 7 discusses and concludes the paper.

2 Related Work

Temporal patterns mining e.g. temporal association rule (Roddick & Spiliopoulou, 2002; Lee et al., 2003; Jin et al., 2004) and sequential pattern (Srikant & Agrawal, 1996) mining, has drawn much attention in recent years. For example, Mannila et al. (1997) reported a technique for finding frequent episodes from a given event sequence, where an episode is defined as a partial order on a set of events. SPADE (Zaki, 2001) was a method employing efficient lattice search techniques and simple joins that needs to perform only three passes over the database. PrefixSpan (Pei et al., 2000) was a more efficient database projection based algorithm for mining sequential patterns when compared with GSP (Srikant & Agrawal, 1996), as there is no need for candidates generation. In this paper, we adopt the concept of *support* of a pattern as in temporal association rule and sequential pattern mining, that is, the percentage of sequences in a database that contain the pattern.

Regarding mining temporal patterns for rare events, Weiss & Hirsh (1998) described *timeweaver*, a genetic algorithm based machine learning system that predicted rare events by identifying predictive temporal and sequential patterns. Zaki et al. (2000) provided a sequential pattern algorithm that could predict failures in databases of plan executions. Recently, Li & Ma (2004) reported algorithms for discovering pairwise temporal patterns without predefined time windows. Sun et al. (2004) proposed a framework to find interesting patterns from a single long temporal event sequence. In order to improve the prediction accuracy of target events in a long sequence, Sun et al. (2005) further investigated the temporal features of an event-oriented pattern i.e. the minimal size of interval that makes a pattern interesting. In this paper, we are interested in handling more complicated temporal sequences, namely exposure and outcome sequences for disease and non-disease entities, with the awareness of difference between inside and outside hazard windows.

For the purpose of understanding differences between several contrasting groups, Dong and Li (1999) introduced the mining of emerging patterns defined as itemsets whose supports increase significantly from one dataset to another. For the sake of anomaly detection, sequences were reduced and filtered for the profiling of users (Lane & Brodley, 1999). To perform early detection of disease outbreaks, Wong et al. (2003) used an anomaly detection algorithm to detect groups with specific characteristics whose recent pattern of illness was anomalous relative to historical patterns, but it was restricted to two items in a single rule. In contrast to these approaches, the goal of this paper is to explore temporal associations, e.g. possible patterns of exposure leading to a given disease from large temporal sequences data sets.

The problem of large numbers of rules has been extensively studied (Liu et al., 1999; Huang & Webb, 2005). Existing techniques mainly pruned off those qualitative or quantitative association rules that contained little extra information as compared to their ancestors. Recently, Pang-Ning Tan (2004) studied a modified Hedge algorithm to address the pattern ordering problem by combining the rank information gathered from disparate sources. In contrast, we will present an effective collaborative ranking algorithm that takes into account not only the rank of patterns by the proposed interestingness measure but also the support in a target population. The interestingness measure is different from the existing ones reviewed in (Tan et al., 2002).

3 Problem Description

For a set of patients (or entities) $E = \{e_i\}_1^N$ where N is the total number of patients in E, we suppose there is a data set of sequences $D = \{s_i = < (e_{i,1},t_{i,1}),(e_{i,2},t_{i,2}),\ldots,(e_{i,m_i},t_{i,m_i}) >\}_1^N$, and time stamp $t_{i,1} \geq T_{START}$ and $t_{i,m_i} \leq T_{END}$ which means that all sequences are bounded in a constant time period $[T_{START}, T_{END}]$. We consider the occurrences of so-called **target events**, which are hospitalisation events specified by domain experts, e.g. hospital admissions due to Angioedema[1]. For example, the following sequence describes a set of medical services received by a patient:

<(G03CA,1), (J01DA,7),(C08CA,10),(C09AA,10),(Angioedema,31),(C08CA,50),…>

On the first day the patient was dispensed the medicine G03CA (identified using its ATC[2] code, which here is code for estrogen). They then received J01DA on the 7th day, and C08CA and C09AA on the 10th day. Twenty-one days later they were admitted to hospital with Angioedema.

Exposure information is recorded in each sequence of patients. We intend to discover possible patterns of exposures leading to the occurrences of target events. As discussed above, it is not a trivial problem. These target events can occur anywhere within the study period in terms of a population of patients who suffer the disease, which is called the **target population** T. That is, T is the collection of patients in E with at least one target event occurring within $[T_{START}, T_{END}]$. The **non-target population** $\overline{T}$ is the collection of patients without any target event within $[T_{START}, T_{END}]$, and it supplies necessary background information for the assessment of risk of various exposure patterns. To control confounding factors e.g. age and gender, we concentrate on two cohorts ("Older Females: aged 60+" and "Older Males: aged 60+") in the case study of Section 5, which are the majority of Angioedema patients. Thus both T and $\overline{T}$ are from a same age and gender group.

In the following, the problem of quantifying the impact of temporal patterns on multiple occurrences of target events and facilitating the mining of the temporal patterns is formalised. The basic idea is to consider the frequency count of patterns by using a non-overlapped sliding window.

Definition 1: $< (e_{i,p},t_{i,p}),(e_{i,p+1},t_{i,p+1}),\ldots,(e_{i,q},t_{i,q}) >$ *is a* ***w-segment*** *of sequence* s_i, *if* $t_s \leq t_{i,p} \leq t_{i,p+1} \leq \ldots \leq t_{i,q} < t_e \leq t_{i,m_i}$, $t_{i,p-1} < t_s$ *and* $t_{i,q+1} \geq t_e$, *where* $[t_s, t_e]$ *is a sliding time window and its size* $w = t_e - t_s$ *is a constant usually specified by domain experts.*

Definition 2: *For the sequence data set* D, p *is defined as a* ***windowed pattern*** *if*

1. *It is a conjunction* (*or ordered list*) *of items* (*medicines*)[3]
2. It occurs at least once in a w-segment in D, which is called a **matched w-segment** of p.

To make an efficient search of possible associations for target event, we do not consider all possible windowed patterns in D which may incur substantial computational overhead. The solution is to heuristically generate a set of candidate patterns from the sequences of the target population. This reduces the search space greatly in two aspects. Firstly, windowed patterns that only occur in $\overline{T}$ are not characteristic of ADRs, and the total number of windowed patterns for T is usually much less than that for $\overline{T}$ since T is a quite small population compared with $\overline{T}$.

Secondly, only the windowed patterns within a sliding time window ending with a target event are of more interest

[1] *Angioedema is a swelling (large welts or weals), that occurs beneath the skin rather than on the surface. There are a number of case series in the literature demonstrating that ACE inhibitors-related angioedema is responsible for as many as 40% of angioedema episodes (Reid et al., 2002).*

[2] *This uses the Anatomical Therapeutic Chemical (ATC) classification system.*

[3] *It is also called an existence (or sequential) pattern hereinafter.*

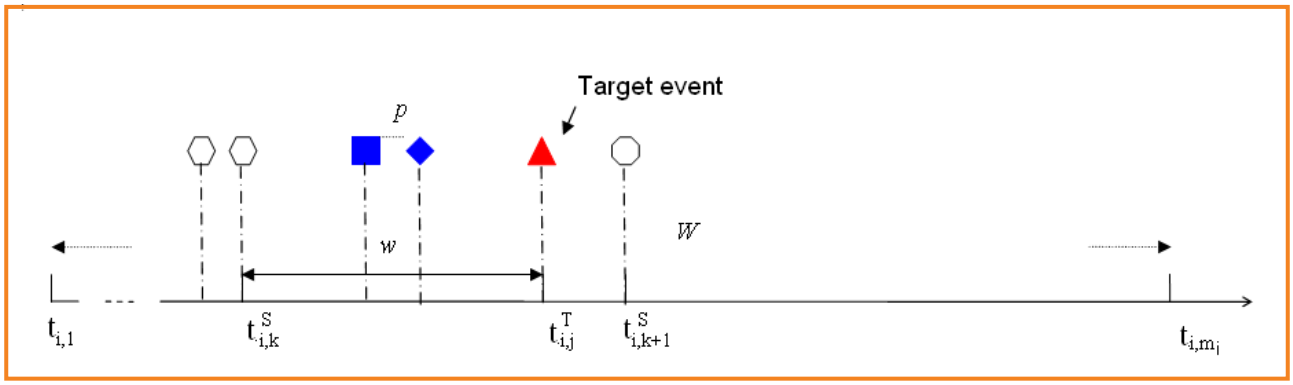

Figure 1: Illustration of jump condition for target population

because our primary objective is to explore possible short-acting medicine exposures associated with a target disease. Specifically, we can deploy existing association pattern mining algorithms to generate the candidate set of windowed patterns after the construction of a sub-database D_{Tw} for T. The idea is to treat each *w*-segment ending with a target event (i.e. t_e is set to be a time stamp of a target event) as a transaction, and if there are multiple target events for a patient, only non-overlapped *w*-segments in s_i will be considered. For each sequence of *T*, we first scan from the start of sequence to get the first target event, and then get the next target event if the window ending with it does not overlap with its previous one, and so on.

If we impose a **jump condition for target population** in the process of counting a windowed pattern *p* in the sequences of *T*, we will observe that **the frequency of** *p* **in** T **is equal to the total number of matched w-segments in** D_{Tw}. This observation is valuable because it enables us to generate a set of candidate windowed patterns by using existing association pattern mining algorithms, such as the OPUS_AR algorithm (Webb, 2000) used in this paper.

The Jump condition for target population is illustrated by Figure 1. In the counting process of p in s_i of T, a sliding window is moved from left to right[4]. We denote the start time stamp of the k^{th} sliding window as $t^S_{i,k}$. The jump condition for target population means that $t^S_{i,k+1}$ will be set to the next consecutive time stamp in s_i except that 1) p is matched in the current sliding window starting from $t^S_{i,k}$ and 2) $t^S_{i,k}$ is the first time stamp such that $t^S_{i,k} \geq t^T_{i,j} - w$, where $t^T_{i,j}$ is one of the time stamps of target events. If this condition is satisfied, $t^S_{i,k+1}$ should jump to the first time stamp in s_i such that $t^S_{i,k+1} \geq t^S_{i,k} + w$, i.e. jump a window ahead to continue the scan. For example, a sliding window with a length of 30 days first starts from *Day 1*, and then it jump to next start time stamp *Day 50* for the counting of pattern "G03CA C09AA" in the above sample sequence. If Angioedema did not occur on *Day 31*, the sliding window would start from *Day 7*.

Thus, we have the following definition of frequency of *p* in the target population.

> **Definition 3:** *For the sequences of* T, $freq_T(p)$ *is the total number of matched w-segments under the jump condition for target population.*

Similarly, we impose a **jump condition for non-target population**. For s_i in $\overline{T}$, each time $t^S_{i,k+1}$ should be set to the next consecutive time stamp in s_i unless p is matched in the current sliding window starting from $t^S_{i,k}$. If this condition

[4] *It can be proved that any non-empty w-segment of* s_i *can be accessed through the slideing window moving event by event (Chen et al., 2004)*

is satisfied, $t^s_{i,k+1}$ should be set to the first time stamp in s_i such that $t^s_{i,k+1} \geq t^s_{i,k} + w$.

Definition 4: *For the sequences of* $\overline{T}$, $freq_{\overline{T}}(p)$ *is the number of matched w-segments under the jump condition for non-target population.*

These two jump conditions ensure that for any s_i in D, the matched w-segments of p being counted are not overlapped. Nonetheless, they are different with respect to target and non-target populations. To compare fairly the occurrences of a windowed pattern in the two populations, we define another frequency metric.

Definition 5: *For the sequences of* T, $freq_{T'}(p)$ *is the total number of matched w-segments under the jump condition for non-target population.*

In summary, $freq_T(p)$ provides a measure about how frequently p appearing in non-overlapped windows ending with target events (called **hazard windows**). $freq_{T'}(p)$ implies how frequently p appears in non-overlapped windows without consideration of target events. It can be derived that $freq_T(p) \leq freq_{T'}(p)$. A higher ratio of $freq_T(p)$ to

Output: Ranked patterns.

Method:

1. output = *NULL*;
2. **for** $r \in R$
3. $T, \overline{T}$ = *PartitionPopulation*(*r*); /*partition entities */
4. $\{p\}, \{freq_T(p)\}$ = GenPattern(*D*,*T*,*w*, s_0);
   /* generate candidate patterns and get $freq_T(p)$*/
5. $\{freq_{T'}(p)\}$ = *CountFreq*(*D*, {*p*}, *T*, *w*);
   /* counting patterns on target population again */
6. $\{freq_{\overline{T}}(p)\}$ = *CountFreq*(*D*, {*p*}, $\overline{T}$, *w*);
   /* counting patterns on non-target population*/
7. ranked-patterns = *ColRank*({*p*, $freq_T(p)$ }); /* collaborative ranking of patterns*/
8. output = output ∪ {ranked-patterns};
9. **return** output;

Figure 2: Pseudo code of the FREM algorithm

$freq_{T'}(p)$ should result in more occurrences of p right in the hazard windows. Based on the above definitions, we define a ***discriminability*** measure to describe how strong a temporal pattern associated with target events, given the information of inside and outside the hazard windows of target population and the sequences of non-target populations.

$$\text{discriminablity}(p) = \frac{w}{W}\left(\frac{freq_{T'}(p)}{|T|} - \frac{freq_{\bar{T}}(p)}{|\bar{T}|}\right)\frac{freq_{\bar{T}}(p)}{freq_{T'}(p)} \qquad (1)$$

where the length of study period is $W = T_{END} - T_{START}$, and the estimated upper bounds of $freq_{T'}(p)$ and $freq_{\bar{T}}(p)$ are $\frac{|T|W}{w}$ and $\frac{|\bar{T}|W}{w}$ respectively. Also note that $freq_T(p) \le freq_{T'}(p)$, we have

$$-1 \le \text{discriminablity}(p) \le 1 \qquad (2)$$

Temporal patterns that are more likely to appear inside hazard windows rather than outside hazard windows or in non-target populations can be highlighted through relatively higher positive values of this interestingness measure. For example, suppose $\frac{w}{W}\left(\frac{freq_{T'}(p)}{|T|} - \frac{freq_{\bar{T}}(p)}{|\bar{T}|}\right) = 0.5$ and $\frac{freq_T(p)}{freq_{T'}(p)} = 1$, which means that the frequency p appearing in hazard windows is the same as the frequency it appears without the limitation of hazard windows, and the normalised difference of frequency of p in target and non-target population is as high as 0.5. Thus it might be of interest in terms of

Input: $\{p, freq_T(p), \text{discriminability}(p)\}$

Output: Ranked patterns

Method:

```
1.  PatternsSorted = SortByDiscriminability( {p, discriminability(p)} );
2.  RankedPatternsSoFar = Null;
3.  for p in PatternsSorted:
4.      if p intersect with x in RankedPatterSoFar:
5.          if freq_T(p) ≥ freq_T(x):
6.              prune(p);
7.          else:
8.              RankedPatternsSoFar.append(p);
9.      else:
10.         RankedPatternsSoFar.append(p);
11. return RankedPatternsSoFar;
```

Figure 3: Pseudo code of ColRank *algorithm*

mining temporal patterns associated with target events. In principle, this interestingness measure has incorporated both the support and strength of a pattern.

4 Frequency-Based Windowed Patterns Mining Algorithm

Figure 2 illustrates the framework of our Frequency-Based Rare Events Mining (FREM) algorithm.

We first partition the whole population according to their demographics and hospitalisation situations with respect to a target disease (Step 3). Then we generate candidate patterns from D_{Tw} of the target population (Step 4). Both existence pattern and sequential pattern algorithms can be integrated in *GenPattern*. The counting of these patterns in $\overline{T}$ requires an efficient algorithm due to the large number of non-target patients (Step 6). Thus, we design an efficient algorithm for the existence patterns in *CountFreq*. Here the basic idea is to update a dynamic data structure for the partially matched items of a pattern, and drop outdated and partially matched items when the sliding window updates. Moreover, we use the set difference between the pattern and a sequence so as to save the search for the pattern that can not appear in the sequence. In our experiment, this algorithm is over five times faster than an intuitive counting process without any optimisation.

Since there are usually a lot of patterns with high *discriminability* values, we need to highlight the most interesting ones for further investigation. Usually the patterns are simply ranked by their interestingness measures. Our idea of ranking interesting patterns is to take into account both the interestingness measure and frequency of patterns in the target population. We propose a pruning method for a collaborative ranking to make the list of interesting patterns shorter (Step 7). A pattern p_1 is pruned if there exists pattern p_2 such that all the following three conditions hold.

(3) $freq_T(p_1) \geq freq_T(p_2)$

(4) $\text{discriminability}(p_1) \leq \text{discriminability}(p_2)$

(5) $\text{intersection}(p_1, p_2) \neq \phi$

It means that pattern p_1 will be pruned if the frequency of p_1 in D_{Tw} is greater than or equal to the frequency of p_2 with the same or higher interestingness measure and p_1 , p_2 also have common items. The underlying reason is that patterns with lower *discriminability* but higher support in D_{Tw} are thought to be less interesting.

Equation (5) can prevent excluding some potential signals from consideration, and improve the chance of detection of most interesting patterns. The ranking algorithm *ColRank* given in Figure 3 illustrates the pruning process of this method. Experimental results in the next section will show that the algorithm can reduce the number of patterns substantially.

5 Mining on Real Health Data: A Case Study

The CSIRO, through its Division of Mathematical and Information Sciences, was commissioned by the now Australian Government Department of Health and Ageing in August 2002 to analyse a linked data set produced from MBS, PBS and Queensland Hospital morbidity data, more commonly referred to as the Queensland Linked Data Set (QLDS). The objective was to provide a demonstration of the utility of data mining on de-identified administrative health data to investigate patterns of utilisation, adverse events and other health outcomes.

The QLDS was made available to CSIRO under a negotiated agreement between the now Australian Government Department of Health and Ageing and Queensland Health. The data set contained de-identified and confidentially linked patient level hospital separation data (1 July 1995 to 30 June 1999), Medicare Benefits Scheme (MBS) data and Pharmaceutical Benefits Scheme (PBS) data (both 1 January 1995 to 31 December 1999). All data were de-identified, and actual dates of service were removed, so that time sequences were indicated only by time from first admission. This process provided strong privacy protection, consistent with the requirements of the relevant Federal and State legislation. CSIRO held the QLDS in a secure computer environment and limited access to authorised staff directly involved in the data analysis.

The QLDS is based on the collection of patients hospitalised in Queensland between 1 July 1995 to 30 June 1999, with linked PBS and MBS data. Because the linkage relied on a valid Medicare number, around 30% of hospital records (those without a valid Medicare Number associated) were discarded. The QLDS therefore contains 3,087,454 hospital records, corresponding to 1,176,294 individuals, which represents about 35% of the Queensland population. The issues of selection bias and data quality of the QLDS are discussed in the report (Williams et al., 2002).

Two datasets are extracted in the following experiments. One contains all 400 patients with hospital admissions due to Angioedema (the target event). The other contains 682,958 patients who have no hospitalizations due to

[5]Actually, it is ranked as the most interesting single drug with the discriminability values of 0.0208 and 0.0241 for the two cohorts, respectively.

[6] Scattered reports suggest the possibility of angioedema associated with the use of estrogens, antihypertensive drugs other than ACE inhibitors, and psychotropic drugs (Agostoni and Cicardi, 2001)

$Rank_c$	$Rank_e$	discriminability(p)	$freq_T(p)$	$freq_{\overline{T}}(p)$	Pattern
1	**1**	**0.0179**	**30418**	**22**	**C09AA G03CA**
2	**4**	**0.0094**	**13714**	**11**	**G03CA C03CA**
3	**5**	**0.0084**	**53844**	**16**	**C09AA ——**
4	**6**	**0.0078**	**44251**	**14**	**C09AA ——**
5	**7**	**0.0078**	**40426**	**13**	**C09AA ——**
6	10	0.0072	67019	15	—— ——
7	**12**	**0.0069**	**25141**	**10**	**G03CA ——**
8	**15**	**0.0068**	**55186**	**14**	**C03CA ——**
9	17	0.0066	26598	11	—— ——
10	**18**	**0.0065**	**31011**	**11**	**C03CA C08CA**
11	**19**	**0.0065**	**24707**	**11**	**C09AA ——**
12	22	0.0062	39230	12	—— ——
13	28	0.0059	21250	10	—— ——
14	**34**	**0.0053**	**37728**	**10**	**C08CA ——**
15	**40**	**0.0049**	**28612**	**10**	**C09AA ——**
16	41	0.0047	41997	10	—— ——
17	49	0.0039	55912	13	—— ——
18	**55**	**0.0037**	**44154**	**10**	**C03CA ——**
19	57	0.0037	41958	10	—— ——
20	60	0.0035	73093	10	—— ——

Table 1: Ranked patterns for females aged 60+ ($|T|/|\overline{T}|$ for this cohort is 105/128558)

$Rank_c$	$Rank_e$	discriminability(p)	$freq_T(p)$	$freq_{\overline{T}}(p)$	Pattern
1	**1**	**0.0125**	**54782**	**12**	**C09AA C03CA**
2	**3**	**0.0118**	**46736**	**11**	**C09AA C08CA**
3	7	0.0108	66975	13	—— ——
4	17	0.0092	18097	7	—— ——
5	21	0.0082	25253	6	—— ——
6	26	0.0079	17578	6	—— ——
7	**32**	**0.0077**	**16351**	**5**	**C08CA ——**
8	37	0.0069	8873	5	—— ——
9	42	0.0064	25958	6	—— ——
10	53	0.0057	35041	5	—— ——
11	55	0.0055	18106	5	—— ——
12	71	0.0042	33191	5	—— ——
13	73	0.0033	26361	5	—— ——

Table 2: Ranked patterns for males aged 60+ ($|T|/|\overline{T}|$ for this cohort is 58/102796)

	Window size	Non-target population	C09AA G03CA	C09AA C03CA
discriminability(p)	180 days	682,958	0.0179	0.0125
$Rank_c$			1	1
discriminability(p)	180 days	85,229	0.0187	0.0136
$Rank_c$			1	1
discriminability(p)	90 days	682,958	0.0076	0.0076
$Rank_c$			1	1
discriminability(p)	90 days	85,229	0.0077	0.0075
$Rank_c$			1	1

Table 3: Results for the top patterns in Tables 1 and 2 with four different settings

Angioedema. We stratify the population into age and gender groups. The study period is four years from 1995 to 1999, and we choose a hazard window of 180 days as suggested by contributing medical experts. The minimum support for the generation of candidate patterns for both the cohorts is 8%.

We then apply the FREM algorithm on these data sets. The generated interesting patterns for the females and males aged 60+ cohorts are shown in Tables 1 and 2 respectively. Here we only consider patterns involving two PBS events to make the results easier for interpretation. Note that the $Rank_d$ in the tables denotes the order of a pattern sorted by *discriminability*, and $Rank_c$ in the tables denotes the order of a output pattern by using the *ColRank* algorithm. The numbers of resulting patterns have been reduced from 79 and 77 to 20 and 13 for the two cohorts, respectively. Among these ranked patterns, *ACE inhibitors* (ATC code: C09AA) has appeared as the most interesting medicine in both tables[5] , which is consistent with medical knowledge[1]. The first pattern in Table 1 is "C09AA G03CA", which means the combination usage of *ACE inhibitors*[1] and *Estrogen*[6] (ATC code: G03CA) within 180 days is highly associated with the occurrence of Angioedema. This is consistent with the findings in (Chen et al., 2004). For males aged 60+, the most interesting pattern "C09AA C03CA" suggests that the combination usage of *ACE inhibitors* and *Sulfonamides* within 180 days is highly associated with the occurrence of Angioedema. Interestingly, *Furosemide* (C03CA01) as one subcategory of *Sulfonamides* has been reported to cause acute reaction of Angioedema (Furosemide, n.d.; Hansbrough et al., 1987). In addition, *Amlodipine besylate* (C08CA01) as one sub-category of *Dihydropyridine derivatives* (C08CA in Rows 10 and 14 of Table 1, and Rows 2 and 7 of Table 2) has been reported to cause allergic reactions including Angioedema, pruritis, rash, and erythema multiforme (NORVASC, n.d.).

To further examine the algorithm, we ran it with different settings, i.e., changing the window size to 90 days, and using a random sampled non-target population. Specifically, the whole non-target population is recursively and randomly bi-partitioned to derive a sampled non-target population with 85,229 patients, about 1/8 of the whole 682,958. Table 3 shows the results of four different settings, for the top patterns in Tables 1 and 2 respectively. For all the settings in Table 3, the minimum support for the generation of candidate patterns for both the cohorts is 8%. For a window size setting of 90 days (the last two rows in Table 3), the number of resulting patterns of the two cohorts have been reduced from 24 and 39 to 8 and 9 by using the *ColRank* algorithm respectively. Clearly, both patterns "C09AA G03CA" and "C09AA C03CA" for the two cohorts respectively are steadily ranked as No1 by the FREM algorithm. We also check all other patterns in Tables 1 and 2. We find that the *discriminability* values and the ranked results based on the sampled population are very close to those based on the whole non-target population under the settings in Table 3, although the *discriminability* values decrease as the window size decreases.

Since the generation of candidate patterns in a small target population is quite fast, we can get the results in a short time if the algorithm is applied to a sampled non-target population. Actually the FREM algorithm was implemented in Python, and all experiments were run on an Intel Pentium 4 (3.2GHz)/Linux, and the runtime for the last two rows in Table 3 was 2 minutes 13 seconds. It is clear that the FREM algorithm is linearly scalable in the number of patients if the integrated existence and sequential pattern mining algorithms are linearly scalable in the number of patients.

As pointed out in (Chen et al., 2004), our intent is to generate hypotheses identifying potentially interesting patterns, while we realise that further validation and examination are necessary.

6 Discussion

The FREM algorithm can easily integrate existing existence and sequential pattern mining algorithms, for which there is no easy solution to the problem in this paper since sequence data of both target and non-target population need to be mined and target events can occur anywhere in the sequence of a patient. It is also not trivial to modify previous methods e.g. (Sun et al., 2004; Sun et al., 2005) on predicting target events in a long temporal sequence. In the above experiments, the temporal association patterns were ranked by the interestingness measure which exploits the difference of frequencies of a pattern inside and outside hazard windows in both target and non-target sequences. It also enables that multiple occurrences of patterns in the whole period of target and non-target data can be measured.

For both the cohorts, ACE inhibitors (C09AA) was successfully highlighted as the most suspected medicine associated with the target event Angioedema. *Amlodipine besylate* (C08CA01) as one sub-category of *Dihydropyridine derivatives* (C08CA), and *Furosemide* (C03CA01) as one sub-category of *Sulfonamides* (C03CA) were also reported to cause allergic reactions including Angioedema. Regarding the highlighted combination "C09AA G03CA", it was pointed out in (Agostoni and Cicardi, 2001) that *Estrogen* and *ACE inhibitors* should be avoided in a patient with congenital or acquired C1-INH deficiency. Moreover, a report on the QLDS also indicates that individuals taking the medicine of a high level category of G03CA are 1.7 times more likely to have Angioedema than who do not (Gu et al., 2003). The report applied Logistic regression to the profile data of ACE inhibitors users, with the consideration of age, gender, indigenous status, location, sickness, the total number of ACE inhibitors scripts, 8 hospital diagnosis flags and 13 ATC level-1 drug flags. Consequently, it might be worth investigating medical implication of this combination of exposures.

7 Conclusion

We have proposed the FREM algorithm to address the problem of mining temporal associations with multiple occurrences of target events in terms of ADRs. The experimental results of using an efficient counting algorithm on real administrative health data have demonstrated the effectiveness and efficiency of the algorithm. This paper can be extended in a variety of ways. For example, we can consider medicine prescription events rather than hospitalisation events as target events. This framework could be applied to other applications where mining temporal sequences of contrast entities is of interest.

Acknowledgements

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

References

Agostoni, A. and Cicardi, M. (2001). Drug-induced angioedema without urticaria. Drug Safety, 24(8):599–606.

Bates, D. *et al.* (2003). Detecting adverse events using information technology. *Journal of American Medical Informatics Association*, 10(2):115–128.

Chen, J., He, H., Williams, G., & Jin, H. (2004). Temporal sequence associations for rare events. In *Proceedings of PAKDD'04*, pages 235–239.

Dong, G. & Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52, San Diego.

Fram, D. M., Almenoff, J. S., & DuMouchel, W. (2003). Empirical bayesian data mining for discovering patterns in post-marketing drug safety. In *Proceedings of SIGKDD'03*, pages 359–368.

Furosemide. (n.d.) Chemical Safety Information from Intergovernmental Organizations, from *http://www.inchem.org/documents/pims/pharm/pim240.htm*

Gu, L., He, H., Williams, G., Hawkins, S., Kelman, C., & Li, J. (2003). QLDS: Methods for identifying rare adverse drug reactions. Client Report 03/113, CMIS.

Hansbrough, J. R., Wedner, H. J., & Chaplin, D. D. (1987). Anaphylaxis to intravenous furosemide. *J Allergy Clin Immunol*, 80(4):538–41.

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.

Huang, S. & Webb, G. (2005). Discarding insignificant rules during impact rule discovery in large databases. In *Proceedings of the SIAM 2005 Data Mining Conference (SDM'05)*.

Jin, H., Chen, J., Williams, G., & He, H. (2004). A survey on temporal/streaming data mining. Technical Report CMIS 04/105, CSIRO.

Lane, T. & Brodley, C. E. (1999). Temporal sequence learning and data reduction for anomaly detectionn. *ACM Transactions on Information and System Security*, 2(3):295–311.

Lazarou, J., Pomeranz, B., & Corey, P. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *The Journal of the American Medical Association*, 279(15):1200–1205.

Lee, C.-H., Chen, M.-S., & Lin, C.-R. (2003). Progressive partition miner: An efficient algorithm for mining general temporal association rules. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):1004–1017.

Li, T. & Ma, S. (2004). Mining temporal patterns without predefined time windows. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 451–454.

Liu, B., Hsu, W., & Ma, Y. (1999). Pruning and summarizing the discovered associations. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 125 – 134.

Mannila, H., Toivonen, H., & Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289.

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.

NORVASC. (n.d.) Inhouse Pharmacy, from *http://www.inhousepharmacy.com/heart-health/norvasc-information.html*.

Pang-Ning Tan, R. J. (2004). Ordering patterns by combining opinions from multiple sources. In *Proceedings of the Tenth ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining (KDD-2004)*.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M.-C. (2000). PrefixSpan mining sequential patterns efficiently by prefix projected pattern growth. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 215–226.

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

Roddick, J. F. & Spiliopoulou, M. (2002). A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):750–767.

Srikant, R. & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. In *Proceeding 5th International Conference on Extending Database Technology, EDBT*, pages 3–17.

Strom, B. L. & Velo, G., editors (1992). *Drug Epidemiology and Post-Marketing Surveillance*. Plenum Press, New York.

Sun, X., Orlowska, M. E., & Li, X. (2004). Finding negative event-oriented patterns in long temporal sequences. In *Proceedings of PAKDD'04*, pages 212–221.

Sun, X., Orlowska, M. E., & Li, X. (2005). Finding temporal features of event-oriented patterns. In *Proceedings of PAKDD'05*, number 778-784.

Tan, P.-N., Kumar, V., & Srivastava, J. (2002). 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.

The Adverse Drug Reactions Advisory Committee (ADRAC) (2005). Australian ADR bulletin. DoHA, Australia, from http://www.tga.gov.au/adr/aadrb.htm.

Webb, G. I. (2000). Efficient search for association rules. In *Proceedings of SIGKDD'00*, pages 99–107.

Weiss, G. M. & Hirsh, H. (1998). Learning to predict rare events in event sequences. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 359–363, New York.

Williams, G., Vickers, D., Rainsford, C., Gu, L., He, H., Baxter, R., & Hawkins, S. (2002). Bias in the Queensland Linked Data Set. Technical Report 02/117, 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.

Wong, W.-K., Moore, A., Cooper, G., & Wagner, M. (2003). WSARE: What's strange about recent events? *Journal of Urban Health*, 80(2):i66–i75.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60.

Zaki, M. J., Lesh, N., & Ogihara, M. (2000). PLANMINE: Predicting plan failures using sequence mining. *Artificial Intelligence Review*, 14(6):421–446.

Correspondence

Jie Chen
CSIRO Mathematical and Information Sciences
GPO Box 664
Canberra ACT 2601
Australia

Jie.Chen@csiro.au