Mining Unexpected Temporal Associations: Applications in Detecting Adverse Drug Reactions

Huidong (Warren) Jin, *Member, IEEE*, Jie Chen, *Member, IEEE*, Hongxing He, Graham J. Williams, *Member, IEEE*, Chris Kelman, *Member, IEEE*, and Christine M. O'Keefe

***Abstract*—In various real-world applications, it is very useful mining *unanticipated episodes* where certain event patterns unexpectedly lead to outcomes, e.g., taking two medicines together sometimes causing an adverse reaction. These *unanticipated episodes* are usually unexpected and infrequent, which makes existing data mining techniques, mainly designed to find frequent patterns, ineffective. In this paper, we propose unexpected temporal association rules (UTARs) to describe them. To handle the unexpectedness, we introduce a new interestingness measure, *residual-leverage*, and develop a novel case-based exclusion technique for its calculation. Combining it with an event-oriented data preparation technique to handle the infrequency, we develop a new algorithm MUTARC to find pairwise UTARs. The MUTARC is applied to generate adverse drug reaction (ADR) signals from real-world healthcare administrative databases. It reliably shortlists not only six known ADRs, but also another ADR, flucloxacillin possibly causing hepatitis, which our algorithm designers and experiment runners have not known before the experiments. The MUTARC performs much more effectively than existing techniques. This paper clearly illustrates the great potential along the new direction of ADR signal generation from healthcare administrative databases.**

***Index Terms*—Adverse drug reaction (ADR), data mining, healthcare administrative databases, pharmacovigilance, unanticipated episode, unexpected temporal association.**

Manuscript received January 9, 2007; revised April 23, 2007.

H. Jin was with the Commonwealth Scientific and Industrial Research Organization (CSIRO), Mathematical and Information Sciences, Canberra, A.C.T. 2601, Australia. He is now with the National Information and Communications Technology Australia (NICTA), Canberra, A.C.T. 2601, Australia (e-mail: huidong.jin@nicta.com.au).

J. Chen, retired, was with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Mathematical and Information Sciences, Canberra, A.C.T. 2601, Australia. He is now with SigNav Pty Ltd., Australia (e-mail: jiechen@ieee.org).

H. He was the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Mathematical and Information Sciences, Canberra A.C.T. 2601, Australia (e-mail: hongxing.he@hotmail.com).

G. J. Williams was the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Mathematical and Information Sciences, Canberra A.C.T. 2601, Australia. He is now with the Australian Taxation Office, Canberra, A.C.T. 2601, Autralia, and also with the University of Canberra, Canberra, A.C.T. 2601, Australia, and the Australian National University, Canberra, A.C.T. 2601, Australia (e-mail: Graham.Williams@togaware.com).

C. Kelman is with the National Centre for Epidemiology and Population Health, Australian National University, Canberra, A.C.T. 0200, Australia (e-mail: chris.kelman@anu.edu.au).

C. M. O'Keefe was with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Mathematical and Information Sciences, Canberra A.C.T. 2601, Australia. She is now with the CSIRO Preventive Health National Research Flagship, Canberra, A.C.T. 2601, Australia, and also with the University of Adelaide, Adelaide 5005, Australia (e-mail: Christine.O'Keefe@csiro.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2007.900808

I. Introduction

IN VARIOUS real-world applications, it is quite useful to find *unanticipated episodes* where event patterns unexpectedly lead to outcomes. For example, users unexpectedly turn to use Yahoo! search engine after navigating through a topic path in Yahoo! Web directory, which may imply a difficulty on finding Web content pages by following the topic path. Another example is taking the drug rofecoxib to relieve signs and symptoms of arthritis, but then unexpectedly experiencing myocardial infarction [1]. Detecting these *unanticipated episodes* is of great value in correction or prevention, especially if outcomes are life threatening. Due to unexpectedness, such an episode may not necessarily occur as an event pattern confidently implying an outcome as in temporal association rules (TARs) [2], [3] or sequential patterns [4]–[6]. In addition, *unanticipated episodes* normally occur infrequently, otherwise they become expected. The infrequency makes existing frequent itemsets/sequential patterns mining techniques ineffective. Thus, finding these unexpected and infrequent episodes necessitates innovative knowledge representations and mining techniques. In this paper, we introduce unexpected temporal association rules (UTARs) to describe these *unanticipated episodes*, and provide an algorithm for discovering them.

Promptly generating adverse drug reaction (ADR) signals from healthcare administrative databases provides a real-world example to illustrate the usefulness of our techniques. As recommended by the International Committee on Harmonization (ICH), "all noxious and unintended responses to a medicinal product related to any dose should be considered ADRs. The phrase 'responses to a medicinal product' means that a causal relationship between a medicinal product and an adverse event is at least a possibility [7]." Most ADRs are infrequent. Such low incidence is of course expected due to the fact that drugs are tested prior to release onto the market. However, due to limited patient numbers and trial duration of this screening process, ADRs with incidence rates less than 0.1% are normally not detected [8], [9], such as myocardial infarction caused by rofecoxib [1]. As a whole, e.g., it is estimated that millions of patients are hospitalized due to adverse events in the USA each year [10] and more than 80 000 in Australia [11]. ADRs are a major cause of morbidity and mortality worldwide [10], and 30%–60% ADR cases are believed to be preventable by careful prescribing and monitoring [12]. Thus, such ADR patterns as a drug probably causing a symptom/condition, can play a key role in the prevention or correction. Using these ADR patterns, e.g., computerized systems can search health records to monitor

adverse events [12], to find patient groups at risk [13], and to help general practitioners (GPs) ameliorate their diagnoses and prescriptions [8].

Existing postmarket ADR detection techniques, known as *signal generation* in *pharmacovigilance*, mainly work on spontaneous ADR case reports, submitted voluntarily by medical practitioners about observed suspected causalities between drug usage and adverse reactions [9]–[14]. However, in spontaneously reporting systems like the Australian ADR Reporting System [15], medical practitioners significantly underreport ADR cases, typically by a factor of about 20 [12], [16]. Adverse reactions may go unnoticed until large numbers of users have been affected [17]. In contrast, healthcare administrative data routinely record events about patients' interactions with a healthcare system for management and accounting purposes. In Australia, e.g., almost all medical services for almost entire population are included in these databases [18]. It is desirable to develop techniques to promptly and systematically signal (and then validate) ADRs from these databases. They can complement the existing postmarket ADR detection techniques, especially on rare ADRs resulting in serious outcomes such as hospitalization or disability. Our proposed techniques make a successful attempt to signal ADRs from healthcare administrative data, which is a brand new ADR signal generation direction in the literature.

We propose UTAR, denoted by $A \overset{T}{\hookrightarrow} C$, to describe an *unanticipated episode* where an event pattern *A* unexpectedly occurs in a *T*-sized period prior to another event pattern *C*. The period length *T* constrains the temporal relation between *the antecedent A* and *the consequent C*, and so, ensures the UTARs' plausibility. To handle the unexpectedness, we introduce an interestingness measure, *residual-leverage* and give a case-based exclusion technique for its calculation. The basic idea is to exclude "expected" events in individual *T*-sized subsequences, and then, aggregate unexpectedness over all the remaining *T*-sized subsequences. We further use an event-oriented data preparation technique to handle the infrequency. We then establish a new algorithm, MUTARC, to discover pairwise UTARs. Our proposed techniques are in principle extendible to longer patterns, such as drug–drug interactions causing symptoms. We apply MUTARC to signal ADRs from a healthcare administrative data set of prescribed drugs and diagnoses. It shortlists six known ADRs and another ADR, flucloxacillin $\overset{T}{\hookrightarrow}$ hepatitis, that has been previously unknown to our algorithm designers and experiment runners before the experiments. The MUTARC also empirically outperforms OPUS_AR$^+$ (extended from OPUS_AR [19]) for signaling ADRs. Medical experts believe that the proposed techniques are promising in the brand new direction of ADR signal generation.

The rest of the paper is organized as follows. We propose UTARs to represent *unanticipated episodes* in Section II, and establish the MUTARC to discover the most interesting UTARs in Section III. Typical results and reliability examination are presented in Section IV. Related work is discussed in Section V, followed by concluding comments in Section VI.

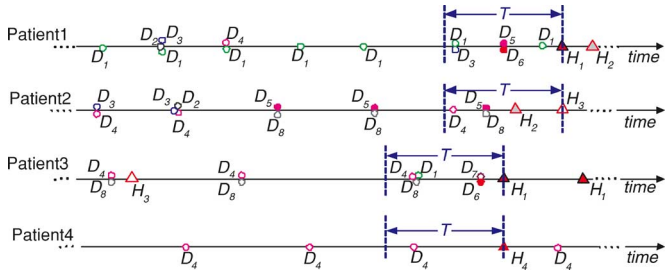

Fig. 1. Illustration of temporal event sequences and *T*-constrained subsequences. D_i and H_j indicate different categories of event types. For example, D_i indicates taking drug i, while H_j indicates an occurrence of condition j in healthcare administrative databases.

II. Problem Formulation

We first characterize the concept of *unanticipated episodes*. We assume that a number of subjects have their own sequences of events, each of which is a list of event types together with their timestamps as exemplified in Fig. 1. For simplicity, we use *event* and *event type* interchangeably hereafter. The effect of most types of events normally lasts for a short period of time, e.g., a Web visit session lasts minutes to hours, and drug usage usually produces adverse events within weeks to months [8]. We introduce a period length *T* to consider limited effect periods of related events in order to ensure the plausibility of *unanticipated episodes*. Then, an *unanticipated episode* involves two event patterns *A* and *C* such that *A* is unexpectedly followed by *C* relatively frequently within the *T*-sized periods. Relatively high frequency is emphasized because it is practicable to design algorithms to identify them from large real-world data that usually contain some noises, such as healthcare administrative data [18]. In addition, such an *unanticipated episode* is applicable to more subjects and, as knowledge, is of more value. Both *A* and *C* can be existence patterns (i.e., a set of events) or sequential patterns (i.e., an ordered list of events). For simplicity, we will only discuss existence patterns hereafter.

We then try to use TARs to describe these *unanticipated episodes*. Association rules are implications in the form of $A \rightarrow C$, meaning that the presence of *A* implies the presence of *C*, where *A* and *C* are mutually disjoint [20]. By embedding temporal constraints into association rules, we suggest a category of TARs denoted by $A \overset{T}{\rightarrow} C$. The notation $\overset{T}{\rightarrow}$ is used to indicate explicitly that *the antecedent A* and/or *the consequent C* occur within subsequences constrained by time windows of length *T*. To simplify temporal constraints, we only choose one subsequence within a *T*-sized time window from each sequence, and call it *T-constrained subsequence*. The *T*-constrained subsequences for patients 1 and 2 in Fig. 1, e.g., are $\{D_1, D_3, D_5, D_6, H_1\}$ and $\{D_4, D_5, D_8, H_2, H_3\}$, respectively. Given a set of *T*-constrained subsequences Θ_T, the *support* of a TAR $supp(A \overset{T}{\rightarrow} C)$ is the proportion of *T*-constrained subsequences in which A occurs prior to C at least once. For the four subsequences in Fig. 1, e.g., $supp(D_1 D_6 \overset{T}{\rightarrow} H_1) = 2/4$. Similarly, its *confidence* $conf(A \overset{T}{\rightarrow} C) = supp(A \overset{T}{\rightarrow} C)/supp(A \overset{T}{\rightarrow})$, where $supp(A \overset{T}{\rightarrow})$ indicates

the proportion of *T*-constrained subsequences that contain A in Θ_T. As another measure of association strength [19]–[21], *leverage* can be defined as the proportion of *T*-constrained subsequences that exhibit the association between the antecedent A and the consequent C in excess of those that would be expected if A and C were independent of each other. That is,

$$leverage(A \xrightarrow{T} C) = supp(A \xrightarrow{T} C) - supp(A \xrightarrow{T}) \cdot supp(\xrightarrow{T} C) \quad (1)$$

where $supp(\xrightarrow{T} C)$ indicates the proportion of *T*-constrained subsequences that contain C in Θ_T. For the four *T*-constrained subsequences in Fig. 1, e.g., $leverage(D_1 D_6 \xrightarrow{T} H_1) = 2/4 - 2/4 \times 2/4 = 1/4$. Similarly, $leverage(D_1 \xrightarrow{T} H_1) = leverage(D_6 \xrightarrow{T} H_1) = 1/4$. D_1 and D_6 have the same association strength with respect to H_1 based on leverage. Using (1) and noting that $0 \leq supp(\xrightarrow{T} C)$ and $supp(A \xrightarrow{T}) \leq 1$, we have the following inequality between the three measures:

$$leverage(A \xrightarrow{T} C) \leq supp(A \xrightarrow{T} C) = conf(A \xrightarrow{T} C). \quad (2)$$

There are two different strategies for mining TARs in order to discover *unanticipated episodes*. The first one is to find *valid TARs* whose support and confidence exceed prespecified thresholds θ_s and θ_c, respectively [20]. The second one is to find the most interesting rules based on some interestingness measure [19].

As discussed in Section I, *unanticipated episodes* including ADRs in pharmacotherapy occur normally at low frequency because of their unexpectedness as well as human regulatory processes. Thus, there are several drawbacks for the first strategy.

1) The support threshold θ_s and the confidence threshold θ_c should be set very small. This leads to innumerable valid TARs, and makes any mining algorithms unmanageable and the computation cost disproportionally high [20].
2) It will generate overwhelming volume of possibly useless results.
3) It is still not easy to set these threshold values appropriately [19].

Thus, it is quite complicated to identify *unanticipated episodes* from valid TARs. The situation is similar for other temporal data mining models, e.g., sequential patterns [4]–[6] or event-driven sequential patterns [22], [23].

Following the second strategy, we can identify *unanticipated episodes* through shortlisting most interesting rules. We may, e.g., simply apply OPUS_AR [19] on the *T*-constrained subsequence set Θ_T to generate the most interesting TARs. We call this algorithm OPUS_AR^+. It can return a prespecified number of TARs that maximize an association quality measure such as leverage.

The existing interestingness measures, such as support, confidence, risk ratio, odds ratio [8], lift, leverage [19], etc., are not suitable for highlighting TARs for detecting *unanticipated episodes*. As discussed earlier, due to their infrequency and unexpectedness, supports and confidences for TARs corresponding to *unanticipated episodes* are normally low. It seems impracticable to identify *unanticipated episodes* by choosing TARs with high support or confidence. The risk and the odds ratios appeal to domain experts, and are commonly used in effect evaluation [16]. For example, the risk ratio for $A \xrightarrow{T} C$ is $RR(A \xrightarrow{T} C) = [supp(A \xrightarrow{T} C)/supp(A \xrightarrow{T})]/[supp(\neg A \xrightarrow{T} C)/supp(\neg A \xrightarrow{T})]$, where $\neg A$ indicates that A does not occur. This describes the degree to which, within the *T*-constraint, the occurrence of C increases with the occurrence rate of A. However, the risk and the odds ratios are not suitable for detecting *unanticipated episodes* like ADRs from healthcare administrative databases. The first reason is that they have been widely used in effect evaluation, especially premarket drug testing; so, drugs having high risk or odds ratios may not be approved to release onto the market. Thus, the risk or odds ratios for ADRs in healthcare administrative databases are relatively low. This, as also empirically shown in Section IV, makes them inappropriate for highlighting *unanticipated episodes* like ADRs. The second reason is that, due to data noises, biases, or incompleteness,[1] the ranking based on these ratios may not be reliable. This point is also applicable to *lift* that may be defined as $lift \quad (A \xrightarrow{T} C) = supp(A \xrightarrow{T} C)/supp(A \xrightarrow{T}) supp(\xrightarrow{T} C)$. Furthermore, leverage can not express the unexpectedness that we need for detecting *unanticipated episodes*. Based on the leverage values, e.g., paracetamol (N02BE01 in Tables II and VI in Section IV-B is found to be strongly associated with two conditions, esophagitis and angioedema. However, paracetamol is widely used in the treatment of mild to moderate pain and fever, and may have been given as part of the treatment for these two conditions. These associations are not unexpected.

Thus, we need to introduce a knowledge representation for *unanticipated episodes*. Our strategy is to embed unexpectedness into rules directly. To clearly indicate temporal unexpectedness, we introduce an *UTAR*, denoted by $A \overset{T}{\hookrightarrow} C$, which means that *the antecedent* A occurs unexpectedly within a *T*-sized period prior to *the consequent* C. Rather than defining unexpectedness explicitly, we aggregate it from individual sequences.

Definition 1: The *support* of the UTAR, $supp(A \overset{T}{\hookrightarrow} C)$, is the proportion of *T*-constrained subsequences that unexpectedly contain A followed by C among all of the *T*-constrained subsequences. Its confidence is given by $conf(A \overset{T}{\hookrightarrow} C) = supp(A \overset{T}{\hookrightarrow} C)/supp(A \overset{T}{\hookrightarrow})$, where $supp(A \overset{T}{\hookrightarrow})$ is the proportion of *T*-constrained subsequences that unexpectedly contain A.

According to Definition 1, only the subsequences that unexpectedly contain A will contribute to support or confidence of $A \overset{T}{\hookrightarrow} C$. There exist various ways to check whether A is unexpectedly contained in a subsequence. Instead of judging this directly, we may remove "expected" event types and roughly keep others as "unexpected ones." For example, if we know that a drug A is prescribed to treat a condition C, A can

[1]For example, the Queensland Linked Data Set, which is used for testing this research, does not contain any other symptoms/conditions/diagnoses except diagnoses from inpatient episodes during a 4-year period. It does not contain any drug prescriptions within inpatient episodes [18], [24].

be removed for the purpose of finding ADRs with respect to C, as C is not unexpected with respect to taking A. Another example is, for patient 1 in Fig. 1, a drug D_1 is taken frequently within a T-sized period prior to, and also, far before the unique condition H_1, and it is difficult to say that this sequence favors the unexpected temporal association $D_1 \overset{T}{\hookrightarrow} H_1$ without prior knowledge. But D_5 and D_6 only occur just before H_1, and it is reasonable to say that this sequence favors $D_5 D_6 \overset{T}{\hookrightarrow} H_1$. Thus, the sequence information outside of a subsequence can be used to prune "expected" event types. The remaining subsequences can then be aggregated together to express the unexpectedness of UTARs. For the four subsequence in Fig. 1, e.g., we can have $supp(D_1 \overset{T}{\hookrightarrow} H_1) = 1/4$ while $supp(D_1 \overset{T}{\rightarrow} H_1) = 2/4$. A possible calculation method based on a simple case-based exclusion operation will be given in Section III.

Note that $0 \leq supp(A \overset{T}{\hookrightarrow}) \leq 1$, we have

$$supp(A \overset{T}{\hookrightarrow} C) \leq supp(A \overset{T}{\rightarrow} C) \tag{3}$$

$$supp(A \overset{T}{\hookrightarrow}) \leq supp(A \overset{T}{\rightarrow}) \tag{4}$$

$$supp(A \overset{T}{\hookrightarrow} C) \leq conf(A \overset{T}{\hookrightarrow} C). \tag{5}$$

We now introduce a new interestingness measure, residual-leverage.

Definition 2: The *residual-leverage* (*resilev*) of the UTAR, $A \overset{T}{\hookrightarrow} C$, is the proportion of *T*-constrained subsequences that exhibits the unexpected association between A and C in excess of those that would be supposed if unexpected A and C were independent of each other. That is

$$resilev(A \overset{T}{\hookrightarrow} C) = supp(A \overset{T}{\hookrightarrow} C) - supp(A \overset{T}{\hookrightarrow}) \times supp(\overset{T}{\rightarrow} C). \tag{6}$$

It intuitively indicates the degree to which the observed unexpected associations between A and C exceeds supposed associations based on an independence assumption. It considers observed unexpected associations between A and C, while leverage considers observed associations between them. We have

$$conf(A \overset{T}{\hookrightarrow} C) \geq supp(A \overset{T}{\hookrightarrow} C) \geq resilev(A \overset{T}{\hookrightarrow} C). \tag{7}$$

Similar to support and confidence, we can set a threshold for residual-leverage to define a *valid UTAR*. This threshold may increase the likelihood of finding interesting UTARs. Once again, the number of such valid UTARs can be too large if the threshold is set too low. Conversely, really interesting UTARs are missed if it is set too high.

To relieve the problem of setting thresholds completely, we simply select a prespecified number of, say ten, UTARs with the highest residual-leverage values. The second reason to rank UTARs only according to residual-leverage is that large residual-leverage also indicates large support and confidence as indicated by (7) and

$$\begin{aligned} conf(A \overset{T}{\rightarrow} C) &\geq supp(A \overset{T}{\rightarrow} C) \geq supp(A \overset{T}{\hookrightarrow} C) \\ &\geq resilev(A \overset{T}{\hookrightarrow} C). \end{aligned} \tag{8}$$

The inequality in (8) is based on (2), (3), and (7). This guarantees that the generated UTARs will have reasonable support, and they will correspond better with relatively frequent *unanticipated episodes*, as discussed at the beginning of this section. In addition, large residual-leverage will also imply large leverage, as shown in Theorem 1 in Section III, when the case-based exclusion operation is used.

III. Searching for Unexpected Temporal Association Rules

In this section, we develop a simple but effective algorithm to search for the most interesting UTARs. We concentrate on pairwise UTARs, such as an ADR, where one single drug possibly induces one condition, i.e., drug $A \overset{T}{\hookrightarrow}$ condition C. Such pairwise UTARs are of great application value, and some successful experience with them can pave the way for us to discover more sophisticated UTARs in the future.

Our proposed algorithm, the MUTARC, is outlined in Algorithm 1.

Algorithm 1: Mining UTARs given the Consequent (MUTARC)

1) initialize parameters, such as the consequent C, event types of interest, the study period $[t_S, t_E]$, time period lengths T_h, T_r, T_b, and T_c, and the number of output UTARs k;
2) prepare case subsequences from case sequences that have the first occurrence of C during the study period: for each case, choose events within its hazard period, and exclude some of them based on the case-based exclusion with respect to the consequent C;
3) choose noncase subsequences within control periods from noncase sequences;
4) calculate supports and residual-leverage of each event; and
5) rank the events in the descending order of residual-leverage, and return the top k interesting UTARs.

Its basic idea is to choose subsequences around a given consequent, remove "expected" events according to a case-crossover design, and then, calculate residual-leverage values. The techniques for mining UTARs around a given antecedent are discussed in [24]. We take the sequences in Fig. 1 as examples to explain MUTARC later.

First, we initialize parameters that are explained as follows.

1) The consequent C is specified to restrict the search space so as to facilitate mining *unanticipated episodes*. The sequences containing C are called *case sequences* while other sequences are called *noncase sequences*. In Fig. 2, e.g., H_1 is specified, and then, patients 1 and 3 in Fig. 1 are cases and patients 2 and 4 are noncases, respectively.
2) Event types of interest are to limit the possible candidates for the antecedent A, e.g., $D_1 - D_8$ in Fig. 2.
3) A study period is determined by $[t_S, t_E]$. A case sequence whose first occurrence of the consequent C is not in the study period will simply not be considered. We restrict ourselves to the first occurrence in this paper in order to facilitate the implementation of the case-based exclusion and choosing at most one *T*-constrained subsequence for

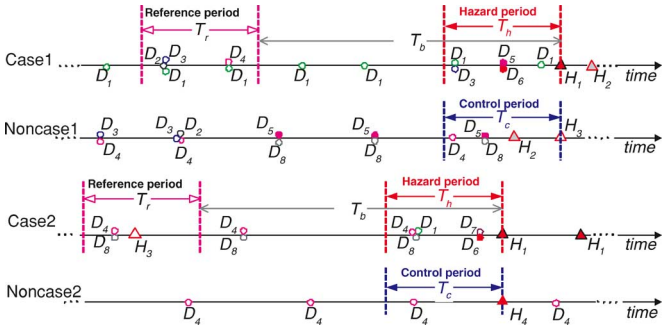

Fig. 2. Illustration of the event-oriented data preparation and case-based exclusion in the MUTARC. D_i and H_j indicate a drug-taking event and a condition occurrence, respectively. H_1 indicates the specified consequent in this example.

each case sequence. Another reason is that there is only a small population of patients with multiple occurrences of the consequence C in the data set that we will study.

4) The time lengths T_h, T_r, T_b, and T_c indicate, as illustrated in Fig. 2, *the hazard period length*, *the reference period length*, *the period length between the hazard and the reference periods*, and *the control period length*, respectively.

We set the hazard period for each case sequence as the T_h-sized period before the first occurrence of the consequent C that occurs in the study period. Clearly, the events within the hazard period might cause the consequent C. Among them, some are less likely to unexpectedly lead to C. If domain experts can provide a list of event types that do not unexpectedly lead to C, we may simply exclude them from the case subsequences. But, such kind of domain knowledge is often unavailable or out of date [25]. Fortunately, for each case sequence, we can use the events that occurred outside of the hazard period to deduce some event types that do not unexpectedly lead to C. For example, if an event occurs repeatedly in the case sequence, say, if the subject often takes one drug like D_1 for Case 1 in Fig. 2, this event is less likely to induce the first occurrence of the consequent C. The underlying reason is that most events, such as taking a drug, have short-term effects [26], and the subject has similar responses to a certain event. Thus, we may disregard this event from the subsequence, and deduce that the remaining ones are more likely to induce C unexpectedly. This exclusion operation is carried out only based on a single case sequence, and is termed as *case-based exclusion*. To further simplify this exclusion operation, we borrow the concept of the reference period from case-crossover studies [26]. The reference period is a T_r-sized period that is a T_b-sized interval before the hazard period as illustrated in Fig. 2. If the event (e.g., taking a drug) in the reference period is protective/therapeutic to the consequent C, the subject is not surprised about the occurrence of C. In contrast, the subject (or his/her GPs) will believe that the events within the reference period are basically "safe" to him/her because there is no occurrence of an unexpected outcome, like C, even long after the reference period. Thus, the events within the reference period are probably expected to the subject with respect to the consequent C, and they can be excluded for mining UTARs. For case 1 in Fig. 2, e.g., $\{D_1, D_2, D_3, D_4\}$ are in the reference period, and $\{D_1, D_3, D_5, D_6\}$ in the hazard period. D_1 and D_3 are excluded, and only $\{D_5, D_6\}$ are kept for the subsequence for case 1. Similarly, $\{D_1, D_6, D_7\}$ are left for case 2.

In Step 3 of the MUTARC, for each noncase sequence, we have two different methods to set a T_c-sized control period in order to choose a noncase subsequence. We may randomly choose the control period within $[t_S - T_c, t_E]$. The other one is motivated by matched case-control studies [8] in order to avoid some possible impact from other factors such as age, gender, and seasonality. We set the control period to match a case according to, say, demographic data. A noncase is chosen from the same demographic (e.g., age-gender) stratum as the case. In addition, the noncase has an event similar to C that also occurs temporally closely to C in the case, say, the onset of another condition in the same month of the onset of the condition C. For example, noncase 1 in Fig. 2 has H_3 that occurs in the same month with H_1 of case 1; thus, the T_c-sized period before H_3 is set as the matched control period for noncase 1. With respect to case 2, the matched control period for noncase 2 is the T_c-sized period before H_4. The events within the control periods, say, D_4, D_5, and D_8 for noncase 1, and D_4 for noncase 2, compose noncase subsequences. These noncase subsequences roughly provide baseline frequency information for these event types.

In Step 4, putting these case subsequences after exclusion and these noncase subsequences together, we then calculate supports for each event with respect to C. Now $supp(A \overset{T}{\hookrightarrow} C)$ takes into account only the case subsequences that more likely unexpectedly contain A followed by C. We then compute the residual-leverage value of each event according to (6). We simply assume that a noncase subsequence "unexpectedly" contains A once it contains A. For the four sequences in Fig. 2, e.g., $resilev(D_1 \overset{T}{\hookrightarrow} H_1) = supp(D_1 \overset{T}{\hookrightarrow} H_1) - supp(D_1 \overset{T}{\hookrightarrow}) \times supp(\overset{T}{\hookrightarrow} H_1) = 1/4 - 1/4 \times 2/4 = 1/8$ and $resilev(D_6 \overset{T}{\hookrightarrow} H_1) = 2/4 - 2/4 \times 2/4 = 1/4$. Thus, residual-leverage roughly indicates the situation where D_6 is more unexpectedly associated with H_1 than D_1.

Finally, the algorithm outputs the k most interesting event types. Together with the consequent C, we have the k most interesting UTARs.

In the MUTARC, the hazard and the control periods are set according to the consequent C, and restrict the events in the calculation of supports. This *event-oriented data preparation* makes it possible to highlight the usually infrequent UTARs. The *case-based exclusion* in Step 2 is designed to prune expected events from a case's viewpoint. It is simple and easy for implementation, but it plays a key role in the good performance of the MUTARC. To highlight its contribution, we implement and parameterize OPUS_AR$^+$ same as MUTARC except without the case-based exclusion operation. That is, the only difference between them is with or without the operation. Then, we compare MUTARC and OPUS_AR$^+$ comprehensively. The following theorem indicates that, using this case-based exclusion, the residual-leverage value is not greater than the leverage value.

Theorem 1: With the case-based exclusion, $resilev(A \overset{T}{\hookrightarrow} C) \leq leverage(A \overset{T}{\rightarrow} C)$.

Proof: Let the proportion of case subsequences where A is excluded be $supp_{\rm ex}(\geq 0)$. Because of the case-based exclusion, we observe

$$supp(A \xrightarrow{T} C) - supp(A \overset{T}{\hookrightarrow} C) = supp_{\rm ex} \quad \text{and} \tag{9}$$

$$supp(A \xrightarrow{T}) - supp(A \overset{T}{\hookrightarrow}) = supp_{\rm ex}. \tag{10}$$

Considering the definition of support, we have another inequality $0 \leq supp(\xrightarrow{T} C) \leq 1.0$. Simply, according to the definitions of residual-leverage and leverage, we have

$$\begin{aligned} leverage(A \xrightarrow{T} C) - resilev(A \overset{T}{\hookrightarrow} C) &= supp(A \xrightarrow{T} C) \\ &- supp(A \overset{T}{\hookrightarrow} C) - (supp(A \xrightarrow{T}) - supp(A \overset{T}{\hookrightarrow})) \\ &\times supp(\xrightarrow{T} C) = supp_{\rm ex} - supp_{\rm ex} \times supp(\xrightarrow{T} C) \geq 0. \end{aligned} \tag{11}$$

The proof is completed. ∎

Combining Theorem 1 with (2), we also have $conf(A \xrightarrow{T} C) \geq supp(A \xrightarrow{T} C) \geq resilev(A \overset{T}{\hookrightarrow} C)$. This theorem says that large residual-leverage indicates large confidence. For example, if $resilev(A \overset{T}{\hookrightarrow} C) \geq \theta$, then $leverage(A \xrightarrow{T} C) \geq \theta$, too. Thus, as discussed in Section II, we can simply choose the most interesting UTARs only based on residual-leverage, and relieve the problem of setting appropriate thresholds.

IV. Experimental Setting and Results

A. Linked Healthcare Administrative Data: The QLDS

The Commonwealth Scientific and Industrial Research Organization (CSIRO), through its Division of Mathematical and Information Sciences, was commissioned by the now Australian Government Department of Health and Ageing (DoHA) in August 2002 to analyze a linked data set produced from Medicare Benefits Scheme (MBS), Pharmaceutical Benefits Scheme (PBS), and Queensland hospital morbidity data, more commonly referred to as the Queensland Linked Data Set (QLDS) [18]. The objective was to provide a demonstration of the utility of data mining on deidentified administrative health data to investigate patterns of utilization, adverse events, and other health outcomes.

The QLDS was made available to CSIRO under a negotiated agreement between the DoHA and Queensland Health. The data set contained deidentified and confidentially linked patient level hospital separation data (July 1, 1995 to June 30, 1999), MBS data and PBS data (both January 1, 1995 to December 31, 1999). All data were deidentified, and actual dates of service were removed, so that time sequences were indicated only by time from first admission that was perturbed up to 16 days. This process provided strong privacy protection, consistent with the requirements of the relevant Federal and State legislations. The CSIRO held the QLDS in a secure computer environment and limited access to authorized staff directly involved in the data analysis.

The QLDS provides a real-world data set appropriate for testing the proposed techniques to generate ADR signals. Each record in the hospital separation data corresponds to one inpatient episode, and each diagnosis is coded in the International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) system, e.g., 530 is for esophagitis. Each record in the MBS data corresponds to a medical service for one patient. Similarly, each record in the PBS data corresponds to one prescription drug supplied to one patient, and the 3842 distinct prescription items are mapped into 758 distinct codes in the WHO anatomical therapeutic chemical (ATC) classification system [27]. For example, the ATC codes for alendronate,[2] nefazodone,[3] and Angiotensin Converting Enzyme (ACE) inhibitors[4] are M05BA04, N06AX06, and C09AA??[5] respectively [27]. The QLDS contains records for 1 176 294 patients. For convenience, we refer to January 1, 1995 as the first day hereinafter. Thus, the time period for the whole data set is [1, 1826].

To speed up data access, for each patient, a temporal event sequence was generated to record sequentially his/her hospitalization, PBS, and MBS events, as well as their timestamps. The head of each sequence also included demographic data such as age and gender. These sequences were stratified into six age-gender strata according to their demographic data. Ages were stratified into young ([0, 20), middle-aged ([20, 59]), and older (59) groups. Thus, we directly accessed these sequences rather than the three linked administrative databases in the QLDS. Both the MUTARC and OPUS_AR$^+$ were implemented in Python. All experiments were run under Linux on a machine with an Intel Pentium 4 3.2 GHz processor.

B. Typical Experimental Results

We first describe typical results generated by the MUTARC. To highlight the significance of the case-based exclusion operation, we compare the MUTARC with OPUS_AR$^+$ where the only difference is without the operation. We concentrate on three types of diagnoses, esophagitis, hepatitis, and angioedema, which are suspected to be sometimes induced by alendronate, nefazodone, and ACE inhibitors, respectively [15], [28]. These are all the ADRs that our algorithm designers and experiment runners (all the authors except Chris) have known before these experiments.

Like other data mining results, it is unrealistic to expect every highly ranked UTAR to be of value or significance to domain experts, especially considering the intrinsic data biases, noises,

[2] Alendronate (Fosamax) is an aminobisphosphonate, which specifically inhibits osteoclast-mediated bone resorption. It was approved for treatment of osteoporosis in postmenopausal women and Paget's disease of bone [28].

[3] Nefazodone, (Serzone) is one of the new antidepressants. It was marketed in mid-1997 and withdrawn in 2004 in Australia. It is related to the selective serotonin reuptake inhibitors (SSRIs), a widely used group of antidepressants), but has a dual action in that it works on both sides of the serotonin synapse [15].

[4] ACE inhibitors (ACE inhibitors) are a commonly used treatment of high blood pressure and heart disease [29].

[5] Here, "?" is a wildcard. There are about nine different ATC codes for ACE inhibitors, from C09AA01 to C09AA10 except C09AA07. For example, C09AA03 represents lisinopril [27].

TABLE I
TOP TEN DRUGS UNEXPECTEDLY ASSOCIATED WITH Esophagitis SHORTLISTED BY MUTARC FOR OLDER FEMALE PATIENTS (THE BOLD ROW IS FOR THE SUSPECTED DRUG Alendronate)

Rank in		Drug in	Drug	Resi	Leve	$supp(A \overset{T}{\hookrightarrow})$	$supp(A \overset{T}{\hookrightarrow} C)$	Risk
Resilev	Leverage	ATC code	name	-lev	-rage	$\times N(N{=}58172)$	$\times N(N{=}58172)$	ratio
1	4	A03FA02	*Cisapride*	4.37E-3	1.25E-2	1790	534	3.245
2	15	J07AL01	Pneumococcal	3.79E-3	3.92E-3	1854	510	1.829
3	13	A02BA04	Nizatidine	2.48E-3	4.06E-3	995	300	2.434
4	**31**	**M05BA04**	**Alendronate**	**1.51E-3**	**2.28E-3**	**628**	**186**	**2.279**
5	2	A02BC01	Omeprazole	1.50E-3	1.55E-2	1528	326	3.692
6	44	C09CA04	Irbesartan	1.50E-3	1.54E-3	658	190	1.886
7	35	B01AC06	Aspirin	1.43E-3	1.85E-3	1035	245	1.667
8	14	A03FA01	*Metoclopramide*	1.31E-3	4.00E-3	2288	434	1.645
9	49	C10AA05	Atorvastatin	1.25E-3	1.47E-3	763	192	1.718
10	42	A03FA03	*Domperidone*	7.75E-4	1.56E-3	524	127	2.023

and incompleteness of the QLDS[1] [18]. One realistic goal is to reliably shortlist the unexpected associations between prescribed drugs and diagnoses among the 10 or 20 highest ranked UTARs. These shortlisted UTARs can be regarded as ADR hypotheses, and have to be further evaluated and validated by pharmacovigilance experts [8], [9]. For example, statistical tests like Fisher exact test on independent data can help us remove UTARs caused by biases or noises. We only discuss those results consistent with existing domain knowledge in this paper.

As for the parameters of the MUTARC and OPUS_AR$^+$, we set $T_h = 180$, $T_r = 3T_h, T_c = T_h$, and $T_b = 365$ in days by default. T_h and T_c were set as about six months for acute or subacute ADRs [8]. T_b was set as about 1 year in order to eliminate seasonality in drug usage. We set the study period as [730, 1645], where 1645 is for the last hospitalization event in the QLDS. For esophagitis, we focused on the older female group, because 9092 out of 33094 esophagitis patients are female and older. For each older female esophagitis patient, we chose up to six (only five are available for some of them) different matched non-esophagitis patients who were from the same age-gender group as the esophagitis patient and hospitalized in the same month as the esophagitis onset. There were 49080 non-esophagitis subsequences. Then, the total number of subsequences N was 58172. For 2912 hepatitis patients, we were interested in 1034 middle-aged female hepatitis patients. We chose up to 21 matched non-hepatitis subsequences for each of them and got 21660 non-hepatitis subsequences in total. For 286 angioedema patients, we focused on 75 female and 41 male angioedema patients older than 59. Up to 21 matched non-angioedema patients were chosen for each of them. There were 1515 female and 795 male non-angioedema patients, respectively.

Table I lists the top ten drugs having the highest residual-leverage values with respect to esophagitis generated by the MUTARC for the older female patients. After the case-based exclusion, there are 186 ($= supp(A \overset{T}{\hookrightarrow} C) \times N$) esophagitis patients from 628 ($= supp(A \overset{T}{\hookrightarrow}) \times N$) alendronate drug users. Comparing the supports listed in Table II where some drugs are ranked based on leverage generated by the OPUS_AR$^+$, we can observe that there are only 53 ($= 239 - 186$) patients taking alendronate within both reference and hazard periods. The UTAR alendronate $\overset{T}{\hookrightarrow}$ esophagitis has the residual-leverage of 1.51×10^{-3}, and, based on this, is ranked as 4 among 758 different kinds of drugs. As a comparison, the TAR alendronate $\overset{T}{\rightarrow}$ esophagitis is ranked as low as 31 in Table II. This is partially because there are only 239 older female patients who took alendronate within their hazard periods, which is much smaller than other drug users in Table II. The support for this TAR is 0.41%, and it is ranked as low as 91. Though its confidence is as high as 35.1% (due to the event-oriented data preparation), this association is ranked as low as 43 based on the confidence value. Its risk ratio is 2.279. It means an older female patient, if taking alendronate, is about 2.279 times more likely to suffer from esophagitis. The rank simply based on the risk ratio is 44. Thus, the residual-leverage can highlight this ADR much better than these other measures. Similar situations can be observed for the other ADRs. On the other hand, paracetamol (N02BE01) is ranked as no. 6 based on its leverage value, and is thought to be strongly associated with esophagitis, as shown in Table II. According to residual-leverage, it is ranked as low as 546, because about 82.2% cases took paracetamol in the reference periods. Note that similar interesting situations happen with several therapeutic drugs like ranitidine, famotidine, and aluminium hydroxide in Table II. Furthermore, the MUTARC had a runtime of 57.2 s that is only 3.3 s longer than OPUS_AR$^+$. These comparisons empirically support the proposed measure, residual-leverage, is useful in removing some protective/therapeutic drugs[6] automatically and effectively for ADR signal generation.

With respect to hepatitis for the middle-aged female patients, Table III lists the ten drugs having the highest residual-leverage values generated by the MUTARC, while Table IV lists drugs having the highest leverage values generated by

[6]Besides the simplified design of the case-based exclusion in the MUTARC and the data quality issues like noises, biases, and incompleteness[1], there exists another interesting reason why the MUTARC cannot remove all protective/therapeutic drugs from the shortlisted UTARs: "treatment failures" may not be distinguished from ADRs only based on data without any prior knowledge. For example, the promotility drugs, including cisapride, metoclopramide, and domperidone in Table I, are reserved either for patients who do not respond to other treatments or are added to enhance other treatments for gastroesophageal reflux disease (see http://www.medicinenet.com/gastroesophageal_reflux_disease_gerd/page6.htm). Thus, it is not surprising to see that there are still many patients suffering esophagitis after taking these drugs. These are not ADRs but "treatment failures" according to our medical experts.

TABLE II
SOME DRUGS STRONGLY ASSOCIATED WITH Esophagitis GENERATED BY OPUS_AR^+ FOR OLDER FEMALE PATIENTS

Rank in		Drug in	Drug	Leve	Resi	Risk	$supp(A \xrightarrow{T})$	$supp(A \xrightarrow{T} C)$
Leverage	Resilev	ATC code	name	-rage	-lev	ratio	$\times N(N{=}58172)$	$\times N(N{=}58172)$
1	488	A02BA02	Ranitidine	2.03E-2	-2.32E-3	2.263	8223	2468
2	5	A02BC01	Omeprazole	1.55E-2	1.50E-3	3.692	2491	1289
3	545	J07BB02	Influenza Vaccine	1.42E-2	-1.63E-2	1.708	10128	2407
4	1	A03FA02	Cisapride	1.25E-2	4.37E-3	3.245	2348	1092
5	402	A02BA03	Famotidine	1.03E-2	-4.76E-4	2.142	3848	1198
6	546	N02BE01	Paracetamol	8.46E-3	-2.93E-2	1.282	17095	3164
7	433	A02AD	Aluminium Hydroxide	6.71E-3	-7.46E-4	1.917	3006	860
8	530	C01DA02	Glyceryl Trinitrate	5.49E-3	-5.98E-3	1.406	5808	1227
9	535	R03AC02	Salbutamol	4.33E-3	-7.37E-3	1.334	5483	1109
10	537	N02AA59	Codeine with Paracetamol	4.26E-3	-7.53E-3	1.274	6732	1300
...	...	...	...	...	...	...	...	...
31	**4**	**M05BA04**	**alendronate**	**2.28E-3**	**1.51E-3**	**2.279**	**681**	**239**

TABLE III
TOP TEN DRUGS UNEXPECTEDLY ASSOCIATED WITH Hepatitis SHORTLISTED BY MUTARC FOR THE MIDDLE-AGED FEMALE PATIENTS (THE BOLD AND THE SLANTED ROWS ARE FOR Nefazodone AND Flucloxacillin, RESPECTIVELY)

Rank in		Drug in	Drug	Resi	Leve	$supp(A \xrightarrow{T})$	$supp(A \xrightarrow{T} C)$	Risk
Resilev	Leverage	ATC code	name	-lev	-rage	$\times N(N{=}22694)$	$\times N(N{=}22694)$	ratio
1	7	N05CD02	Nitrazepam	5.03E-4	1.22E-3	276	24	3.156
2	15	N02AC	Diphenylpropylamine Derivatives	4.30E-4	6.40E-4	49	12	7.008
3	**25**	**N06AX06**	**Nefazodone**	**3.78E-4**	**3.78E-4**	**53**	**11**	**4.593**
4	*20*	*J01CF05*	*Flucloxacillin*	*3.51E-4*	*5.19E-4*	*352*	*24*	*1.746*
5	19	P01BC01	Quinine Bisulphate	3.30E-4	5.40E-4	187	16	2.429
6	3	N05BA04	Oxazepam	3.26E-4	3.23E-3	562	33	3.826
7	10	N06AB05	Paroxetine	3.11E-4	8.99E-4	394	25	2.140
8	21	N06AG02	Moclobemide	3.05E-4	5.15E-4	375	24	1.694
9	26	C02AC01	Clonidine	2.94E-4	3.78E-4	51	9	4.593
10	12	N06AB06	Sertraline	2.66E-4	6.87E-4	504	29	1.691

TABLE IV
SOME DRUGS ASSOCIATED WITH Hepatitis GENERATED BY THE OPUS_AR^+ FOR THE MIDDLE-AGED FEMALE PATIENTS

Rank in		Drug in	Drug	Leve	Resi	Risk	$supp(A \xrightarrow{T})$	$supp(A \xrightarrow{T} C)$
Leverage	Resilev	ATC code	name	-rage	-lev	ratio	$\times N(N{=}22694)$	$\times N(N{=}22694)$
1	36	N05BA01	Diazepam	5.39E-3	4.64E-5	3.910	1109	173
2	17	N05CD07	Temazepam	3.74E-3	4.75E-5	2.671	1296	144
3	6	N05BA04	Oxazepam	3.23E-3	5.80E-5	3.826	631	102
4	305	N02AA59	Codeine with Paracetamol	3.06E-3	2.13E-5	1.727	2557	186
5	20	A03FA01	Metoclopramide	1.46E-3	4.83E-5	1.853	919	75
6	276	J01DA01	Cephalexin	1.36E-3	3.87E-5	1.390	1958	120
7	1	N05CD02	Nitrazepam	1.22E-3	8.89E-5	3.156	293	41
8	306	J01CA04	Amoxycillin	1.13E-3	2.08E-5	1.247	2643	146
9	303	R03AC02	Salbutamol	9.84E-4	2.12E-5	1.300	1814	105
10	7	N06AB05	Paroxetine	8.99E-4	6.30E-5	2.140	408	39
...	...	...	...	...	...	...	...	...
20	*4*	*J01CF05*	*Flucloxacillin*	*5.19E-4*	*6.81E-5*	*1.746*	*356*	*28*
25	**3**	**N06AX06**	**Nefazodone**	**3.78E-4**	**2.44E-4**	**4.593**	**53**	**11**

the OPUS_AR^+. Eleven patients suffer hepatitis soon after taking nefazodone. Thus, the support for the UTAR nefazodone $\xrightarrow{T}$ hepatitis is as low as 0.05%. Its residual-leverage is 3.78×10^{-4}, and it is ranked 3 among 758 different kinds of drugs. As a comparison, the OPUS_AR^+ ranks the TAR nefazodone $\xrightarrow{T}$ hepatitis as low as 25. It is worth pointing out that rank 4 in Table III indicates another very interesting ADR flucloxacillin $\xrightarrow{T}$ hepatitis. The algorithm designers and experiment runners have not been aware of this ADR before medical experts checked the experimental results. According to the Australian Adverse Drug Reactions Bulletin, flucloxacillin is the most commonly reported to the Adverse Drug Reactions Advisory Committee (ADRAC) in association with hepatic reaction up to March 1996 [30]. As a comparison, the OPUS_AR^+ ranks this association as low as 20 in Table IV. On the other hand, the runtime of the MUTARC and OPUS_AR^+ is 18.6 and 18.3 s, respectively.

Table V lists the top ten suspected drugs that unexpectedly lead to angioedema generated by the MUTARC, while Table VI lists some drugs strongly associated with angioedema generated by the OPUS_AR^+. The runtime of the two algorithms is 1.9 and 1.8 s, respectively. Interestingly, there is only one ACE inhibitor, lisinopril (C09AA03), among the top ten drugs strongly associated with angioedema in Table VI. However, three from nine distinct ACE inhibitors, i.e., lisinopril, perindopril (C09AA04), and fosinopril (C09AA09) are within the top ten drugs shortlisted by the MUTARC. They are ranked 1, 8,

TABLE V
TOP TEN DRUGS UNEXPECTEDLY ASSOCIATED WITH Angioedema SHORTLISTED BY THE MUTARC FOR THE OLDER FEMALE PATIENTS (THE THREE BOLD ROWS ARE FOR THREE DISTINCT ACE Inhibitors)

Rank in		Drug in	Drug	Resi	Leve	$supp(A \overset{T}{\hookrightarrow})$	$supp(A \overset{T}{\hookrightarrow} C)$	Risk
Resilev	Leverage	ATC code	name	-lev	-rage	$\times N(N{=}1590)$	$\times N(N{=}1590)$	ratio
1	**2**	**C09AA03**	**Lisinopril**	**4.49E-3**	**5.68E-3**	**82**	**11**	**3.759**
2	12	G03CA01	Ethinyloestradiol	3.45E-3	3.45E-3	11	6	12.48
3	22	R06AD02	Promethazine	2.98E-3	2.98E-3	48	7	3.306
4	20	J01EA01	Trimethoprim	2.50E-3	3.10E-3	43	6	3.616
5	33	M02AC	Methyl Salicylate	2.37E-3	2.37E-3	26	5	4.296
6	17	J01DA08	Cefaclor	2.12E-3	3.32E-3	77	7	2.608
7	44	J01CA04	Amoxycillin	2.05E-3	2.05E-3	143	10	1.556
8	**29**	**C09AA04**	**Perindopril**	**1.99E-3**	**2.59E-3**	**39**	**5**	**3.369**
9	**48**	**C09AA09**	**Fosinopril**	**1.90E-3**	**1.90E-3**	**42**	**5**	**2.632**
10	49	A01AB04	Amphotericin	1.89E-3	1.89E-3	21	4	4.209

TABLE VI
SOME DRUGS ASSOCIATED WITH Angioedema GENERATED BY THE OPUS_AR^+ FOR THE OLDER FEMALE PATIENTS

Rank in		Drug in	Drug	Leve	Resi	Risk	$supp(A \overset{T}{\hookrightarrow})$	$supp(A \overset{T}{\hookrightarrow} C)$
Leverage	Resilev	ATC code	name	-rage	-lev	ratio	$\times N(N{=}1590)$	$\times N(N{=}1590)$
1	42	C03CA01	Frusemide	9.16E-3	7.71E-4	3.270	200	24
2	**1**	**C09AA03**	**Lisinopril**	**5.68E-3**	**4.49E-3**	**3.759**	**84**	**13**
3	199	C01DA02	Glyceryl Trinitrate	5.41E-3	-1.78E-3	2.475	157	16
4	210	J07BB02	Influenza vaccine	5.29E-3	-4.30E-3	1.731	373	26
5	193	G03CA57	Oestrogens conjugated	5.20E-3	-1.38E-3	3.643	79	12
6	93	N05CD02	Nitrazepam	4.48E-3	2.84E-4	3.856	61	10
7	204	R03AC02	Salbutamol	4.02E-3	-2.57E-3	2.171	140	13
8	100	A02AD	Aluminium hydroxide	3.86E-3	2.61E-4	2.829	82	10
9	211	N02BE01	Paracetamol	3.78E-3	-9.37E-3	1.436	466	28
10	18	J01FA06	Roxithromycin	3.65E-3	1.26E-3	2.312	110	11
...	...	...	...	...	...	...	...	...
29	**8**	**C09AA04**	**Perindopril**	**2.59E-3**	**1.99E-3**	**3.369**	**40**	**6**
48	**9**	**C09AA09**	**Fosinopril**	**1.90E-3**	**1.90E-3**	**2.632**	**42**	**5**

TABLE VII
TOP TEN DRUGS UNEXPECTEDLY ASSOCIATED WITH Angioedema SHORTLISTED BY THE MUTARC FOR THE OLDER MALE PATIENTS (THE TWO BOLD ROWS ARE FOR TWO DISTINCT ACE Inhibitors)

Rank in		Drug in	Drug	Resi	Leve	$supp(A \overset{T}{\hookrightarrow})$	$supp(A \overset{T}{\hookrightarrow} C)$	Risk
Resilev	Leverage	ATC code	name	-lev	-rage	$\times N(=836)$	$\times N(=836)$	ratio
1	**7**	**C09AA03**	**Lisinopril**	**4.57E-3**	**5.71E-3**	**24**	**5**	**5.561**
2	10	H02AB06	Prednisolone	4.02E-3	5.16E-3	13	4	8.155
3	21	A04AD	Other antiemetics	3.58E-3	3.58E-3	41	5	2.693
4	11	C03EA01	Hydrochlorothiazide agents	3.49E-3	4.63E-3	22	4	4.909
5	12	M01AB01	Indometacin	3.32E-3	4.45E-3	25	4	4.327
6	3	J01DA01	Cefalexin	3.31E-3	6.72E-3	66	6	3.126
7	**16**	**C09AA02**	**Enalapril**	**2.95E-3**	**4.09E-3**	**72**	**6**	**2.152**
8	24	A03FA01	Metoclopramide	2.36E-3	3.49E-3	21	3	4.00
9	33	R06AD02	Promethazine	2.36E-3	2.36E-3	21	3	3.064
10	34	D07AA02	Hydrocortisone	2.30E-3	2.30E-3	22	3	2.921

and 9, respectively. In addition, these ranks are much higher than their ranks in terms of leverage. For example, the rank of fosinopril improves from 48 to 9. Even for lisinopril, its rank advances from 2 to 1. Again paracetamol is ranked 9 based on leverage, but as low as 211 based on residual-leverage.

Similar comparison results can be found for older-males with respect to angioedema, as shown in Tables VII and VIII. For example, lisinopril, enalapril (C09AA02), and codeine with paracetamol (N02AA59) are ranked as, respectively, 1, 7, and 92 by the MUTARC and 7, 16, and 1 by the OPUS_AR^+. The runtime of the two algorithms is 1.3 and 1.2 s, respectively. It is interesting to point out that enalapril is shortlisted for the older male stratum, not for the older female stratum, partially because this drug was more carefully prescribed to females (because it should not be used in pregnancy or lactation) [29]. Since only several patients were prescribed perindopril and fosinopril , the MUTARC could not shortlist these two drugs for this relatively small stratum.

In summary, the MUTARC can shortlist six known ADRs such as alendronate $\overset{T}{\hookrightarrow}$ esophagitis, nefazodone $\overset{T}{\hookrightarrow}$ hepatitis, lisinopril $\overset{T}{\hookrightarrow}$ angioedema, perindopril $\overset{T}{\hookrightarrow}$ angioedema, fosinopril $\overset{T}{\hookrightarrow}$ angioedema, and enalapril $\overset{T}{\hookrightarrow}$ angioedema within the ten most interesting UTARs. It can also highlight another ADR, flucloxacillin $\overset{T}{\hookrightarrow}$ hepatitis, which is unknown to our algorithm designers and experiment runners. In addition, it performs much more effectively than the OPUS_AR^+, which is implemented and parameterized same as the MUTARC except without the case-based exclusion operation.

TABLE VIII
TOP DRUGS ASSOCIATED WITH Angioedema GENERATED BY THE OPUS_AR$^+$ FOR THE OLDER MALE PATIENTS

Rank in		Drug in	Drug	Leve	Resi	Risk	$supp(A \xrightarrow{T})$	$supp(A \xrightarrow{T} C)$
Leverage	Resilev	ATC code	name	-rage	-lev	ratio	$\times N(=836)$	$\times N(=836)$
1	92	N02AA59	Codeine with Paracetamol	8.15E-3	1.85E-4	3.826	65	10
2	137	C03CA01	Furosemide	7.76E-3	-1.33E-3	2.965	92	11
3	6	J01DA01	Cefalexin	6.72E-3	3.31E-3	3.126	69	9
4	138	D07AC01	Betamethasone	6.60E-3	-1.35E-3	3.03	71	9
5	17	N05CD07	Temazepam	6.17E-3	1.62E-3	3.252	58	8
6	128	C01DA14	Isosorbide mononitrate	6.11E-3	-7.06E-4	3.193	59	8
7	**1**	**C09AA03**	**Lisinopril**	**5.71E-3**	**4.57E-3**	**5.561**	**25**	**6**
8	152	A02BA02	Ranitidine	5.49E-3	-2.48E-3	2.331	90	9
9	74	C08CA01	Amlodipine	5.26E-3	7.13E-4	3.042	53	7
10	2	H02AB06	Prednisolone	5.16E-3	4.02E-3	8.155	14	5
...	...	...	...	...	...	...	...	...
16	**7**	**C09AA02**	**Enalapril**	**4.09E-3**	**2.95E-3**	**2.152**	**73**	**7**

TABLE IX
RANKS OF THE SUSPECTED DRUGS GENERATED BY THE MUTARC AND OPUS_AR$^+$ FOR 18 DIFFERENT PARAMETER SETTINGS

Strata	Control period setting	Hazard period (in days)	Study period (in days)	Rank of association between alendronate and esophagitis		nefazodone and hepatitis		flucloxacillin and hepatitis	
				Resilev	Leverage	Resilev	Leverage	Resilev	Leverage
age-gender stratum: older-female for esophagitis; middle -aged -female for hepatitis	matched control period	180	[910, 1645]	7	30	1	16	4	12
			[730, 1645]	4	31	3	25	4	20
			[550, 1645]	8	35	6	38	9	23
		120	[730, 1645]	7	32	7	30	3	19
		90	[730, 1645]	5	30	6	33	7	22
		60	[730, 1645]	5	27	6	33	16	34
	random control period	180	[910, 1645]	8	41	3	22	8	17
			[730, 1645]	7	37	7	36	10	23
			[550, 1645]	9	30	8	45	4	18
All strata	matched control period	180	[910, 1645]	18	74	10	38	12	23
			[730, 1645]	16	69	14	48	16	23
			[550, 1645]	19	76	17	60	19	23
		120	[730, 1645]	17	68	17	45	15	23
		90	[730, 1645]	17	69	18	51	16	25
		60	[730, 1645]	16	62	19	53	23	38
	random control period	180	[910, 1645]	18	96	12	36	15	16
			[730, 1645]	20	106	18	47	14	21
			[550, 1645]	19	119	20	62	9	22

C. Reliability Examination

It is important to study the influence of the different parameter settings of the MUTARC on the performance of the ADR signal generation [9]. This shows the reliability of the algorithm. For simplicity, we take the three ADRs, alendronate $\xrightarrow{T}$ esophagitis, nefazodone $\xrightarrow{T}$ hepatitis, and flucloxacillin $\xrightarrow{T}$ hepatitis, as examples to illustrate its reliable performance.

Besides using the specific age-gender strata, as in Section IV-B, we also examined our algorithm on all the strata for a given diagnosis, as shown in the first column of Table IX. We used two different approaches to choose the noncase subsequences:

1) *Using matched control period:* for each case, a prespecified number (e.g., 6 for esophagitis patients and 21 for hepatitis patients) of different noncases are selected as those who have a hospitalization event in the month of the first occurrence of the consequent in the case and are in the same age-gender stratum as the case and

2) *Using random control period:* where a T_c-sized subsequence is chosen randomly from each noncase sequence within $[t_S - T_c, t_E]$. Various hazard period lengths T_h were tested such as 180, 120, 90, and 60 d. We also set different study periods such as from 550 to 1645, 730 to 1645, and 910 to 1645, as shown in the fourth column of Table IX. Note that $T_b = 365$ by default, $t_S = 550$ was the minimal value we might choose in order to have at least half-a-year time period available for the reference period.

For the older female patients, the ranks of alendronate based on residual-leverage with respect to esophagitis range from 4 to 9 for the nine different parameter settings. The average rank is 6.7. The ranks based on leverage range from 27 to 41, which lie at least 21 behind. On average, they lie 25.9 behind. Similarly, for middle-aged-female patients, the ranks of nefazodone $\xrightarrow{T}$ hepatitis generated by the MUTARC range from 1 to 8. The average is 5.2. The ranks of nefazodone $\xrightarrow{T}$ hepatitis generated by the OPUS_AR$^+$ are from 16 to 45, and the MUTARC sorts this ADR 15 to 37 higher than the OPUS$_{}_A R^+$ does. The ranks of flucloxacillin $\xrightarrow{T}$ hepatitis generated by the MUTARC range from 3 to 16. The average is 7.3. The ranks of flucloxacillin $\xrightarrow{T}$ hepatitis generated by the OPUS_AR$^+$ are from 12 to 34, and the MUTARC ranks the ADR 7 to 18 higher than the OPUS_AR$^+$ does.

If all the strata are included, the ranks of alendronate $\overset{T}{\hookrightarrow}$ esophagitis range from 16 to 20, while those of alendronate $\overset{T}{\rightarrow}$ esophagitis vary from 62 to 119. Ranks generated by the MUTARC are 46 to 100 higher than those generated by the OPUS_AR$^+$, and 64.3 higher on average. The ranks of nefazodone $\overset{T}{\hookrightarrow}$ hepatitis are from 10 to 20, and those of nefazodone $\overset{T}{\rightarrow}$ hepatitis are from 36 to 62. The MUTARC ranks this association at least 24 higher than the OPUS_AR$^+$. Similarly, the ranks of flucloxacillin $\overset{T}{\hookrightarrow}$ hepatitis generated by the MUTARC range from 9 to 23, and those of flucloxacillin $\overset{T}{\rightarrow}$ hepatitis generated by the OPUS_AR$^+$ are from 16 to 38. The MUTARC ranks the association 8.3 higher than the OPUS_AR$^+$ on average. In a word, the MUTARC stably ranks the three ADRs within the top ten for their risky age-gender strata of patients, and within the top 20 for all strata, for almost all the different parameter settings. The only exception is for the association between flucloxacillin andhepatitiswhen $T_h = 60$. It is worth noting that the MUTARC reliably shortlisted alendronate $\overset{T}{\hookrightarrow}$ esophagitis based on the data up to June 1999, while this ADR has not been announced until August 1999 in Australia [28]. That illustrates that our techniques can be used to signal ADRs promptly.

V. Related Work

Our work is closely related to the problem of mining TARs where a consequent and an antecedent occur together frequently within a data subset specified by temporal constraints. Along this direction, Li *et al.* [3] studied TARs during time intervals specified by a user-given calendar schema. Lee *et al.* [2] explored the problem of mining TARs in publication databases where time intervals rather than timestamps were used. Harms and Deogun [31] also presented an efficient method for finding frequent TARs in one or more sequences that precede the occurrence of patterns in other sequences. Different from TARs, as a new knowledge representation, our proposed UTAR $A \overset{T}{\hookrightarrow} C$ indicates the antecedent A *unexpectedly* occurs within a T-sized period prior to the consequent C.

Our pairwise UTARs can also be viewed as sequential patterns [4] (or episodes in [5]), where a collection of events occur relatively close together in a given partial order. There were several efficient algorithms on searching for sequential patterns that were more frequently than a threshold [4], [6], or were further constrained [5]. For example, Sun *et al*. [23] discussed the discovery, from a single sequence, of negative event-oriented associations in which the antecedent patterns happen frequently all the time but before the consequent.

Clearly, all these techniques concentrate on finding frequent sequential patterns/itemsets. They are not suitable for identifying infrequent *unanticipated episodes* like ADRs, which is the main aim of this work. The existing techniques consider different aspects of temporal data mining; we consider our research as complementary to them.

The problem of discovering user's unexpected rules, explored by Wang *et al.* [25], is closely related to our work. It concentrated on how to embed the user knowledge in the association rule mining procedure, and established promising techniques to find more unexpected associations of user interest [25]. In many areas, such as medication, there is too much domain knowledge to be considered. Furthermore, the user knowledge is not always easily available, as pointed out in [25]. In contrast, our solution is to better use the data and automatically degrade the uninteresting associations during the mining procedure.

Keogh *et al.* proposed efficient techniques, from a time series, for finding unusual subsequences that are maximally different to all the rest of time-series subsequences [32]. Differing from sequential patterns and UTARs, these unusual subsequences only contain contiguous events.

In the medical domain, current postmarket ADR signaling techniques like proportional reporting ratios (PRRs) [9], multi-item Gamma Poisson Shrinker (MGPS) [14], and Bayesian confidence propagation neural network (BCPNN) [33] perform fruitfully on spontaneous ADR case reports [9]. Each ADR case report describes the suspected causality between drugs and conditions for one patient. Thus, different from the MUTARC, these techniques are not suitable for healthcare administrative data that only routinely record drugs and conditions for people accessing a healthcare system. Other ADR monitoring systems, as reviewed in [12], are to identify adverse events by searching for given ADRs like drug-possibly-causing-symptom patterns. From healthcare administrative databases, risk patterns [13] may find patient groups at high risk of a given adverse reaction. The ADR signals shortlisted by our proposed techniques can help these systems move toward automation.

VI. Conclusion and Discussion

Mining *unanticipated episodes* is of great application value. For example, ADR signals generated can be used, after validation, to prevent lots of unnecessary conditions or hospitalization worldwide. In order to discover *unanticipated episodes*, in this paper, we have introduced a knowledge representation, UTARs, and an interestingness measure, residual-leverage. Based on the novel case-based exclusion and event-oriented data preparation techniques, we have developed an effective mining algorithm, the MUTARC, to discover infrequent pairwise UTARs. The MUTARC has been applied in signaling ADRs from healthcare administrative databases. It has reliably shortlisted with various parameter settings, not only the known ADRs, but also an unknown ADR to algorithm designers and experiment runners. It has empirically performed much more effectively than the OPUS_AR$^+$ whose only difference from the MUTARC is without the case-based exclusion operation. These experimental results have illustrated a new promising direction of ADR signal generation based only on linked healthcare administrative databases. This has also justified the usefulness of our proposed techniques.

We have only concentrated on highlighting pairwise UTARs in this paper. However, the proposed concept and its interestingness measure are readily extended to detect more sophisticated UTARs. Another possible extension is to consider some

quantitative information such as dosage of drug usage during exclusion in order to discover *unanticipated episodes* such as conditions induced by cumulative drug toxicity. One general way for mining unexpected temporal associations is to use a data mining algorithm to find some expected temporal associations from data directly, and then, using this mined knowledge to help us discover unexpected ones. These research directions are the subject of our future work.

ACKNOWLEDGMENT

The authors would like to thank the Australian Government Department of Health and Ageing (DoHA) and the Queensland Department of Health for providing data for this research. They also thank the editors and the anonymous reviewers: R. Hill, I. Boyd, K. Mackay, P. Purcell, E. Hoole, J. McEwen, J. Roediger, C. Winfield, and J. Corbett for their constructive comments and suggestions; other experts from DoHA for reviewing and comments for the work; and Dr. R. Sparks, Dr. L. Gu, and D. McAullay from the Commonwealth Scientific and Industrial Research Organization for constructive discussion and data preparation.

REFERENCES

[1] P. E. Langton, G. J. Hankey, and J. W. Eikelboom, "Cardiovascular safety of rofecoxib (Vioxx): Lessons learned and unanswered questions," *Med. J. Aust.*, vol. 181, no. 10, pp. 524–525, 2004.

[2] C.-H. Lee, M.-S. Chen, and C.-R. Lin, "Progressive partition miner: An efficient algorithm for mining general temporal association rules," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 1004–1017, Aug. 2003.

[3] Y. Li, P. Ning, X. S. Wang, and S. Jajodia, "Discovering calendar-based temporal association rules," *Data Knowl. Eng.*, vol. 44, no. 2, pp. 193–218, 2003.

[4] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. ICDE 1995*, pp. 3–14.

[5] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of frequent episodes in event sequences," *Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 259–289, 1997.

[6] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu, "FreeSpan: Frequent pattern-projected sequential pattern mining," in *Proc. KDD 2000*, pp. 355–359.

[7] The ICH Expert Working Group. (Nov. 2003). Post-approval safety data management: Definitions and standards for expedited reporting. ICH Harmonised Tripartite Guideline [Online]. Available: http://www.fda.gov/cber/gdlns/ichexrep.htm

[8] M. Stephens, J. Talbot, and P. Routledge, Eds*., Detection of New Adverse Drug Reactions*. London, U.K.: Macmillan, 1998.

[9] E. Roux, F. Thiessard, A. Fourrier, B. Begaud, and P. Tubert-Bitter, "Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance," *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 4, pp. 518–527, Dec. 2005.

[10] J. Lazarou, B. Pomeranz, and P. Corey, "Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies," *J. Amer. Med. Assoc.*, vol. 279, no. 15, pp. 1200–1205, 1998.

[11] E. Roughead, "The nature and extent of drug-related hospitalisations in Australia," *J. Qual. Clin. Pract.*, vol. 19, no. 1, pp. 19–22, Mar. 1999.

[12] D. W. Bates, R. S. Evans, H. Murff, P. D. Stetson, L. Pizziferri, and G. Hripcsak, "Detecting adverse events using information technology," *J. Amer. Med. Inf. Assoc.*, vol. 10, no. 2, pp. 115–128, 2003.

[13] J. Li, A. W.-C. Fu, H. He, J. Chen, H. Jin, D. McAullay, G. Williams, R. Sparks, and C. Kelman, "Mining risk patterns in medical data," in *Proc. KDD 2005*, pp. 770–775.

[14] D. M. Fram, J. S. Almenoff, and W. DuMouchel, "Empirical Bayesian data mining for discovering patterns in post-marketing drug safety," in *Proc. KDD 2003*, pp. 359–368.

[15] The Adverse Drug Reactions Advisory Committee (ADRAC). (2007). Australian adverse drug reaction bulletin, DoHA [Online]. Available: http://www.tga.gov.au/adr/aadrb.htm

[16] H. J. Murff, V. L. Patel, G. Hripcsak, and D. W. Bates, "Detecting adverse events for patient safety research: A review of current methodologies," *J. Biomed. Inf.*, vol. 36, no. 1/2, pp. 131–143, 2003.

[17] K. Lasser, K. E. Lasser, P. D. Allen, S. J. Woolhandler, D. U. Himmelstein, S. M. Wolfe, and D. H. Bor, "Timing of new black box warnings and withdrawals for prescription medications," *J. Amer. Med. Assoc.*, vol. 287, no. 17, pp. 2215–2220, May 2002.

[18] G. Williams, D. Vickers, C. Rainsford, L. Gu, H. He, R. Baxter, and S. Hawkins, "Bias in the Queensland linked data set," CSIRO Math. Inf. Sci., Canberra, Australia, Tech. Rep. TR 02/117, 2002.

[19] G. I. Webb, "Efficient search for association rules," in *Proc. KDD 2000*, pp. 99–107.

[20] C. Zhang and S. Zhang, *Association Rule Mining: Models and Algorithms*. New York: Springer-Verlag, 2002.

[21] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. New York: AAAI/MIT Press, 1991, pp. 229–248.

[22] J. Chen, H. He, G. Williams, and H. Jin, "Temporal sequence associations for rare events," in *Proc. PAKDD 2004*, May, pp. 235–239.

[23] X. Sun, M. E. Orlowska, and X. Li, "Finding negative event-oriented patterns in long temporal sequences," in *Proc. PAKDD 2004*, May, pp. 212–221.

[24] H. Jin, J. Chen, C. Kelman, H. He, D. McAullay, and C. M. O'Keefe, "Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases," in *Proc. PAKDD 2006*, Apr., pp. 867–876.

[25] K. Wang, Y. Jiang, and L. V. Lakshmanan, "Mining unexpected rules by pushing user dynamics," in *Proc. KDD 2003*, pp. 246–255.

[26] P. Wang, S. Schneeweiss, R. Glynn, H. Mogun, and J. Avorn, "Use of the case-crossover design to study prolonged drug exposures and insidious outcomes," *Ann. Epidemiol.*, vol. 14, pp. 296–303, Apr. 2004.

[27] The Drug Utilisation Sub-Committee (DUSC), *Australian Statistics on Medicines, 1999-2000.* DoHA, Canberra, Australia, 2003.

[28] The Adverse Drug Reactions Advisory Committee, "A gut feeling for alendronate," *Aust. Adverse Drug React. Bull.* vol. 18, no. 3, p. 11, Aug. 1999.

[29] MedlinePlus. (2007). [Online]. Available: http://medlineplus.gov/

[30] The Adverse Drug Reactions Advisory Committee, "Drug-induced liver disease," *Aust. Adverse Drug React. Bull.* vol. 15, no. 2, pp. 1–2, May 1996.

[31] S. K. Harms and J. S. Deogun, "Sequential association rule mining with time lags," *J. Intell. Inf. Syst.*, vol. 22, no. 1, pp. 7–22, 2004.

[32] E. Keogh, J. Lin, A. Fu, and H. VanHerle, "Finding unusual medical time-series subsequences: Algorithms and applications," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 3, pp. 429–439, Jul. 2006.

[33] J. Almenoff, J. M. Tonning, A. L. Gould, A. Szarfman, M. Hauben, R. Ouellet-Hellstrom, R. Ball, K. Hornbuckle, L. Walsh, C. Yee, S. T. Sacks, N. Yuen, V. Patadia, M. Blum, M. Johnston, C. Gerrits, H. Seifert, and K. LaCroix, "Perspectives on the use of data mining in pharmacovigilance," *Drug Saf.*, vol. 28, no. 11, pp. 981–1007, 2005.

Huidong (Warren) Jin (S'02–M'03) received the B.Sc. degree in applied mathematics from the Department of Applied Mathematics and the M.Sc. degree in applied mathematics from the Institute of Information and System Sciences, both from Xi'an Jiaotong University, China, in 1995 and 1998, respectively, and the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Shatin, Hong Kong, in 2002.

He was with the Division of Mathematical and Information Sciences, Commonwealth Scientific and Industrial Research Organization, Canberra, Australia, and Lingnan University, Hong Kong. He is currently a Researcher at the Nation Information and Communications Technology Australia (NICTA) Laboratory, Canberra, Australia. He is also an Adjunct Research Fellow at the Australian National University, Canberra. He is the author or coauthor more than 30 papers. His current research interests include data mining, pattern recognition, health informatics, security, and privacy.

Dr. Jin is a member of the Association for Computing Machinery (ACM), the IEEE Computer Society, and the program committees of various international conferences and workshops including the 2006 IEEE International Conference on Data Mining.

Jie Chen (M'04) received the Dr.Eng. degree in underwater acoustic engineering from the Northwestern Polytechnical University, Xi'an, China, in 2000.

He was with the Commonwealth Scientific and Industrial Research Organization (CSIRO), Mathematical and Information Sciences, as a Postdoctoral Research Fellow. From 2000 to 2002, he was also with the National Laboratory on Machine Perception, Peking University, Beijing, China. His current research interests include data mining and signal processing.

Mr. Chen is a member of the Association for Computing Machinery (ACM).

Hongxing He received the Ph.D. degree in theoretical condensed matter physics from the Department of Physics and Astronomy, Michigan State University, MI, in Aug. 1985, and the Master degree in computer science from the University of New South Wales, Australia, in Aug. 2003. He was with the Commonwealth Scientific and Industrial Research Organization (CSIRO), Mathematical and Information Sciences, Canberra, Australia.

His current research interests include data mining and its applications.

Graham J. Williams (M'87) received the B.Math.Sc. (Hons.) degree from the University of Adelaide, Adelaide, Australia, in 1984, and the Ph.D. degree in machine from the Australian National University, Canberra, Australia, in 1991.

He was a Principle Research Scientist in Data Mining at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Division of Mathematical and Information Sciences. He is currently the Director and a Senior Data Miner at the Australian Taxation Office, Canberra, Australia. He is also an Adjunct Professor at the University of Canberra, Canberra, and an Adjunct Associate Professor at the Australian National University. He is the author or coauthor of more than 60 papers, including edited volumes, conference proceedings, and journal papers. His current research interests include data mining and text mining, machine learning, spatial information systems, expert systems, and artificial intelligence.

Dr. Williams is a member of the Association for the Advancement of Artificial Intelligence (AAAI), the Association for Computing Machinery (ACM), and the steering committees of the Pacific Asia Conference on Knowledge Discovery and Data Mining and the Australian Artificial Intelligence Conference. He is also a Co-Chair of the Australasian Data Mining Conferences.

Chris Kelman (M'92) received the M.B.B.S. degree in 1979 from Sydney University, Sydney, N.S.W., Australia, and the Ph.D. degree in 2000 from Australian National University, Canberra, A.C.T., Australia, where he was engaged in assessing the potential for using linked electronic health data in evaluating medical devices and monitoring health care outcomes.

For ten years, he was a Hospital Clinician and in general practice. During this time, while studying electrical engineering at the University of Southern Queensland, he became interested in the use of information technology and expert systems in medicine. He was a Chief Medical Adviser in the Australian Therapeutic Goods Administration. Currently, he is an Adjunct Associate Professor at the University of Western Australia, Perth, W.A., Australia, and the Australian National University. He has been closely involved in the development and application of the use of linked electronic health data and has published in the areas of health service research, health economics, travel health, privacy protection, and new methodologies for event detection and outcome monitoring in linked pharmaceutical data. Recently, he has been involved in assisting in the development of a national population-based pharmacovigilance system and has promoted a number of proposals for the rationalization of the medicines regulation in Australia.

Dr. Kelman is a Fellow of the Australian Faculty of Public Health Medicine.

Christine M. O'Keefe received the B.Sc. (Hons.) and Ph.D. degrees in pure mathematics from the University of Adelaide, Adelaide, Australia, in 1982 and 1988, respectively.

She was a Queen Elizabeth II Fellow and held several lecturing positions in pure mathematics at the Universities of Adelaide and the University of Western Australia, Perth, Australia. She has been with the Health Informatics and Information Security and Privacy, Commonwealth Scientific and Industrial Research Organization (CSIRO) Mathematical and Information Sciences and the CSIRO Information and Communication Technologies (ICT) Center, as the Leader. She was with the University of Ghent, Belgium, and Rome, Italy, as a Visiting Professor. She is currently with the CSIRO Preventative Health National Research Flagship, Canberra, Australia, as the Research and Business Leader of Health Data and Information. She is also an Affiliate Associate Professor at the University of Adelaide. She is the author or coauthor of more than 70 papers published in international journals and refereed conference proceedings. Her current research interests include privacy-enhancing technologies, including privacy-preserving linkage and disclosure risk and data utility associated with remote analysis servers.

Dr. O'Keefe is a Fellow of the Australian Mathematical Society and the Institute of Combinatorics and its Applications. She was the recipient of the Australian Mathematical Society Medal 2000 for distinguished research in mathematical sciences and the Hall Medal of the Institute for Combinatorics and its Applications 1996 for outstanding contributions to the field.