Frequency-Based Temporal Pattern Mining in 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, GPO Box 664, Canberra ACT 2601, Australia*

[2] *Australian Taxation Office, 51 Allara Street, Canberra ACT 2601, Australia* (Current affiliation)

[3] *National Centre for Epidemiology and Population Health, the Australian National University, Canberra ACT 0200, Australia*

Abstract

The low occurrence rate of adverse drug reactions makes it difficult to identify the risk factors from straightforward application of frequent pattern discovery in large databases. In this paper, we are interested in developing a data mining strategy that can fully utilize the information around rare events in sequence data in order to measure the multiple occurrences of patterns in the whole period of target and non-target data. We define an interestingness measure which exploits the difference between frequency of patterns in target and non-target sequence data. The proposed strategy guarantees the easy generation of candidate patterns from the target sequence data by applying existing association mining algorithms. Then these patterns can 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 patterns as determined by the interestingness measure and the support in the target population, which 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.

1 Introduction

Adverse drug reactions (ADRs) occur infrequently but may lead to serious or life threatening conditions requiring hospitalisation. At present, adverse drug reactions resulting from new medications and their interactions with other medicines, are often detected only if there exists either dramatic or widespread reactions. 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. Thus, systematic monitoring of health data to more quickly identify possible ADRs is of financial and social importance. In general, the early detection of unexpected adverse reactions relies on a local voluntary reporting system and collated statistics from overseas agencies. The use of a population-based prescribing data set, 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. From a data mining prospective, the low occurrence rate of ADRs in large databases often makes it difficult to identify the risk factors from a straightforward application of frequent 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 drug exposure; (3) Rare events are associated with a small proportion of patients yet all data for all patients are required to assess the risk.

Often, we can not identify, in advance, appropriate hypotheses. For example, for adverse drug reactions we usually have little prior knowledge on which drug or drug 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 [1], only the information in the time window before the first target event was considered for the mining of temporal associations. In this paper, we are interested in developing a data mining strategy that can fully utilize 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 concludes the paper.

2 Related Work

Temporal patterns mining has drawn much attention in recent years [12, 11, 6]. Regarding mining patterns for rare events, [16] describes *timeweaver*, a genetic algorithm based machine learning system that predicts rare events by identifying predictive temporal and sequential patterns. [19] provides an sequential pattern algorithm that can predict failures in databases of plan executions. The framework proposed by [13] finds interesting patterns from a single long temporal event sequence. In this paper, we are interested in handling more complicated temporal sequences, namely the exposure and outcome sequences for disease and non-disease entities, with the awareness of difference between inside and outside hazard windows.

Following the goal of understanding differences between several contrasting groups, [2] introduces the emerging patterns mining. [18] uses anomaly detection algorithm to detect groups with specific characteristics whose recent pattern of illness is anomalous relative to historical patterns, but it limits itself to two items in a single rule. In contrast, the goal of this paper is to explore temporal associations from large temporal sequences datasets.

The problem of large number of rules has been studied by many researchers [7, 5]. They mainly prune off those qualitative or quantitative association rules that contain little extra information as compared to their ancestors. Recently, [9] studies a modified Hedge algorithm to address the pattern ordering problem by combining the rank information gathered from disparate sources. We present an effective collaborative ranking algorithm that takes into account not only the rank of patterns by the interestingness measure but also the support in the target population. The interestingness measure is also different from the general ones reviewed by [14].

3 Problem Description

Let $E = \{\epsilon_i\}$ be a set of entities (patients). Suppose there is a database of sequences $D = \{s_i =< (e_{i1}, t_{i1}), (e_{i2}, t_{i2}), ..., (e_{ij}, t_{ij}), ..., (e_{im_i}, t_{im_i}) > \}$ and for any s_i, $t_{i_1} \geq T_{START}$ and $t_{im_i} \leq T_{END}$ which means that all sequences are bounded in a constant time period $[T_{START}, T_{END}]$. We care about the occurrences of events called *target events*, which are user specified hospitalisation events in our case. For these target events, we try to explore the associations between these and other events, or to identify the high risk exposures associated with the outcome. The population E is partitioned into two subsets T and $\overline{T}$, where T are the patients or entities that have at least one target event occurring within $[T_{START}, T_{END}]$ and $\overline{T}$ is for the others.

Definition 1 $\langle (e_{ip}, t_{ip}), (e_{i,p+1}, t_{i,p+1}), ..., (e_{iq}, t_{iq}) \rangle$ *is a* ***windowed segment*** *of sequence s_i with time window $[t_s, t_e]$ if $t_s \leq t_{ip} \leq t_{i,p+1} \leq \ldots \leq t_{iq} < t_e \leq t_{im_i}$, $t_{i,p-1} < t_s$ and $t_{i,q+1} \geq t_e$, $w = t_e - t_s$, where $w = t_e - t_s$ is constant, and usually specified by a domain expert.*

Definition 2 *For sequence data D, p is defined as a* ***windowed pattern*** *if 1) It is a conjunction (or ordered list)of items(drugs) 2) There exists at least one windowed segment so that there is at least one occurrence of pattern p within in the windowed segment The windowed segment is called a* ***matched windowed segment*** *of p.*

To make an efficient search of possible associations for target events, we do not consider all possible windowed patterns in D. We generate a candidate set of windowed patterns directly from the sequences in T. The idea is to construct a sub-database D_{Tw} for T, i.e. treat each windowed segment exactly prior to each target event (t_e is the time stamp of a target event) as a transaction, and if there are multiple target events for a patient (entity), non-overlapped windowed segments in s_i are considered. Namely, for each sequence of T, we first scan from the start of sequence to get the first target event, then get the next target event, and so on, if the window ending with it is not overlapped with its previous one.

Target events may appear in one sequence of a patient (entity) multiple times. For simplicity, we impose a **jump condition for target population**, which can be illustrated by Figure 1. For any s_i, in the search process of a pattern p, we use sliding windows event by event according to the order of time stamps. It can be proved that any windowed segment of s_i can be accessed in such a way [1]. We denote the start time stamp of the kth sliding window as t^S_{ik}. For s_i of any entity in T, each time $t^S_{i,k+1}$ will be set to the next consecutive time stamp of s_i except that 1) p is matched in the current sliding window starting from t^S_{ik} and 2) t^S_{ik} is the first time stamp in s_i that $t^S_{ik} \geq t^T_{ij} - w$, where t^T_{ij} is one of the time stamps of target events. If this exception happens, $t^S_{i,k+1}$ will be set to the first time stamp in s_i that $t^S_{i,k+1} \geq t^S_{ik} + w$, i.e. jump a window ahead to continue the scan. We can make the following definition of **frequency** and observation.

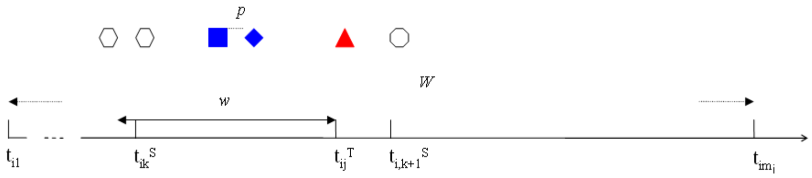

Figure 1: Illustration of jump condition for target population

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

Observation 1. *$freq_T(p)$ is equal to the total number of sequences in D_{Tw} for any windowed pattern p generated from D_{Tw}.*

This observation is useful because it enables us to generate a set of candidate windowed patterns by using existing frequent patterns mining algorithms, say OPUS [15] used in this paper. Similarly, we impose a **jump condition for non-target population**. For s_i of any entity in $\overline{T}$, each time $t^S_{i,k+1}$ will be set to the next consecutive time stamp of s_i except that 1) p is matched in the current sliding window starting from t^S_{ik}. If this exception happens(here no second condition is specified as there is no any target events in the sequence), $t^S_{i,k+1}$ will be set to the first time stamp in s_i that $t^S_{i,k+1} \geq t^S_{ik} + w$, i.e. jump a window ahead to continue the scan.

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

These two jump conditions ensure that for any s_i in D, the counted matched windowed segments of p are not overlapped. Nonetheless, they are different for 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 windowed segments under the jump condition for non-target population.*

In summary, $freq_T(p)$ provides a measure about how frequent p appears in non-overlapped windows exactly prior to 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 $freq_T(p) \leq freq_{T'}(p)$. The higher the ratio of $freq_T(p)$ to $freq_{T'}(p)$, there should be more occurrences of p right in the hazard windows. Based on the above definitions of frequencies, we define a **discriminability** measure to describe a temporal pattern associated with target events, given the information of the whole sequences inside and outside the hazard winows of target population and non-target populations as well.

$$discriminability(p) = \frac{w}{W}\left(\frac{freq_{T'}(p)}{|T|} - \frac{freq_{\overline{T}}(p)}{|\overline{T}|}\right)\frac{freq_T(p)}{freq_{T'}(p)} \tag{1}$$

Input: Two datasets of all entities including their event sequences D and demographics, a window size w and study period size W, a minimum support for target dataset and demographical rules R for generating sub-populations.
Output: Ranked patterns.
Method:

```
output=NULL;
for $r \in$ R
    $T$, $\overline{T}$ =PartitionPopulation($r$) /*partition entities */
    {$p$}, {$freq_T(p)$} = GenPattern($D$,$T$,$w$,$s_0$); /* generate candidate patterns and get $freq_T(p)$*/
    {$freq_{T'}(p)$} = CountFreq($D$, {$p$},$T$, $w$); /* counting patterns on target population again */
    {$freq_{\overline{T}}(p)$} = CountFreq($D$, {$p$},$\overline{T}$, $w$); /* counting patterns on non-target population */
    ranked-patterns = ColRank({$p$,$freq_T(p)$}); /* collaborative ranking of patterns*/
    output = output $\cup$ {ranked-patterns};
return output;
```

Figure 2: Pseudo code of the FREM algorithm

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

$$-1 \leq discriminability(p) \leq 1 \qquad (2)$$

Temporal patterns that are more likely to appear inside hazard windows rather than outside hazard windows or in non-target populations will be highlighted through relative higher positive values of this interestingness measure. For example, suppose $\frac{w}{W}(\frac{freq_{T'}(p)}{|T|} - \frac{freq_{\overline{T}}(p)}{|\overline{T}|}) = 0.5$ and $\frac{freq_T(p)}{freq_{T'}(p)} = 1$, i.e. we have $discriminability(p) = 0.5$ which means that the the frequency p appears 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 our interest for 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 the target disease. Then we generate candidate patterns from D_{Tw} of the target populations. Both existence patterns, which ignore the order of events, and sequential patterns algorithms can be integrated in *GenPattern*. The counting of these patterns in $\overline{T}$ needs an efficient algorithm due to the large number of non-target patients. Thus, we design an efficient algorithm for the existence patterns in *CountFreq*, which is illustrated in Figure 3. Here the general idea is to update the dynamic data structure for the partially matched items of a pattern, dropping outdated, partially matched items when the

```
Input: {p}, sequences of T or $\overline{T}$
Output: $freq_{T'}(p)$ or $freq_{\overline{T}}(p)$
Method:

Patterns = PreselectPatterns({p}, sequence $s_i$);
Patterns.PartiallyMatched = Null;
IndexCurrentEvent = ValidSlidingWindow.start = 0;
Patterns.LastMatched = $-\infty$
while ValidSlidingWindow :
  Patterns.PartiallyMatched.drop();
  if IndexCurrentEvent - ValidSlidingWindow.start < $w$ :
    IndexCurrentEvent += 1;
  else: break
  if Patterns.PartiallyMatch():
    Patterns.PartiallyMatched.update();
    if ValidSlidingWindow.start - Patterns.LastMatched > $w$:
      if Patterns.MatchedCheck():
        Patterns.count += 1;
        Patterns.LastMatched = ValidSlidingWindow.start;
  ValidSlidingWindow.start += 1;
return Patterns.count;
```

Figure 3: Pseudo code of *CountFreq* algorithm

sliding window updates. Moreover, *PreselectPatterns* uses the set difference between pattern and 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 a intuitive counting process without these optimisations.

Since there are usually many patterns with high discriminability values. We need to highlight the most interesting ones for further investigation.Usually the patterns are 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 condition for the collaborative ranking to shortlist interesting patterns.

$$freq_T(p_1) \geq freq_T(p_2) \tag{3}$$

but

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

and

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

It means that a pattern p_1 will be pruned if the frequency of p_1 in D_{Tw} is greater than or equal to any pattern p_2 with the same or higher interestingness measure and p_1 and p_2 also

```
Input: {$p$,$freq_T(p)$, $discriminability(p)$}
Output: Ranked patterns
Method:

PatternsSorted = SortByDiscriminability({$p$,$discriminability(p)$});
RankedPatternsSoFar = Null;
for $p$ in PatternsSorted:
    if $p$ intersect with $x$ in RankedPatterSoFar:
        if $freq_T(p) \geq freq_T(x)$:
            prune($p$);
        else:
            RankedPatternsSoFar.append($p$);
    else:
        RankedPatternsSoFar.append($p$);
return RankedPatternsSoFar;
```

Figure 4: Pseudo code of *ColRank* algorithm

have common items. Equation 5 can prevent excluding some potential signals from consideration, and achieve the goal of improving the chance of detection of most significant patterns. The ranking algorithm *ColRank* given in Figure 4 will do the pruning process according to this condition. Experiment result 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 Queensland Linked Data Set [17] links hospital admissions data from Queensland Health with the pharmaceutical prescription data from Commonwealth Department of Health and Ageing, providing a de-identified dataset for analysis. The record for each patient includes demographic variables and a sequence of PBS and hospitalisation events. Two datasets are extracted. One contains all 400 patients with hospital admissions due to angioedema [1](the target event). The other contains 682,958 patients who have no angioedema hospitalisations. 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.

We used our FREM algorithm on this data set. The ranked interesting patterns for the female and male aged 60+ cohorts are shown in Table 1 and 2 respectively. The minimum support for the generation of candidate patterns for both cohorts is 8%. Here we only consider patterns involving two drugs at a time, to make results easier to interpret. Note that

[1] Angioedema is a swelling (large welts or weals), that occurs beneath the skin rather than on the surface [10]. 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 [10].

No.	$discriminability(p)$	$freq_{\overline{T}}(p)$	$freq_T(p)$	Pattern
1	0.0179	30418	22	C09AA G03CA
4	0.0094	13714	11	G03CA C03CA
5	0.0084	53844	16	C09AA N05CD
6	0.0078	44251	14	C09AA C07AB
7	0.0078	40426	13	C09AA R03AC
10	0.0072	67019	15	N02BE C01DA
12	0.0069	25141	10	G03CA N05CD
15	0.0068	55186	14	C03CA C01DA
17	0.0066	26598	11	J01DA H02AB
18	0.0065	31011	11	C03CA C08CA
19	0.0065	24707	11	C09AA M01AB
22	0.0062	39230	12	N05CD C01DA
28	0.0059	21250	10	C01DA J01FA
34	0.0053	37728	10	C08CA C07AB
40	0.0049	28612	10	C09AA J01CA
41	0.0047	41997	10	A02BA N06AA
49	0.0039	55912	13	N02BE R03AC
55	0.0037	44154	10	C03CA A12BA
57	0.0037	41958	10	J07BB R03AC
60	0.0035	73093	10	N02BE J01DA

Table 1: Ranked patterns for females aged 60+ ($|T|/|\overline{T}|$ for this cohort is 101/12858

No.	$discriminability(p)$	$freq_{\overline{T}}(p)$	$freq_T(p)$	Pattern
1	0.0125	54782	12	C09AA C03CA
3	0.0118	46736	11	C09AA C08CA
7	0.0108	66975	13	A02BA N02BE
17	0.0092	18097	7	N05CD R03BA
21	0.0082	25253	6	R03AC N02AA
26	0.0079	17578	6	N05CD D07AC
32	0.0077	16351	5	C08CA D07AC
37	0.0069	8873	5	J01CA A03FA
42	0.0064	25958	6	H02AB R03BA
53	0.0057	35041	5	J07BB C01DA
55	0.0055	18106	5	N05CD C07AB
71	0.0042	33191	5	N02BE R03BA
73	0.0033	26361	5	A02BA H02AB

Table 2: Ranked patterns for males aged 60+ ($|T|/|\overline{T}|$ for this cohort is 53/102796)

the "No." in tables denotes the order of a pattern sorted by their discriminabilities. The number 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 [2] code: C09AA) has appeared as the most interesting drug in both tables, which is consistent with the knowledge of medical practitioners. The first pattern in Table 1 is "C09AA G03CA", which means the combination usage of *ACE inhibitors* and *estrogen* within 180 days is highly associated with the occurrence of angioedema. This result is consistent with our previous discovery in [1]. For males aged 60+, the most interesting pattern "C09AA C03CA" suggests that the combination usage of *ACE inhibitors* and *Sulfonamides, Plain* within 180 days is highly associated with the occurrence of angioedema reactions. Interestingly, *Furosemide* (C03CA01) as one sub-categorty of *Sulfonamides, Plain* has been reported to cause acute reaction of angioedma [3, 4].*Amlodipine besylate* (C08CA01) as one sub-categorty of *Dihydropyridine derivatives* (C08CA) has been reported to cause allergic reactions including pruritis, rash, angioedema and erythema multiforme [8].

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

6 Discussion and Conclusions

We have defined an interestingness measure which exploits the difference of frequency of patterns in target and non-target sequence data, so that multiple occurrences of patterns in the whole period of target and non-target data can be measured. The proposed strategy guarantees the generation of candidate temporal patterns from the target sequence data by integrating conventional frequent pattern mining algorithms. Then, these patterns can be evaluated in conjunction with the frequency of them in non-target data. We have also proposed a collaborative ranking algorithm that takes into account both the rank of patterns by the interestingness measure and the support in the target population, which can prune the patterns greatly and highlight more interesting results. The experimental results by using an efficient counting algorithm on real health data show the usefulness of the proposed approach. This paper can be extended in a variety of aspects. For example, we can consider drug prescription events rather than hospitalization events as target events for our ongoing work. We suggest this framework could be applied to other applications where mining temporal sequences of contrast entities is of interest.

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] 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.

[2] G. Dong and J. Li. 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, August 1999.

[3] Furosemide. *http://www.inchem.org/documents/pims/pharm/pim240.htm.* Chemical Safety Information from Intergovernmental Organizations.

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

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

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

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

[8] NORVASC. *http://www.inhousepharmacy.com/heart-health/norvasc-information.html.* Inhouse Pharmacy.

[9] R. J. Pang-Ning Tan. 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)*, 2004.

[10] M. Reid, B. Euerle, and M. Bollinger. Angioedema, 2002. http://www.emedicine.com/med/topic135.htm.

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

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

[13] X. Sun, M. E. Orlowska, and X. Li. Finding negative event-oriented patterns in long temporal sequences. In *Proceedings of PAKDD'04*, pages 212–221, May 2004.

[14] 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.

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

[16] G. M. Weiss and H. Hirsh. 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, August 1998.

[17] 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.

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

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