A Potential Causal Association Mining Algorithm for Screening Adverse Drug Reactions in Postmarketing Surveillance

Yanqing Ji, Hao Ying, *Senior Member, IEEE*, Peter Dews, Ayman Mansour, John Tran, Richard E. Miller, and R. Michael Massanari

***Abstract*—Early detection of unknown adverse drug reactions (ADRs) in postmarketing surveillance saves lives and prevents harmful consequences. We propose a novel data mining approach to signaling potential ADRs from electronic health databases. More specifically, we introduce potential causal association rules (PCARs) to represent the potential causal relationship between a drug and ICD-9 (CDC. (2010). *International Classification of Diseases, Ninth Revision (ICD-9)*. [Online]. Available: http://www.cdc.gov/nchs/icd/icd9.html) coded signs or symptoms representing potential ADRs. Due to the infrequent nature of ADRs, the existing frequency-based data mining methods cannot effectively discover PCARs. We introduce a new interestingness measure, *potential causal leverage*, to quantify the degree of association of a PCAR. This measure is based on the computational, experience-based fuzzy recognition-primed decision (RPD) model that we developed previously (Y. Ji, R. M. Massanari, J. Ager, J. Yen, R. E. Miller, and H. Ying, "A fuzzy logic-based computational recognition-primed decision model," *Inf. Sci.,* vol. 177, pp. 4338–4353, 2007) on the basis of the well-known, psychology-originated qualitative RPD model (G. A. Klein, "A recognition-primed decision making model of rapid decision making," in *Decision Making in Action: Models and Methods,* 1993, pp. 138–147). The *potential causal leverage* assesses the strength of the association of a drug–symptom pair given a collection of patient cases. To test our data mining approach, we retrieved electronic medical data for 16 206 patients treated by one or more than eight drugs of our interest at the Veterans Affairs Medical Center in Detroit between 2007 and 2009. We selected enalapril as the target drug for this ADR signal generation study. We used our algorithm to preliminarily evaluate the associations between enalapril and all the ICD-9 codes associated with it. The experimental results indicate that our approach has a potential to better signal potential ADRs than risk ratio and leverage, two traditional frequency-based measures. Among the top 50 signal pairs (i.e., enalapril versus symptoms) ranked by the *potential causal-leverage* measure, the physicians on the project determined that eight of them probably represent true causal associations.**

***Index Terms*—Adverse drug reactions (ADRs), data mining, fuzzy logic, postmarketing surveillance, potential causal association rules (PCARs), recognition-primed decision model (RPD).**

I. Introduction

ADVERSE drug reactions (ADRs) represent a serious problem worldwide. They refer to drug-associated adverse incidents in which drugs are used at an appropriate dose and indication. They can complicate a patient's medical condition or contribute to increased morbidity, even death. Studies have shown that ADRs contribute to about 5% of all hospital admissions and represent the fifth most common cause of death in hospitals [4]. In year 2000, for example, about 7000 deaths in the U.S. were attributed to ADRs [5].

Even though premarketing clinical trials are required for all new drugs before they are approved for marketing, these trials are necessarily limited in size and duration, and thus are not capable of detecting rare ADRs. In general, if the occurrence rate of a potential ADR is less than 0.1%, it cannot be recognized by the premarketing randomized controlled trials due to their limitation in size [6]. Hence, significant, life-threatening, unsuspected ADRs that occur infrequently may never be observed before the drug is introduced into the market where thousands of patients are exposed to the drug.

Drug safety depends heavily on postmarketing surveillance. In the U.S., current postmarketing methods primarily rely on FDA's (Food and Drug Administration) spontaneous reporting system MedWatch. The limitations of this system are well described [7]. MedWatch is a passive system that depends on voluntary, spontaneous reports of suspected ADRs filed by healthcare professionals, drug manufactures, and/or consumers. Detection of an ADR generally relies on FDA's retrospective or concurrent review of patient cases. Because ADR reports are filed at the discretion of the users of the system, it was estimated that less than 10% of all ADR cases were reported to MedWatch [8]. Moreover, the current surveillance system is limited by latency and inconsistency [9]. Consequently, it may require years to identify and withdraw problematic drugs from the market, and results in unnecessary mortality, morbidity, and cost.

Manuscript received May 7, 2010; revised December 25, 2010; accepted March 9, 2011. Date of publication March 24, 2011; date of current version May 4, 2011. This work was supported in part by National Institutes of Health under Grant 1 R21 GM082821-01A1.

Y. Ji is with the Department of Electrical and Computer Engineering, Gonzaga University, Spokane, WA 99258 USA (e-mail: ji@gonzaga.edu).

H. Ying and A. Mansour are with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202 USA (e-mail: hao.ying@wayne.edu; el_akaber@yahoo.com).

P. Dews is with the Department of Medicine, St. Mary Mercy Hospital (Trinity Health), Livonia, MI 48154 USA (e-mail: pdews@attglobal.net).

J. Tran is with the Spokane Mental Health, Spokane, WA 99202 USA (e-mail: jtran@smhca.org).

R. E. Miller was with the John D. Dingell Veterans Affairs Medical Center, Detroit, MI 48201 USA. He is now with Wayne State University, Detroit, MI 48202 (e-mail: remiller@med.wayne.edu).

R. M. Massanari is with Critical Junctures Institute, Bellingham, WA 98225 USA (e-mail: Michael.Massanari@wwu.edu).

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.2011.2131669

Data mining methods for the detection of suspected safety problems from spontaneous reports have been studied and practically implemented [10]. For example, the FDA currently adopts a data mining algorithm called multiitem gamma Poissson shrinker [11] for detecting potential signals from its spontaneous reports. Another important signal-detection strategy is known as the Bayesian confidence propagation neural network that has been used by the Uppsala Monitoring Center in routine pharmacovigilance with its World Health Organization database [12]. However, a large number of cases are needed in order for the neural network approach to work. Various other methods such as proportional reporting ratios [13], empirical Bayes screening [14], reporting odds ratios [15], and incidence rate ratios [16] have been used in the spontaneous reporting centers of other nations (e.g., England and Australia). By utilizing data mining techniques, these methods have shown better performance than traditional methods even though the ability of detecting a signal varies among them [17]. However, the performance of these techniques could be highly situation dependent due to the weaknesses and potential biases inherent in spontaneous reporting [18]. In addition, early generation of a new signal can be very difficult because a large number of interesting cases cannot be timely collected due to the underreporting nature of the current reporting system.

As electronic patient data become more easily accessible in various healthcare organizations, they may provide a new source of information from which ADR signals could be generated much earlier [19]. In this paper, we mine potential ADR signals using an extensive data source that contains administrative data, pharmacy, and clinical laboratory data. We propose a new data mining method where the association between a drug and a symptom for each patient case is a value between [0, 1], instead of a binary number used by other mining methods in the literature [20]. Rather than simply mining the temporal association between drug–symptom pairs, we attempted to mine the more difficult potential causal association rules (PCARs). The word "potential" is necessary because whether or not a mined association portrays a real causal relationship is uncertain. Confirming a causal relationship between a drug and an ADR is challenging because multiple unrelated causes may result in similar outcomes. We defined a new interestingness measure called *potential causal leverage* where the contribution of each case to the measure depends on the degree of potential causal association between the drug and the symptom within that case.

In our approach, the degree of causality for a drug–symptom pair within each case is evaluated by a computational fuzzy recognition-primed decision (RPD) model [2] that we developed earlier on the basis of psychology-originated qualitative RPD model [3]. In this model, experiences played a key role. Four ADR detection experiences were acquired through the joint efforts of our engineering and medical team members after careful analysis of the relevant literature. Each experience corresponded to one of the four causality categories (i.e., very likely, probable, possible, and unlikely) and was characterized by the values of four cues: temporal association, dechallenge, rechallenge, and other explanation. The critical temporal association assumed that cause preceded effect, or that drug treatment preceded occurrence of a presumed adverse response. Dechallenge was defined as the relationship between withdrawal of the drug and abatement of the adverse effect. Rechallenge described the relationship between reintroduction of the drug followed by recurrence of the adverse event. Other explanations denoted alternative explanations by concurrent disease or other drugs. The degree of potential causal association for a drug–symptom pair within a particular case was obtained by matching the cue values extracted from the case with the cue values in each experience. For more details of the fuzzy RPD model as well as concrete examples, the reader is referred to our previous work [2].

The effectiveness of our data mining strategy was evaluated using electronic patient data retrieved from the Veterans Affairs Medical Center in Detroit. Our preliminary results are presented in Section IV followed by Section V. Note that the scope of this paper is to provide an automated, efficient method that expedites prioritization of all possible signal pairs that require extensive, labor-intensive analytical exercises. The ADR signals generated using our new data mining algorithm will require more thorough epidemiological analyses in order to sort through possible causative factors, including drugs, and to quantify the boundaries of uncertainty around suspect drugs.

II. Problem Formulation and a Novel Potential Causal-Leverage Measure

An association rule is an implication expression in the form of $X \rightarrow Y$, where X and Y are two event sets and they are disjoint (i.e., $X \cap Y = \phi$), meaning that they share no common events [21]. Both X and Y may contain one or more events. An association rule indicates that the presence of X implies the presence of Y. If X and Y have temporal relationship, a temporal constraint is often applied to the association rule. Such an association rule, represented as $X \stackrel{T}{\rightarrow} Y$, is called the temporal association rule, where $\stackrel{T}{\rightarrow}$ denotes that Y occurs after X within a time window T in the same event sequence.

In the literature, association rule mining is usually based on two measures: support and confidence. The support of an association rule supp($X \rightarrow Y$) is the proportion of sequences in which both X and Y occur at least once, among all the event sequences. This measure indicates how often a rule is applicable to a given set of event sequences. The importance of this measure lies in the fact that if a rule has very low support, it may occur simply by chance and thus would be uninteresting. The confidence of an association rule is defined as conf($X \rightarrow Y$) = supp($X \rightarrow Y$)/supp($X \rightarrow$), where supp($X \rightarrow$) is the proportion of sequences that contain X. This measure determines how frequently Y appears in those event sequences that contain X. That is, given an association rule $X \rightarrow Y$, the higher the confidence, the more likely it is for Y to be present in sequences that contain X. Given these two measures, the association rule mining problem can be formalized as finding those rules whose support and confidence are greater than prespecified thresholds minsupp and minconf, respectively.

In the context of ADR signal generation using electronic patient data, a health database often contains many patient records,

each of which can be considered as an event sequence where various events such as drug prescription, occurrence of a symptom, and laboratory test occur at different times. However, the previous two frequency-based measures cannot be used to effectively find the association between a drug and an ADR because of the inherently low frequency of ADRs. For example, bromfenac (Duract) was a nonsteroidal antiinflammatory agent that was removed by the FDA from the market in 1998, less than 1 year after it was introduced. Bromfenac caused serious hepatotoxicity in only 1 in 20 000 patients taking the drug for longer than 10 days [22]. This type of rare and delayed drug reactions is very difficult to be detected during the premarketing clinical trials because most new drugs are approved after an average of only 1 500 patient exposures and usually for relatively short exposure periods. Thus, detection of these low-frequency ADRs requires many more exposures and can only be possibly achieved in postmarketing surveillance. If the two aforementioned measures were used, their corresponding thresholds minsupp and minconf would have to be set very small in order to detect these ADRs. This would result in a higher rate of false drug-ADR associations and also make the computational cost of the algorithm extremely high [20], [21].

Another pitfall of using the support and confidence measures is that the detected associations with relatively high frequencies may not be interesting. The reason is that symptoms of a potential ADR cannot be easily distinguished from those of an underlying disease in electronic health databases, since both are coded using the same ICD-9 codes [1]. Thus, if a drug is used to treat a disease, the drug and the disease can be easily detected as an association using these measures. However, this association is expected and not interesting.

Interesting measures like leverage and its variations have been used to mine temporal associations between drugs and symptoms [20], [23]. The leverage of an association rule is defined as

$$\text{leverage}(X \rightarrow Y) = \text{supp}(X \rightarrow Y) - \text{supp}(X \rightarrow) \times \text{supp}(\rightarrow Y) \quad (1)$$

where supp($\rightarrow Y$) is the proportion of sequences that contain *Y*. This measure indicates the proportion of sequences exhibiting the association between *X* and *Y* in excess of those that would be expected if *X* and *Y* were independent of each other. To mine temporal association rules, a temporal constraint can be embedded into the measure. That is,

$$\text{leverage}(X \overset{T}{\rightarrow} Y) = \text{supp}(X \overset{T}{\rightarrow} Y) - \text{supp}(X \overset{T}{\rightarrow}) \times \text{supp}(\overset{T}{\rightarrow} Y) \quad (2)$$

where supp($X \overset{T}{\rightarrow} Y$) represents the proportion of sequences in which *Y* occurs after *X* within a *T*-sized window. supp($X \overset{T}{\rightarrow}$) and supp($\overset{T}{\rightarrow} Y$) denote the proportion of sequences that contain *X* and *Y*, respectively, among all the *T*-constrained sequences. While this measure can express the temporal association between two events, it cannot effectively represent the potential *causal* relationship between a drug and a symptom. Hence, a new measure needs to be introduced.

TABLE I
MEMBERSHIP OF EACH CAUSALITY CATEGORY FOR SAMPLE SIGNAL PAIRS (PID MEANS PATIENT IDENTIFICATION NUMBER)

PID	Pair	Very likely	Probable	Likely	Unlikely
1	<drugA, symptom1>	0.6	0.2	0.15	0.05
1	<drugA, symptom2>	0.8	0.15	0.05	0.0
2	<drugB, symptom2>	0.3	0.4	0.2	0.1
3	None	0	0	0	0
4	<drugA, symptom1>	0.0	0.1	0.2	0.7
5	<drugA, symptom2>	0.2	0.5	0.2	0.1
5	<drugB, symptom1>	0.1	0.3	0.4	0.2
6	None	0	0	0	0
7	<drugA, symptom1>	0.05	0.15	0.3	0.5
8	None	0	0	0	0

Our strategy is to utilize an experience-based model (i.e., the computational fuzzy RPD model) to capture the potential causality between *X* and *Y*. We introduce PCAR, denoted by $X \overset{C}{\rightarrow} Y$, meaning that *X* potentially causes *Y*. The degree of potential causality of an event pair in each event sequence is determined by the fuzzy RPD model in which temporal association is only one of the four parameters that are used to assess the strength of causal association between two events. Note that the RPD model employs "situation-experience matching" decision rules to determine how well the current situation matches various prior experiences. Specifically, for each interesting drug–symptom pair, the values of four cues (i.e., temporal association, other explanation, dechallenge, and rechallenge) are abstracted from the related electronic patient record. These cue values then match the cue values in the four defined experiences, each of which corresponds to a unique category of causality (i.e., very likely, probable, possible, and unlikely). A similarity value between the current pair and each of the experiences is obtained using a similarity measure developed in our previous work [2]. The similarity measure was defined as the weighted sum of all the local similarities between each pair of cue values associated with the same cue in the two sets of cue values. The calculation of a local similarity depends on the type of the associated cue that can be fuzzy, nominal, or quantitative. After that, these similarity values are normalized so that their sum is equal to 1. These normalized values are then used to represent the membership values of corresponding categories of causality between the drug of interest and symptom. For example, if the normalized similarity value between a pair and the experience exhibiting "probable" association is 0.4, we would say that the causality of the drug–symptom pair is "probable" with a membership value of 0.4. Table I gives several sample drug–symptom pairs as well as their membership values related to each causality category. Note that, given a particular drug, there may exist none, one, or multiple drug–symptom pairs within each patient case. Patient 3, 6, and 7 in the table do not contain drugs and/or symptoms of our interest.

Let μ_v, μ_p, μ_o, and μ_u represent the membership values of causality categories "very likely," "probable," "possible," and "unlikely," respectively. Then, the quadruple $\mu = \{\mu_v, \mu_p, \mu_o, \mu_u\}$ can be used to characterize the potential causality of a drug–symptom pair. In order to define a causality-based leverage measure, we combine the membership values of the four causality categories into one number using weighted sum and define the degree of causality of a pair as

$$C_{<\text{drug, symptom}>} = \mu_v \times w_v + \mu_p \times w_p + \mu_o \times w_o + \mu_u \times w_u \tag{3}$$

where w_v, w_p, w_o, and w_u are the weights for μ_v, μ_p, μ_o, and μ_u, respectively. In this study, we set $\{w_v, w_p, w_o, w_u\}$ as $\{1.0, 0.667, 0.333, 0.0\}$ based on two considerations. First, causality categories representing stronger causal associations should have higher weights. That is, $w_v > w_p > w_o > w_u$ must be satisfied. Second, the range of $C_{<\text{drug, symptom}>}$ should be [0, 1]. That is, $C_{<\text{drug, symptom}>}$ should be 0 for the extreme situation $\mu = \{0,0,0,1\}$ where the evidence in a patient record strongly shows "unlikely" association of the pair. If all the evidence in the patient supports "very likely" association (i.e., $\mu = \{1,0,0,0\}$), $C_{<\text{drug, symptom}>}$ should be 1. Otherwise, $C_{<\text{drug, symptom}>}$ is between 0 and 1. Given the aforementioned weights for μ_v, μ_p, μ_o, and μ_u, the causality of pair <drugA, symptom1> in patient 1 can be calculated as follows:

$$C^1_{<\text{drugA, symptom1}>} = 0.6 \times 1.0 + 0.2 \times 0.667 + 0.15 \times 0.333 + 0.05 \times 0.0 = 0.783$$

where the superscription 1 indicates the patient identification number (PID).

In general, if the experience-based fuzzy RPD model classifies the causality between *X* and *Y* into *m* categories, the degree of causality is defined as

$$C_{<X,Y>} = \sum_{i=1}^{m} \mu_i \times w_i \tag{4}$$

where μ_i is the membership of the *i*th causality category for the pair, and w_i represent the corresponding weight when converging all the causality categories into one. Moreover,

$$\sum_{i=1}^{m} \mu_i = 1.$$

We define the support of a PCAR, $\text{supp}(X \xrightarrow{C} Y)$, as the accumulated votes over all sequences. That is, it is calculated by first summing all votes of each sequence divided by the total number of sequences. The vote of a sequence is a function of causality between *X* and *Y*. That is,

$$\text{supp}(X \xrightarrow{C} Y) = \frac{\sum_{j=1}^{n} \alpha^j_{<X,Y>}}{n} \tag{5}$$

where *n* is the total number of sequences, and $\alpha^j_{<X,Y>}$ represents the vote of the *j*th sequence with respect to the pair <*X,Y*>. $\alpha^j_{<X,Y>}$ is defined as

$$\alpha^j_{<X,Y>} = \begin{cases} C^j_{<X,Y>}, & \text{if } C^j_{<X,Y>} > \theta \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $C^j_{<X,Y>}$ is the degree of causality between *X* and *Y* in the *j*th sequence. θ is the user-specified threshold, which ensures that if the causality of pair <*X, Y*> is low in a sequence, this sequence will not be considered when calculating the support of the PCAR. The purpose of the inclusion of parameter θ is to exclude those false associations as much as possible based on the assumption that they normally obtain lower values for $C^j_{<X,Y>}$. The selection of θ for the ADR problem will rely on either domain experts (e.g., physicians and drug safety officers) or experiments. In the latter case, our interestingness measure will be used to detect one or more known ADR signals when θ takes various values. The appropriate value for θ will be the one that ranks high for the known drug–ADR pairs among all the pairs in a database. Note that if a sequence does not contain both *X* and *Y*, the causality between *X* and *Y* is 0. One can see that the contribution of each sequence to the support measure becomes gradual as opposed to binary. That is, the higher the degree of association between *X* and *Y* in a sequence, the more the sequence contributes to the measure. This definition of the support measure is more reasonable in the ADR detection and likely many other medical applications than its traditional definition where the contribution of sequence to the measure is either 0 or 1.

As an example, we now compute the support of the PCAR drugA $\xrightarrow{C}$ symptom1 using the data in Table I. We assume that θ is equal to 0.2. The degree of causality of pair <drugA, symptom1> in patient 1 was already calculated earlier. Since $0.783 > 0.2$ in patient 1, ${\alpha^1}_{<\text{drugA,symptom1}>} = 0.783$. The causality of pair <drugA, symptom1> in patient 4 is computed as follows:

$$C^4_{<\text{drugA, symptom1}>} = 0.0 \times 1.0 + 0.1 \times 0.667 + 0.2 \times 0.333 + 0.7 \times 0.0 = 0.133.$$

Since the value is less than $\theta = 0.2$, ${\alpha^4}_{<\text{drugA,symptom1}>} = 0.0$. Similarly, we can get ${C^7}_{<\text{drugA,symptom1}>} = 0.467$, which is larger than θ. Thus, ${\alpha^4}_{<\text{drugA,symptom1}>} = 0.467$. None of the other patient cases contains the pair <drugA, symptom1>. Their votes are zeros. Therefore, the support of the PCAR drugA $\xrightarrow{C}$ symptom1 is

$$\text{supp}(\text{drugA} \xrightarrow{C} \text{symptoml}) = \frac{0.783 + 0.467}{8} = 0.156.$$

Next, we introduce a new interestingness measure, potential causal leverage, based on the aforementioned definition of the support of a PCAR. The potential causal leverage of a PCAR is given as

$$\text{potential causal leverage}(X \xrightarrow{C} Y) = \text{supp}(X \xrightarrow{C} Y) - \text{supp}(X \xrightarrow{C}) \times \text{supp}(\rightarrow Y) \tag{7}$$

where supp($X \xrightarrow{C}$) is the proportion of sequences whose votes are not equal to 0 with respect to the pair <*X,Y*>. Based on the earlier discussions, only the votes in patients 1 and 7 are greater than 0 when computing supp(drugA$\xrightarrow{C}$ symptom1). Thus, supp($\text{drugA} \xrightarrow{C}$) = 2/8 = 0.25. Given the data shown in Table I, supp(→symptom1) = 4/8 = 0.5 since four cases contain symptom1. Hence,

$$\text{potential causal leverage}(\text{drugA} \xrightarrow{C} \text{symptom1}) = 0.156 - 0.25 \times 0.5 = 0.031.$$

As a comparison, we also calculated the potential causal leverage of another PCAR—drugA$\xrightarrow{C}$ symptom2:

$$\text{potential causal leverage}(\text{drugA} \xrightarrow{C} \text{symptom2}) = 0.19 - 0.25 \times 0.375 = 0.096.$$

Therefore, drugA has a stronger association with symptom2 than with symptom1 based on our potential causal-leverage measure, even though pair <drugA, symptom1> has a higher frequency than pair <drugA, symptom2>. Note that, in Table I, there are three cases containing <drugA, symptom1>, but only two cases containing <drugA, symptom2>. The supports for traditional association rules drugA→ symptom1 and drugA → symptom2 are 3/8 and 2/8, respectively. Their confidences are 3/5 and 2/5, respectively. That is, drugA has a stronger association with symptom1 than with symptom2, according to the traditional support and confidence measures. We believe that our result is more reasonable because, even though fewer cases contain the pair <drugA, symptom2>, they exhibit higher accumulated degree of causality. Moreover, common high-frequent illnesses (e.g., fever) often coexist with a drug by chance in the databases, but they can be mistakenly classified as ADR signals if the traditional frequency-based measures are used. As mentioned earlier, premarketing clinical trials fail to detect rare ADRs which have to rely on postmarketing surveillance after more patients expose to the drug. Our measure can deal with the infrequent nature of ADRs better than the traditional association rule mining measures.

To see the difference between our potential causal-leverage measure and the traditional leverage measure, we also calculated the latter for the two association rules drugA→ symptom1 and drugA → symptom2 as follows:

$$\begin{aligned}\text{leverage}(\text{drugA} \rightarrow \text{symptom1}) &= \text{supp}(\text{drugA} \rightarrow \text{symptom1}) \\ &\quad - \text{supp}(\text{drugA} \rightarrow) \times \text{supp}(\rightarrow \text{symptom1}) \\ &= \frac{3}{8} - \frac{5}{8} \times \frac{4}{8} = 0.0625 \\ \text{leverage}(\text{drugA} \rightarrow \text{symptom2}) &= \text{supp}(\text{drugA} \rightarrow \text{symptom2}) \\ &\quad - \text{supp}(\text{drugA} \rightarrow) \times \text{supp}(\rightarrow \text{symptom2}) \\ &= \frac{2}{8} - \frac{5}{8} \times \frac{3}{8} = 0.0156.\end{aligned}$$

One can see that the traditional leverage measure comes to the same conclusion (i.e., drugA has a stronger association with symptom1 than with symptom2) as the traditional support and confidence measures. The reason is that all these measures have the same drawback: the *degree* of association of the pair in each case is not considered. That is, each patient case contributes equally to the measures as long as that case contains the drug–symptom pair of interest. Therefore, the traditional leverage measure cannot effectively handle the infrequent nature of ADRs either.

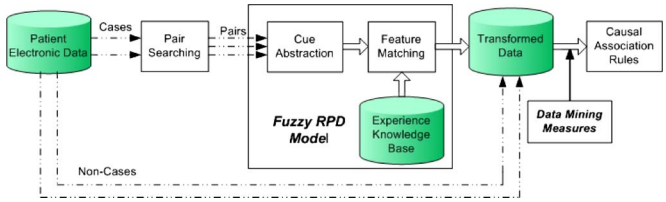


Fig. 1. Algorithm for mining causal association rules (the term *cases* refers to those patient records containing the drug of interest and *noncases* refers to other patient records).

DRUG NAME	START DATE
ENALAPRIL MALEATE 5MG TAB	5/19/2006
LOVASTATIN 10MG TAB	5/23/2007
ENALAPRIL MALEATE 5MG TAB	10/2/2008
SIMVASTATIN 20MG TAB	10/28/2008

ICD-9 CODE	DIAGNOSIS DATE
729.1	8/23/2004
786.2	6/12/2006
276.7	10/13/2008
276.7	9/5/2009

Fig. 2. Sample signal pairs within a patient case.

III. Mining PCARs

In this section, we develop a data mining algorithm to search for PCARs based on our *potential causal-leverage* measure. Specifically, we will explore how to mine a clinically oriented electronic database in order to find out the strengths of associations between a drug of interest and various symptoms. We assume that patient data are stored in various tables in a relational database. In the following descriptions, we use the term *cases* to refer to those records containing the drug of interest and *noncases* to refer to other records.

Fig. 1 shows an overall picture of the algorithm. First, each record is classified as either case or noncase, depending on whether it contains the drug of interest. Next, each case is searched for all possible drug–symptom pairs based on drug start dates and diagnosis dates. As discussed earlier, in a typical electronic health database, one cannot easily differentiate the underlying disease from a potential ADR, both of which are encoded in the ICD-9 codes and have diagnosis dates in the database. For example, cough (ICD-9 code 786.2) could represent either an ADR caused by a drug-like enalapril or simply a cough related to some other conditions. Moreover, since each patient case records the patient's history for a long period of time, the same drug may be prescribed many times and thus multiple drug start dates may exist for it. Similarly, there may exist multiple diagnosis dates for the same condition or ADR. Therefore, we must compare each start date of the drug of interest with each diagnose date of a condition/symptom. A drug–symptom pair is recognized if the symptom occurs at least once after one of the start dates of the drug within a certain period of time (e.g., 90 days). If we assume enalapril is our drug of interest, Fig. 2 shows that two pairs are found within one

case. In this case, since the ICD-9 code 729.1 (myalgia) occurs before all the start dates of enalapril, it does not form a pair with the drug. The ICD-9 code 786.2 (cough) occurs after the first start date of enalapril in 24 days. Thus, a pair is found, i.e., enalapril—786.2. The code 276.7 (hyperkalemia) occurs twice after the second start date of enalapril. A pair is formed between the drug and the closest occurrence of 276.7.

After a pair is found, its degree of association is assessed using the fuzzy RPD model. Four cues (i.e., temporal association, other explanation, dechallenge, and rechallenge) are utilized in this process. Their values are extracted from patient cases using fuzzy rules. For example, the cue *temporal association* is a fuzzy variable whose value is obtained using fuzzy rules like "if the time duration between taking a drug and occurrence of a symptom is *short*, then temporal association is *likely*." Readers are referred to our previous paper [2] for all the fuzzy rules used in this study.

After the cue values of a pair are extracted from a case, we use the feature-matching procedure introduced in Section II to transform the data of a pair to the format shown in Table I. We then compute the support and *potential causal leverage* for all the pairs, each of which implies a possible PCAR. To further increase the accuracy of our data mining algorithm, we also use a frequency threshold δ to exclude those pairs with very high frequencies since they may generate false signal pairs. They are primarily those pairs between the drug of interest and the symptoms caused by the underlying diseases treated by the drug. Since ADRs are infrequent and usually have much lower frequencies than those pairs, the exclusion of the pairs will unlikely eliminate any true signals. Finally, instead of setting an explicit threshold for the *potential causal-leverage* measure, we rank all the PCARs in the decreasing order of the measure and ask the program to return the most interesting rules.

IV. Experiment

A. Experiment Setting

To evaluate the effectiveness of our new data mining strategy, we retrieved the electronic data of all the patients who received at least one of the eight drugs of our interest in the Veterans Affairs Medical Center in Detroit during the time period from September 4, 2007 to September 16, 2009. There were six cholesterol or "statin" drugs (i.e., rosuvastatin, atorvastatin, fluvastatin, lovastatin, pravastatin, and simvastatin) and two angiotensin converting enzyme inhibitor or "ACEI" drugs (i.e., captopril, and enalapril). Event data such as dispensing of drug, office visits, and certain laboratory tests were retrieved for all the patients. For each event, certain details were obtained. For example, the data for dispensing of drug include the name of the drug, quantity of the drug dispensed, dose of the drug, drug start date, drug schedule, and the number of refills.

The total number of patients retrieved was 16 206 (15 605 males and 601 females). Their average age was 68.0. To test our data mining algorithm, we selected a drug whose associated potential ADRs have a relative high occurrence frequency since ADRs generally occur infrequently and the size of our database is moderate. In general, if the occurrence rate of a potential ADR is greater than 1%, it can be easily detected in a premarketing clinical trial. As we mentioned in Section I, if the occurrence rate of a potential ADR is less than 0.1%, it normally cannot be detected. If the rate is between 0.1% and 1%, the ADR may or may not be detected depending on the size of the premarketing clinical trial. Thus, the occurrence rate of a potential ADR is generally less than 1% in postmarketing surveillance. Moreover, to generate a valid signal, the interestingness measure is often used along with another criteria—the count of drug–ADR pairs being greater than or equal to 3 [13]. That is, even if the calculated interesting measure for a pair is high (e.g., higher than a predefined threshold), the pair cannot be considered as a valid signal if there are less than three patient cases that are evaluated as containing the potential ADR. This requires at least 300 ($3/0.01 = 300$) cases in order to generate a valid signal even if the occurrence rate is 1%. Owing to these considerations, we picked the drug enalapril to test the proposed algorithm since 1021 patients were found taking the drug at least once in the database and the number of its known potential ADRs had relative high occurrence rates [24], [25]. One thousand two hundred and ninety distinctive ICD-9 codes were found to be associated with this drug for these patients.

All the data were deidentified and stored in a Microsoft Access database. Java database connectivity (JDBC) [26], an application programming interface for the Java programming language, was used to access the database. JDBC could wrap a structured query language (SQL) statement [27], send it to the database, and retrieve the desired data. Fuzzy rules and fuzzy reasoning were implemented using the freeware FuzzyJess [28], a Java-based fuzzy inference engine. It allowed the user to use the Java language to define membership functions, set antecedent and consequent of a fuzzy rule, and made a fuzzy inference.

B. Experimental Results

It might be necessary to briefly describe the nature of the ground truth for ADRs before we present our experimental results. For some potential ADRs, there only exist a limited number of reports about their associations with particular drugs in the literature due to their infrequent nature and the difficulties establishing causal relationships. In many other cases, even if the statistical information is available, it is still difficult to say whether a drug is associated with a symptom. For example, the premarketing clinical studies for enalapril indicate that the incidence of asthenia for patients who took enalapril and placebo was 1.1 and 0.9, respectively [29]. In this example, we are not certain whether asthenia is an ADR of enalapril since the number 1.1 is only slightly larger than 0.9. Because of these reasons, clinicians normally use linguistic terms such as "possible," "probably," and "very likely" to describe the association between a drug and a potential ADR.

The *potential causal-leverage* values were calculated for each of the pairs between enalapril and all the 1290 distinctive ICD-9 codes involved with this drug. All the pairs were ranked in decreasing order according to their *potential causal-leverage*

TABLE II
ICD-9 CODES FOUND TO INDICATE POSSIBLE ADRS ASSOCIATED WITH DRUG ENALAPRIL AMONG THE TOP 50 SIGNAL PAIRS RANKED ACCORDING TO THE POTENTIAL CAUSAL-LEVERAGE VALUES

Rank in top 50 signal pairs	ICD-9 code	ICD-9 code description	Potential causal-leverage value
6	311.	Depressive disorder NEC	2.05×10^{-4}
10	786.05	Shortness of breath	1.79×10^{-4}
15	586.	Renal failure NOS	1.65×10^{-4}
16	300.4	Neurotic depression	1.62×10^{-4}
20	276.7	Hyperpotassemia	1.53×10^{-4}
25	388.31	Subjective tinnitus	1.42×10^{-4}
29	607.84	Impotence-organic origin	1.28×10^{-4}
42	780.4	Dizziness and giddiness	1.07×10^{-4}

values. The top 50 pairs were evaluated by the physicians on our project team, and eight of them were found to be possible ADRs. Table II shows the ICD-9 codes of these eight pairs as well as their descriptions, and the corresponding *potential causal-leverage* values and ranks. These ICD-9 codes represent "possible" ADRs that might be caused by enalapril except "276.7" (hyperpotassemia) which is considered as a "very likely" ADR by our physicians. Physicians often are uncertain about the causal relationship between a drug and a symptom due to the fact that the same syndrome or ADR may be caused by various other conditions. More details regarding this issue are provided in Section V.

We also examined the ICD-9 codes that occurred more frequently than the threshold δ in the same patient cases (i.e., those pairs excluded by the frequency threshold). We set δ to 0.01 since the frequencies of ADRs are generally lower than 1%. To exclude those high-frequency pairs, we simply assign -1 to the *potential causal leverage* of the pairs to flag them. Note that the normal value of the *potential causal leverage* of a pair is between -1 and 1. Twenty-two out of 1290 ICD-9 codes were found to be excluded. Our physicians determined that none of them were potential ADRs. This indicated that the frequency threshold δ was at an appropriate level and did not eliminate true signal pairs.

Since different ICD-9 codes may represent the same (or similar) diagnoses, we also clustered the data mining results into a manageable number of categories based on the clinical classifications system (CCS) for the ICD-9-CM fact sheet [30]. Developed at the Agency for Healthcare Research and Quality, the CCS can group over 13600 ICD-9 codes into 285 mutually exclusive and clinically meaningful categories. We accumulated the *potential causal-leverage* values of the ICD-9 codes that belonged to the same CCS category and then ranked the resulting categories according to the accumulated leverage values. The 1290 ICD-9 codes coexisting with enalapril in our database were grouped into 127 CCS categories. Among the top 50 categories, 14 of them were found to be possible ADRs associated with enalapril. These categories as well as their corresponding accumulated *potential causal-leverage* values and rankings are shown in Table III. Again, these categories only represent "possible" ADRs. Besides the reason given earlier, another reason is that a category may represent multiple symptoms and only one or some of them are known to be potentially associated with the drug.

TABLE III
CLINICAL CLASSIFICATION CATEGORIES FOUND TO INDICATE POSSIBLE ADRS ASSOCIATED WITH DRUG ENALAPRIL AMONG THE TOP 50 CLASSIFICATION CATEGORIES RANKED ACCORDING TO THE ACCUMULATIVE POTENTIAL CAUSAL-LEVERAGE VALUES

Rank in top 50 signal pairs	Clinical classifications category description	Accumulative potential causal-leverage value
2	Mood disorders	4.45×10^{-4}
5	Other lower respiratory disease	3.10×10^{-4}
15	Other gastrointestinal disorders	2.06×10^{-4}
16	Nutritional deficiencies	1.91×10^{-4}
17	Chronic renal failure	1.87×10^{-4}
19	Other skin disorders	1.74×10^{-4}
24	Acute and unspecified renal failure	1.39×10^{-4}
26	Fluid and electrolyte disorders	1.26×10^{-4}
29	Other diseases of kidney and ureters	1.20×10^{-4}
31	Other upper respiratory infections	1.13×10^{-4}
35	Other male genital disorders	1.02×10^{-4}
40	Dizziness	8.71×10^{-5}
42	Other disorders of stomach and duodenum	8.43×10^{-5}
49	Other upper respiratory disease	6.74×10^{-5}

TABLE IV
RANKS FOR TWO ICD-9 CODES 276.7 AND 786.2 WHEN THRESHOLD θ TAKE DIFFERENT VALUES

ICD-9 code	ICD-9 code description	Ranks for different values of the threshold θ						
		θ=0.0	θ=0.05	θ=0.1	θ=0.15	θ=0.2	θ=0.25	θ=0.3
276.7	Hyperpotas -semia	24	21	20	21	21	26	42
786.2	Cough	123	117	116	113	112	101	108

Our physicians also evaluated the remaining 77 CCS categories. Only two of them (i.e., bronchitis and syncope) were possible ADRs. Their rankings were 69 and 93, respectively.

We also examined how threshold θ affected our interestingness measure in ranking two ICD-9 codes (276.6—hyperpotassemia and 786.2—cough) whose associations with enalapril were considered as "very likely." Table IV gives their ranks when θ takes a value between 0.0 and 0.3. The result indicates that the threshold does improve the performance of our measure when it changes from 0.0 to 0.05 or 0.1. The two ICD-9 codes respond to the threshold a little bit differently. While 276.6 reaches its highest rank at $\theta = 0.1$, 786.2 obtains its highest rank at $\theta = 0.25$. The ranks for both codes become lower when $\theta = 0.3$. Based on the data in the table, any value between 0.1 and 0.25 seems to be appropriate for θ. In general, θ should be assigned a value that makes most known potential ADRs have higher ranks.

C. Comparing Our Measure With Two Traditional Measures—Risk Ratio and Leverage

To establish the value of our new interestingness measure, we compared the ranks generated by our measure and two other measures for the two ICD-9 codes 276.7 and 786.2. These two measures are risk ratio and leverage without considering causality. Risk ratio is defined as the probability that a case coexists with the ICD-9 code relative to the probability that a noncase

TABLE V
COMPARISON OF RANKS GENERATED BY THREE INTERESTINGNESS MEASURES FOR TWO ICD-9 CODES WHOSE ASSOCIATION WITH ENALAPRIL ARE KNOWN TO BE "VERY LIKELY"

ICD-9 code	ICD-9 code description	Risk ratio		Leverage		Potential causal-leverage	
		Rank	Value	Rank	Value	Rank	Value
276.7	hyperpotassemia	571	1.45	28	5.37×10^{-4}	20	1.53×10^{-4}
786.2	cough	1137	0.56	1239	-4.91×10^{-4}	116	5.31×10^{-5}

coexists with the same ICD-9 code. The risk ratio represents a common method for investigating dichotomous variables and potential causal associations between an exposure of interest (drug) and outcome of interest (ADR). It generally relies on estimating the frequency or incidence of the outcome among an exposed cohort relative to the incidence of the outcome among an unexposed cohort. The magnitude of the risk ratio may vary from 0 to infinity. When the ratio is less than 1, the frequency or incidence of the outcome of interest is less in the exposed cohort than in the unexposed cohort and suggests that the exposure is "protective." When the risk ratio is over 1.0, the incidence of the outcome is greater in the exposed cohort than in the unexposed cohort. For drugs that produce potential ADRs, we anticipate that the incidence of ADRs will be greater in the exposed cohort (those receiving drug) than in the unexposed cohort (those not receiving drug). This measure has been used by the Medicines Control Agency in the U.K. to generate signals of possible unrecognized hazards based on their spontaneous reporting database.

Having established an estimate for the risk ratio and the estimate is greater than 1.0, we must ask whether the estimate is significant. Can the estimate that is greater than 1.0 be attributed to chance, or is it sufficiently greater to require further scrutiny? For statistical inference in these experiments, we calculated 95% confidence intervals for the estimated risk ratio [31]. For the experiments described later when the boundaries for the 95% confidence intervals exceed 1, it can be assumed that the estimated risk ratio is unlikely to have occurred by chance. The risk ratio for hyperpotassemia (ICD-9 276.6) is 1.45 (95% confidence interval is [1.00, 2.10]). In the clinical setting, this represents a very weak association between drug and outcome (potential ADR). The risk ratio for cough (ICD-9 786.2) is 0.56 (95% confidence interval [0.26, 1.05]) and would be regarded as no association between drug and outcome (potential ADR). These unimpressive estimates for the risk ratio are reflected in the low ranks reported for risk ratios in Table V. These low estimates of association between drug and potential ADR would not likely warrant labor-intensive studies of causation without other biological evidence for a causal association.

The definition of the traditional leverage measure is already given in Section II. Table V shows the comparison results generated by the three measures. For the leverage measure, while it ranks hyperpotassemia as 28, it ranks cough as low as 1239. We believe the main reason for this discrepancy is that among the 1021 patients who were taking enalapril, only 30 of them are coded as having hyperpotassemia, and only 11 of them coded as having cough. That is, the frequency of hyperpotassemia among the patients taking enalapril is 2.73 times higher than that of cough. This indicates that the traditional leverage measure can still reasonably rank high for those potential ADRs with high frequency. But it fails to distinguish those potential infrequent ADRs from other symptoms. Risk ratio and leverage represent two typical frequency-based interestingness measures. By looking at the ranks of the two "very likely" ADRs (i.e., 276.7 and 786.2), our *potential causal-leverage* measure performed much better when compared to the risk ratio and leverage measures. The results also suggest that our measure has a much stronger ability to identify infrequent potential ADRS than the two frequency-based measures.

V. DISCUSSION

Establishing sound scientific evidence for linkages between cause (drug) and effect (ADR) presents many challenges. For this discussion, we highlight two rate-limiting steps that our approach attempts to address. First, an important step in confirming causal relationships between drug and ADR depends on accumulating multiple cases in which the signal pair (i.e., ADR and drug) has been observed and occurred in an appropriate chronological sequence. Because clinicians ordering a drug may overlook a potential causal link between drug and ADR, or because physicians and pharmacists fail to report suspect signal pairs, this step may be delayed for years following introduction of the new drug. Second, after identifying a sufficient number of suspect cases (signal pairs), more thorough epidemiological analyses must be conducted to sort through possible causative factors, including drugs, and to quantify the boundaries of uncertainty around suspect drugs. This step requires the gathering of more detailed clinical information from each case and the design and execution of analytical studies that are labor intensive and rate limiting. For any given case, there are often multiple factors that might cause or contribute to the outcome of interest (ADR). This complexity is compounded by the observation that the FDA receives more than 400 000 such reports for multiple drugs each year [32]. In short, resources are insufficient to conduct analytical studies for all of the suspect signal pairs.

Our approach addresses two of the rate-limiting steps described earlier. First, the experience-based computational fuzzy RPD model—by circumventing the current passive reporting system—can significantly accelerate this rate-limiting step for identification and accumulation of pertinent cases that exhibit causality with respect to a signal pair. Second, our data mining technique—based on the *potential causal-leverage* measure—can sort and prioritize disparate signal pairs. Prioritizing potentially significant signal pairs should improve the efficiency of identifying and executing appropriate epidemiological analyses that are required to confirm causal relationships. Therefore, we want to emphasize that, like the other data mining methods, our method primarily focuses on shortening the long list of potential drug–ADR pairs. The highly ranked symptoms can be considered as ADR hypotheses. Their values and significance will be subject to further analysis (e.g., epidemiology study) and case review and interpretation by drug safety professionals

experienced in the nuances of pharmacoepidemiology and clinical medicine [33].

The experimental results in Table V demonstrated that our *potential causal-leverage* measure outperformed traditional interestingness measures such as risk ratio and leverage. Compared with traditional data mining methods used in this field, the value of our approach lies in its ability to capture suspect causal relationships. Traditional data mining approaches simply find the statistical correlation between two events X and Y. They do not specify whether X causes Y, or vice versa, or whether a third event causes the coexistence of X and Y. Our approach utilizes an experience-based fuzzy RPD model to capture the potential causal nature of a drug–symptom pair within each patient case. The causalities for all the cases are then incorporated into an interestingness measure (i.e., *potential causal leverage*). This measure can not only capture the potential causality of a pair but also address the infrequency issue. The reason is that, for a drug–symptom pair of interest, the contribution of each patient case to this interestingness measure depends on the strength of causality between the drug and the symptom within that case. Hence, relatively infrequent pairs may get a high value for the interestingness measure as long as each supporting case exhibits high causality.

Despite the advantages of our approach discussed earlier, some frequently occurring symptoms may still be falsely ranked high based on the *potential causal leverage*. The reason is that, in many cases, it is very difficult to capture the actual causality of a signal pair due to the complexity, incompleteness, and potential bias of the data. As examples, we list three situations that complicate the causality capturing process as follows: 1) if two drugs are often prescribed together, it is difficult to determine which one causes a symptom of interest based on the time-related cue values such as temporal association and rechallenge. 2) Frequently prescribed drugs are often temporally associated with symptoms by coincidence. 3) Alternatively, some potential ADRs may not be recorded in a health database, which make any data mining approach ineffective in discovering the potential ADRs.

Our experimental results suggest that existing data mining measures used for generating potential signals based on spontaneous reports will have limited utility when mining electronic health database. The risk ratio, for example, failed to effectively rank the two known signals in our study. One reason for this disparity is that data from spontaneous reports are clean and simple since the data are filtered by reporters and the reported data only include the relevant information at the time point when the suspected adverse event occurred. Electronic health data contain all data collected and entered for each patient including large amounts of data that is irrelevant to the ADR problem. Another important reason is that the causality of each reported drug–symptom pair is already implied by the corresponding report. That is, a report is normally filed only after the reporter has certain evidence (e.g., reasonable temporal association) about the causal relationship of the suspected pair. Therefore, even though measures like risk ratio cannot capture the potential causal relationship of a pair, they can still be used to signal potential signals using spontaneous reports. Electronic health data are much more complex and potentially significant drug–symptom pairs for any given patient will be obscured by multiple irrelevant pairs. For example, one complexity is that symptoms of a potential ADR cannot be differentiated from those of an underlying disease, both of which are encoded using the same ICD-9 codes in electronic health databases. For a drug–symptom pair, the associated symptom may be caused by the disease and be manifest after the drug therapy is initiated.

VI. Conclusion

We have introduced a knowledge representation PCAR and developed a new interestingness measure, *potential causal leverage*, to quantify the degree of possible causality of a PCAR. Based on this novel measure and an experience-based fuzzy RPD model, we have developed a data mining algorithm to search an electronic health database for potential ADR signals. Experimental results showed that our method could improve the efficiency of identifying a reduced list of signal pairs that will more likely prove to be clinically significant causal relationships when subjected to more rigorous analyses. Its value will be further tested by searching for different signal pairs in different electronic health databases.

Acknowledgment

The authors would like to thank E. Floyd for retrieving the patient information from the VAMC Detroit's databases.

References

[1] CDC. (2010). *International Classification of Diseases, Ninth Revision (ICD-9).* [Online]. Available: http://www.cdc.gov/nchs/icd/icd9.htm

[2] Y. Ji, R. M. Massanari, J. Ager, J. Yen, R. E. Miller, and H. Ying, "A fuzzy logic-based computational recognition-primed decision model," *Inf. Sci.*, vol. 177, pp. 4338–4353, 2007.

[3] G. A. Klein, "A recognition-primed decision making model of rapid decision making," in *Decision Making in Action: Models and Methods*, New York City, NY: Ablex Publishing, 1993, pp. 138–147.

[4] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies," *J. Amer. Med. Assoc.*, vol. 279, pp. 1200–1205, 1998.

[5] L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, *To Error is Human: Building a Safer Health System*, Washington, DC: National Academy Press, 2000.

[6] J. Talbot and P. Waller, *Stephens' Detection of New Adverse Drug Reactions*, 5th ed. Hoboken, NJ: Wiley, 2004.

[7] D. F. Klein, "The flawed basis for FDA post-marketing safety decisions: the example of anti-depressants and children," *Neuropsychopharmacology*, vol. 31, pp. 689–699, 2006.

[8] 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, pp. 2215–2220, 2002.

[9] C. Biriell and R. Edwards, "Reasons for reporting adverse drug reactions—Some thoughts based on an international review," *Pharmacol. Drug Safety*, vol. 6, pp. 21–26, 1997.

[10] C. L. Bennett, J. R. Nebeker, P. R. Yarnold, C. C. Tigue, D. A. Dorr, J. M. McKoy, B. J. Edwards, J. F. Hurdle, D. P. West, D. T. Lau, C. Angelotta, S. A. Weitzman, S. M. Belknap, B. Djulbegovic, M. S. Tallman, T. M. Kuzel, A. B. Benson, A. Evens, S. M. Trifilio, D. M. Courtney, and D. W. Raisch, "Evaluation of serious adverse drug reactions: A proactive pharmacovigilance program (RADAR) vs safety activities conducted by the food and drug administration and pharmaceutical manufacturers," *Arch. Intern. Med.*, vol. 167, pp. 1041–1049, 2007.

[11] A. Szarfman, J. M. Tonning, and P. M. Doraiswamy, "Pharmacovigilance in the 21st century: New systematic tools for an old problem," *Pharmacotherapy*, vol. 24, pp. 1099–1104, 2004.
[12] M. Lindquist, I. R. Edwards, A. Bate, H. Fucik, A. M. Nunes, and M. Stahl, "From association to alert—A revised approach to international signal analysis," *Pharmacoepidemiol. Drug Safety*, vol. 1, pp. 15–25, 1999.
[13] S. J. Evans, P. C. Waller, and S. Davis, "Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports," *Pharmacoepidemiol. Drug Safety*, vol. 10, pp. 483–486, 2001.
[14] W. DuMouchel, "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system," *Amer. Stat.*, vol. 53, pp. 177–190, 1999.
[15] P. Purcell and S. Barty, "Statistical techniques for signal generation: The Australian experience," *Drug Safety*, vol. 25, pp. 415–421, 2002.
[16] E. Heeley, L. V. Wilton, and S. A. Shakir, "Automated signal generation in prescription-event monitoring," *Drug Safety*, vol. 25, pp. 423–432, 2002.
[17] K. Kubota, D. Koide, and T. Hirai, "Comparison of data mining methodologies using Japanese spontaneous reports," *Pharmacoepidemiol. Drug Safety*, vol. 13, pp. 387–394, 2004.
[18] M. Hauben, "Early postmarketing drug safety surveillance: Data mining points to consider," *Ann. Pharmacother.*, vol. 38, pp. 1625–1630, 2004.
[19] Y. Ji, H. Ying, M. S. Farber, J. Yen, P. Dews, R. E. Miller, and R. M. Massanari, "A distributed, collaborative intelligent agent system approach for proactive postmarketing drug safety surveillance," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 826–837, May 2010.
[20] H. Jin, J. Chen, H. He, G. Williams, C. Kelman, and C. O'Keefe, "Mining unexpected temporal associations: Applications in detecting adverse drug reactions," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 4, pp. 488–500, Jul. 2008.
[21] C. Zhang and S. Zhang, *Association Rule Mining: Models and Algorithms*. New York: Springer-Verlag, 2002.
[22] M. A. Friedman, J. Woodcock, M. M. Lumpkin, J. E. Shuren, A. E. Hass, and L. J. Thompson, "The safety of newly approved medicines: Do recent market removals mean there is a problem?" *J. Amer. Med. Assoc.*, vol. 281, pp. 1728–1734, 1999.
[23] H. Jin, J. Chen, K. Chris, H. He, D. McAullay, and C. M. O'Keefe, "Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases," presented at the 10th Pacific-Asia Conf. Knowledge Discovery and Data Mining, Singapore, 2006.
[24] A. E. Fletcher, A. J. Palmer, and C. J. Bulpitt, "Cough with angiotensin converting enzyme inhibitors: How much of a problem?" *J. Hypertens. Suppl.: Official J. Int. Soc. Hypertens.*, vol. 12, pp. S43–S47, 1994.
[25] M. A. Raebel, D. L. McClure, S. R. Simon, K. A. Chan, A. Feldstein, S. E. Andrade, J. E. Lafata, D. Roblin, R. L. Davis, M. J. Gunter, and R. Platt, "Laboratory monitoring of potassium and creatinine in ambulatory patients receiving angiotensin converting enzyme inhibitors and angiotensin receptor blockers," *Pharmacoepidemiol. Drug Safety*, vol. 16, pp. 55–64, 2007.
[26] Oracle. (2010). *JDBC Overview.* [Online]. Available: http://java.sun.com/products/jdbc/overview.html
[27] OmegaSphere. (2010). *SQL Database Reference Material*. [Online]. Available: http://www.sql.org/
[28] R. Orchard, "Fuzzy reasoning in Jess: The FuzzyJ toolkit and fuzzyjess," in *Proc. 3rd Int. Conf. Enterprise Inf. Syst.*, Setubal, Portugal, 2001, pp. 533–542.
[29] *Enalapril's Side Effects.* (2010). [Online]. Available: http://www.rxlist.com/vasotec-drug.htm
[30] AHRQ. (2007). *Clinical Classifications Software (CCS) for ICD-9-CM Fact Sheet.* [Online]. Available: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp
[31] B. W. Brown and M. Hollander, *Statistics: A Biomedical Introduction*. New York: Wiley, 1977.
[32] B. L. Strom, "How the US drug safety system should be changed," *J. Amer. Med. Assoc.*, vol. 295, pp. 2072–2075, 2006.
[33] M. Hauben and X. Zhou, "Quantitative methods in pharmacovigilance: Focus on signal detection," *Drug Safety*, vol. 26, pp. 159–186, 2003.

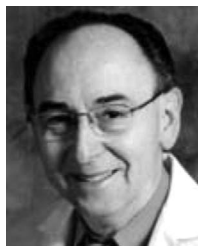
Yanqing Ji, photograph and biography not available at the time of publication.

Hao Ying (S'88–M'90–SM'97) photograph and biography not available at the time of publication.

Peter Dews, photograph and biography not available at the time of publication.

Ayman Mansour, photograph and biography not available at the time of publication.

John Tran, photograph and biography not available at the time of publication.



Richard E. Miller received the B.S. and M.D. degrees from the University of Michigan, Ann Arbor, in 1964 and 1966, respectively.

From 1975 to 1992, he was with Case Western Reserve University, and since 1992, he has been with Wayne State University, Detroit, MI. From 1992 to 2010, he was also an Associate Chief of Staff for R&D at the John D. Dingell Veterans Affairs Medical Center.

Dr. Miller is a member of the Endocrine Society and the American Diabetes Association.

R. Michael Massanari received the M.D. degree in internal medicine and infectious disease from the Feinberg School of Medicine, Northwestern University, Chicago, IL, in 1967, and completed the postgraduate training at the University of Pennsylvania, Philadelphia, and Northwestern University, Chicago. He received the M.S. degree in Preventive Medicine and Epidemiology in 1985 from the University of Iowa, Iowa City.

He is currently an Executive Director of the Critical Junctures Institute, Bellingham, WA. He was also the Director of the Center for Healthcare Effectiveness Research and a Professor of Medicine and Community Medicine, Wayne State University, until his retirement in 2009.

Prof. Massanari is a Fellow of the American College of Physicians and a member of the Academy of Health and American Public Health Association.