Noname manuscript No.
(will be inserted by the editor)

Temporality and Context for Detecting Adverse Drug Reactions from Longitudinal Data

Henry Lo · Wei Ding ⋆ · Zohreh Nazeri

the date of receipt and acceptance should be inserted later

Abstract This paper introduces a method for mining co-occurring events from longitudinal data, and applies this method to detecting adverse drug reactions (ADRs) from patient data. Electronic health records are richer than older data sources (such as spontaneous report records) and thus are ideal for ADR mining. However, current data mining methods, such as disproportionality ratios and temporal itemset mining, ignore certain important aspects of the longitudinal data in patient records. In this paper, we highlight two specific problems with current methods, which we name temporal and contextual sensitivity, and discuss why these two properties are vital to mining patterns from longitudinal data. We also propose two sensitive longitudinal rate comparison measures, which utilize condition occurrence rates and length of drug eras, for mining ADRs from this type of data. These novel methods are then used to rank potential ADRs, along with existing state-of-the-art methods, under many simulated yet realistic datasets. In 48 out of 60 experiments, the proposed longitudinal rate comparison methods significantly outperform other methods in mining known ADRs from other drug / condition pairs.

Keywords data mining · disproportionality methods · adverse drug reactions · longitudinal data

1 Introduction

Both drug availability and drug consumption increase every year, making adverse drug reactions (ADRs) an increasingly pervasive and difficult problem. In the

⋆ Corresponding author.

Henry Lo and Wei Ding ✉
Department of Computer Science
University of Massachusetts Boston, USA
E-mail: henryzlo@cs.umb.edu, ding@cs.umb.edu

Zohreh Nazeri
MITRE Corporation, USA
E-mail: nazeri@mitre.org

United States alone, ADRs send 700,000 patients to emergency rooms and account for 120,000 hospitalizations per year [5]. All of this amounts to an estimated cost of $177.4 billion per year [8]. Mortality due to ADRs continues to grows year after year: just in the FDA reporting system alone, 63,839 ADR-related deaths were reported in 2009; 82,724 in 2010, and 98,518 in 2011 [25].

To detect ADRs as soon as possible, it is imperative to monitor drug usage and condition occurrence in healthcare institutions. Though self-reporting systems (SRS) exist for patients to document their ADRs, data from hospital electronic health records (EHRs) are of much higher quality. These records are entered by trained professionals, and contain much more patient data, such as demographic information and risk-related variables.

Detecting ADRs from patient data is an example of mining co-occurring events, in which events are drug usage and condition occurrence. Many methods exist for this general problem. However, for ADR detection, viable methods must contain the following three properties:

- Temporal sensitivity: must be able to take into account the lengths of time, rather than just counts.
- Contextual sensitivity: must be able to consider the strength of a potential ADR with relation to the condition's contextual information.
- Statistical significance: must be able to provide p-values in order to establish significance.

Temporal and contextual sensitivity are important not only for ADR detection, but also for mining temporally correlated events. Detection of temporal correlation (which implies causality) has been studied in artificial intelligence and data mining literature [16,18]. However, we argue that for a temporal pattern to be meaningful in any problem domain, context and length of time must be taken into account. Without this information, it is impossible to determine whether or not an event following another is truly significant.

Current methods for temporal pattern mining (such as methods based on support and confidence) lack contextual sensitivity, as they do not consider occurrence rates outside of co-occurrences. Others are not grounded in statistics [1, 13, 17]. Current methods to mine ADRs, which are based on proportionality ratios, are grounded in statistics, but these lack temporal sensitivity, as do association rule based methods [27]. Episode mining algorithms, which take into account temporal contexts, do not provide measures of statistical significance [14, 26].

We introduce a Poisson-based method which is able to mine statistically significant event patterns, and is temporally and contextually sensitive. This method is based on the comparison of condition occurrence rates; thus, we call it longitudinal rate comparison (LRC). LRC is contextually sensitive due to its comparisons, and is temporally sensitive due to its reliance on occurrence rates. These properties are not shared in any other method. We provide two techniques to perform the rate comparison, LRC-ratio and LRC-SD, both of which are grounded in statistics and are able to provide p-values. Our experiments on simulated EHR datasets show that the LRC-ratio and LRC-SD methods obtain better ROC curves than existing methods.

The contributions of this paper are thus:

– A systematic analysis of the temporal sensitivity and contextual sensitivity properties, explaining why they are necessary for mining patterns in longitudinal data.
– The contextually and temporally sensitive LRC methods for extracting ADRs from longitudinal data.

The rest of the paper is organized as follows. In section 2, we operationalize the problem of mining ADRs, and introduce existing methods. Section 3 describes temporal and contextual sensitivity, and issues with existing methods. In section 4, we introduce the LRC methods, which bypasses the problems inherent in using proportionality methods in EHR data. Experiments in section 5 then compare the two LRC methods against different versions of disproportionality measures. The paper concludes in section 6.

2 Background

2.1 Problem Definition

The central problem of this paper is extracting potential ADRs from a set of longitudinal patient data. This longitudinal dataset contains information about patients over time, such as condition occurrences and drug usage periods. From this, drug/condition pairs which may be potential adverse drug reactions (ADRs) must be selected and ranked.

Operationally, potential ADRs can be detected by identifying abnormally high rates of condition occurrence among drug users. As will be described, there exist many methods for quantifying this abnormality.

2.2 A Framework for ADR Mining

The general problem of mining itemsets can be summarized in the following framework. Without loss of generality, we use itemsets of size two:

1. For each possible set of items x_1, x_2, extract the co-occurrence counts:
   – Number of data samples in which x_1 and x_2 co-occur, $n_{x_1 x_2}$.
   – Number of data samples which contain x_1 but not x_2, $n_{x_1 \overline{x_2}}$.
   – Number of data samples which contain x_2 but not x_1, $n_{\overline{x_1} x_2}$.
   – Number of data samples which contain neither x_2 nor x_1, $n_{\overline{x_1 x_2}}$.
2. From these counts, extract an interestingness measure.
3. Determine a cutoff for interestingness, usually either fixed or derived from some computation.
4. If itemset interestingness meets this cutoff, then mine out this pattern; otherwise, discard.

In the ADR mining problem domain, x_1 is a drug d and x_2 is a condition c.

The interestingness measures used in ADR mining are called disproportionality (or proportionality) ratios. These measures are calculated from co-occurrence values, and are typically higher for drug / condition pairs more likely to be true ADRs [3, 6, 10, 19].

	Condition	No condition	
Drug	n_{dc}	$n_{d\bar{c}}$	n_d
No drug	$n_{\bar{d}c}$	$n_{\overline{dc}}$	$n_{\bar{d}}$
	n_c	$n_{\bar{c}}$	n

Table 1 Contingency table. For some given drug and condition, n_{dc} refers to the number of subjects who have taken the drug and experienced the condition; $n_{d\bar{c}}$, subjects who have taken the drug without experiencing the condition; $n_{\bar{d}c}$, subjects who have not taken the drug, but experienced the condition; $n_{\overline{dc}}$, subjects who have neither taken the drug nor experienced the condition.

Though these methods were developed for SRS data, Zorych et al. developed co-occurrence counting methods for longitudinal data, and thus extended disproportionality ratios to longitudinal data [27]. Instead of counting non-occurring conditions, Zorych's method counts the occurrence of *other* conditions; likewise, non-drug events are counted by the usage of other drugs.

Cutoffs for proportionality ratios are usually determined either by convention (e.g. when the probability of a type I error $\alpha < 0.05$ [3]) or by an arbitrary threshold. Some methods do not specify a specific threshold, while some studies neglect thresholds altogether. Indeed, many studies compare methods by their rank ordering of potential ADR pairs, and with evaluation methodologies such as ROC curves, cutoffs do not need to be defined. Some work has also been done using the false discovery rate (FDR) as the cutoff statistic, with a cutoff of 0.05 [2].

2.3 Proportionality Ratios

Most accepted ADR methods use proportionality measures. These measures quantify how often a drug and condition co-occur compared to some baseline rate, and mostly differ in the definition of the baseline. Potential ADRs are mined by specifying a cutoff, usually defined using confidence intervals, for these measures.

All disproportionality measures are ratios or combinations of the four count values present in a contingency table, as shown in Table 1. The variable names n_{dc},$n_{d\bar{c}}$,$n_{\bar{d}c}$, and $n_{\overline{dc}}$ used in that table will be used to define each of the disproportionality measures.

2.3.1 Proportional Reporting Ratio

The proportional reporting ratio (PRR), as defined in Equation 1, compares occurrences of a condition c in the presence of a drug d with occurrences in the drug's absence [9, 10]. Larger PRR signifies a positive effect of d on c, smaller numbers signify negative effects, and a PRR of one signifies no effect. This is a widely used measure in many studies, and is used in the UK Medicines Control Agency (MCA) [10].

$$PRR = \frac{n_{dc}/n_d}{n_{\bar{d}c}/n_{\bar{d}}} \tag{1}$$

PRR divides the dataset into two groups depending on the presence of d, then returns the ratio of the measures in both groups. PRRs greater than two, a count of over three, and a χ^2 value over four, are used in the literature to signify potential reactions [10, 11]. The χ^2 value is calculated using the contingency table.

2.3.2 Reporting Odds Ratio

The reporting odds ratio (ROR) [19], shown in Equation 2, uses ratios of condition occurrence to non-occurrence. The test hinges on the proportions of this ratio for drug-taking subjects versus non-drug taking subjects. It is currently being used by the Netherlands Pharmacovigilance Centre to detect ADRs [19].

$$ROR = \frac{n_{dc}/n_{d\bar{c}}}{n_{\bar{d}c}/n_{\overline{dc}}} \tag{2}$$

The drug-condition pair is signaled if the lower limit of the two-sided 95% confidence interval exceeds 1. This confidence interval can be calculated using Equation 3, from [19, 24].

$$ROR_{0.05} = \exp\left(\ln(ROR) \pm 1.96\sqrt{\left(\frac{1}{n_{dc}} + \frac{1}{n_{d\bar{c}}} + \frac{1}{n_{\bar{d}c}} + \frac{1}{n_{\overline{dc}}}\right)}\right) \tag{3}$$

In practice, condition occurrence (n_c) is very low for any given condition. Thus, the ROR and PRR measures tend to give very similar numbers, though their cutoffs differ.

2.3.3 Information Component

Bayesian confidence propagation neural networks (BCPNN) utilize a disproportionality measure known as information component (IC) [3]:

$$\begin{aligned} IC &= \log_2 \frac{p(c,d)}{p(c)p(d)} \\ &= \log_2 \frac{(n_{dc}/n)}{(n_c/n)(n_d/n)} \end{aligned}$$

where d is a drug, and c is a condition. IC is 0 if d and c are totally independent; negative if d causes a decrease in the occurrence probability of c, and positive if it causes an increase.

IC is assumed to be normally distributed. Thus, its credible interval (IC is a Bayesian method) can be determined using its expected value and its standard deviation.

The expected value and variance of the IC are given as follows [3]:

$$\begin{aligned} E(IC) &= \log_2 \frac{(n_{dc}+1)(n+2)^2}{(n+1)(n_d+1)(n_c+1)} \\ Var(IC) &= \frac{1}{(\ln 2)^2}\left(\frac{n-n_{dc}+\gamma-1}{(n_{dc}+1)(1+n+\gamma)} + \frac{n-n_d+1}{(n_d+1)(3+n)} + \frac{n-n_c+3}{(n_c+1)(3+n)}\right) \end{aligned}$$

where:

$$\gamma = \frac{(n+1)^2}{(n_d+1)(n_c+1)}$$

The credible interval for $\alpha = 0.05$ is therefore:

$$IC_{0.05} = E(IC) \pm 1.96 * SD(IC) \tag{4}$$

If the lower bound of this interval is greater than 0, a possible ADR is signaled. All equations can be found in [3].

BCPNNs have been used to mine ADRs from the WHO Uppsala Monitoring Centre database since 1998 [3,4,15]. Derivation of the equations listed and rationale can be found in the original paper by Bate [3].

The IC measure has also been extended for temporal data, using time lengths as well as condition counts. It does so by changing the IC term [21]:

$$\begin{aligned} IC &= \log_2 \frac{p(c,d)}{p(c)p(d)} \\ &= \log_2 \frac{x_1}{x_2/t_{\bar{d}} * t_d} \end{aligned}$$

Here, x_1 is the number of times c occurs in the presence of d, x_2 is the number of times c occurs elsewhere, $t_{\bar{d}}$ is the time spent not under the drug, t_d is the time spent under the drug. All other equations remain the same.

2.3.4 Gamma Poisson Shrinker

The Gamma Poisson shrinker (GPS) uses the following proportionality ratio [6]:

$$\lambda = \frac{n_{dc}}{n_d n_c / n} \tag{5}$$

This ratio is converted into the empirical Bayesian geometric mean (EBGM) [6].

$$EBGM = \exp(E(\log \lambda) \tag{6}$$

The expected value of $\log \lambda$ is obtained by means of a Bayesian update; λ is assumed to be drawn from a Poisson distribution, and its prior is chosen to be a mixture of two Gamma distributions. The credible interval is chosen in the same way. For more details see [6, 7].

2.4 Co-occurrence Counting in Longitudinal Data

Zorych et al. proposed three different methods for defining contingency counts based on longitudinal data [27]. The *SRS* method counts a co-occurrence whenever a drug co-occurs with a condition, and estimates condition non-occurrence using the occurrence of other conditions. In this way, this method attempts to generate reports as seen in SRS data – one report for each co-occurring drug era and condition occurrence.

This method does not count drugs which do not occur with conditions, or conditions which do not occur with drugs. This is rectified in the *modified-SRS* method [27]:

- n_{dc} counts occurrences of c during exposure to d.
- $n_{d\bar{c}}$ counts occurrences of conditions that are not c during exposure to d, plus the exposures to d during which no conditions occur.
- $n_{\bar{d}c}$ counts occurrences of c during exposure to drugs that are not d.

– $n_{\overline{dc}}$ counts occurrences of conditions that are not c during exposure to drugs that are not d, plus the occurrences of c with no corresponding drug, plus the exposures to d with no reported conditions.

This counting method and others were used to obtain contingency values from longitudinal data. These contingency values were used by various methods to detect ADRs. Their results indicated that modified-SRS yielded the best ADR detection performance, though the SRS method was close [27].

3 Analysis

In the context of existing ADR mining methods, we introduce the contexual and temporal sensitivity properties, and discuss how they apply to longitudinal data.

In order to determine what conditions are abnormally frequent, we must compare them to some baseline. In other words, the interestingness of any potential ADR must depend partially on the context of the potential ADR. We call this *contextual sensitivity*. Disproportionality measures, such as PRR and ROR, have this property, due to comparisons between condition occurrences in the presence and absence of drug usage. However, measures utilizing only the absolute quantity of condition occurrences, such as support and confidence, do not.

Any interestingness measure must also take into account the time spent under a drug. This is because determining some occurrence count to be abnormal is not sufficient - we are interested in abnormal frequency, which considers time also. Three conditions occurring within a year of drug usage is not as suspicious as three conditions within a week. We deem this property *temporal sensitivity*.

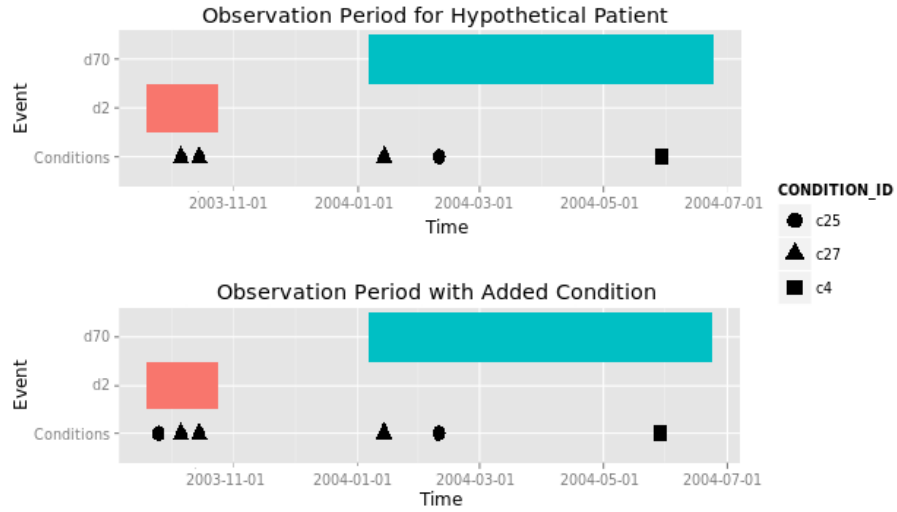
Though disproportionality methods are appropriate for SRS data, their effectiveness is hampered by their reliance on co-occurrence counts when applied to longitudinal data [27], as we show below.

3.1 Contextual Sensitivity

Disproportionality ratio methods are inherently contextually sensitive, as they compare some measure of frequency with some baseline (hence the "ratio"). Various methods define both the frequency measure and the baseline measure in different ways. However, adapting these methods for longitudinal data introduces bias for contextual sensitivity.

Calculating contingency values requires counting the non-occurrence of events. Specificially, $n_{\bar{d}c}$ counts the absence of a drug, $n_{d\bar{c}}$, counts the absence of a condition, and $n_{\overline{dc}}$ counts the absence of both. Zorych's methods count condition non-occurrences whenever a different condition occurs while a patient is taking (or not taking, depending on the count value in question) a certain drug, and similarly for drug non-occurrences.

Though a natural adaptation, using this scheme, $n_{d\bar{c}}$ counts increase with the number of non-c condition events in the time period under consideration. This affects disproportionality ratios, when intuitively, the presence or absence of other conditions should have no bearing on whether or not a drug causes a given condition. Thus, there exists a dependence of the disproportionality measure on occurrences of conditions not under question.



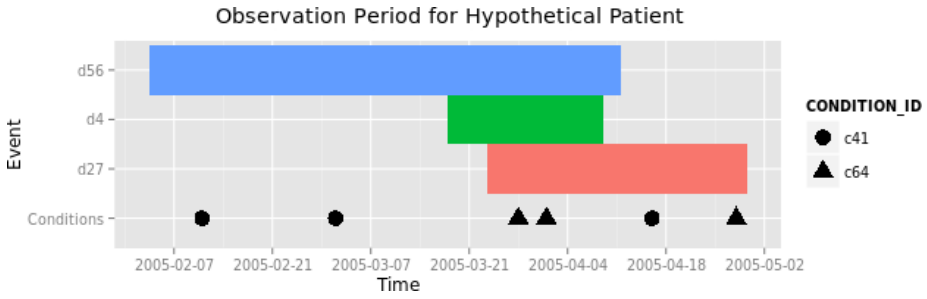
	n_{dc}	$n_{d\bar{c}}$	$n_{\bar{d}c}$	$n_{\bar{d}\bar{c}}$	PRR	Potential ADR?
Observation Period	2	0	1	2	2	Yes
Observation with Added Condition	2	1	1	2	1.3	No

Fig. 1 Diagram demonstrating the influence of unrelated conditions. The pair under consideration is d2 and c27. Rectangles represent drugs; drug labels are on the left side of drug rows.

The effect is shown in Figure 1. In this figure, shapes represent condition occurrences (c4, c25, c27), rectangles (d2, d70) represent different drugs, and the area that rectangles cover represents a period of time during which the drug was used. Two observation periods are shown. Below the graphic is a table showing the contingency counts associated with conditions 41 and 64, when counted using the Zorych methods [27].

Condition 4 co-occurs with drug 2 twice in both conditions, but note that the PRRs (calculated using Equation 1) are different. The only difference between the two conditions is the addition of one unrelated condition occurrence (c25 in the second observation), yet this simple change results in the PRR going from 2 (potential ADR) to 1.3 (not a potential ADR). This demonstrates the potential effect that unrelated conditions have on disproportionality ratios, when using the Zorych SRS counting method [27].

Apart from the problem of unrelated condition influence, longitudinal adaptation also over-counts the number of other events in total. As an example, observe the patient data in Figure 2. Both conditions 41 and 64 occur thrice. However, note how many reports are generated for each. For condition 41, three reports are generated, since each condition occurrence co-occurs with one drug era. For condition 64, however, two condition occurrences are simultaneous with three drug eras. This results in a report for each condition occurrence, for each drug era, yielding 7 reports. Despite the fact that both conditions occurred the same number of times, the marginal count of condition 64 is more than twice as high as that of condition



	n_{dc}	$n_{d\bar{c}}$	$n_{\bar{d}c}$	$n_{\bar{d}\bar{c}}$	PRR	Potential ADR?
Condition 41	2	1	1	5	4	Yes
Condition 64	2	1	5	1	0.8	No

Fig. 2 Diagram demonstrating over-counting. The drug under consideration is d56. Rectangles represent drugs; drug labels are on the left side of drug rows.

41. This results in a much higher PRR for condition 41, as shown in the table in Figure 2.

For patients with enough simultaneous events, this could theoretically lead to a combinatorial explosion of "reports", though the number of coinciding conditions and drug usage eras is in practice low enough to mitigate this. Nevertheless, over-counting non-occurring events biases results, as shown in the example.

3.2 Temporal Sensitivity

Apart from the issues arising from counting non-occurrences, there are issues with merely counting occurrences as well. Longitudinal data contains temporal information such as the duration of a drug era, which cannot be accounted for in contingency counts. Yet drug era duration is important; a patient experiencing a condition three times within one week is much more worrisome than the same number of condition occurrences within a year.

For a more concrete example, observe the two patients in Figure 3, each suffering several occurrences of only one condition. Both patient observations yield the same contingency values; specifically, the n_{dc} and $n_{\bar{d}c}$ counts of drug 15 and the condition are both 3. Thus, any disproportionality method would yield the same result for both the above and below case.

Intuitively, this should not be the case. The bottom patient has had the same number of condition occurrences while taking drug 15, but in a much shorter time span. The top patient's condition occurrences could have been due to chance, but it seems that the bottom patient's conditions are much less likely to be so.

4 Longitudinal Rate Comparison

To attain both temporal and contextual sensitivity, we propose a method which is based on comparing condition occurrence frequencies within different contexts. Essentially, using frequencies allows us to take into account length of time, and

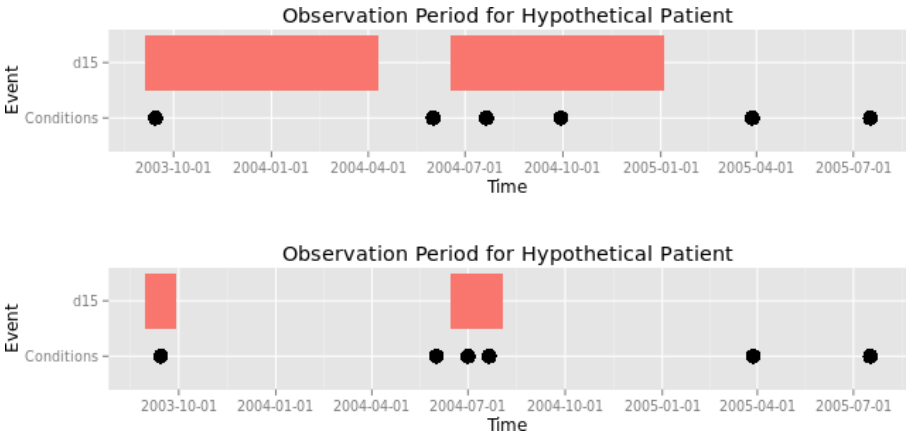


Fig. 3 Diagram demonstrating temporal negligence.

comparing rates gives contextual sensitivity to the method. We call the proposed method longitudinal rate comparison (LRC).

4.1 Counting Longitudinally

Rather than the four contingency values used in disproportionality ratios, LRC depends on four values:

- $n_{dc} = x_1$, the drug/condition co-occurrence count.
- $n_{\bar{d}c} = x_2$, the condition occurrence count in the absence of the drug.
- t_d, the time spent in a drug era.
- $t_{\bar{d}}$, the time spent outside a drug era.

Condition occurrence counts (n_{dc} and $n_{\bar{d}c}$) are necessary for determining the significance of a drug/condition pair; without comparing condition prevalence both in and out of drug eras, it is impossible to know whether a condition is abnormally frequent under the presence of a drug, as we have no baseline. Thus, we retain these counts, renaming them x_1 and x_2.

The time that a subject spends in and out of drug eras for a given drug is required to avoid temporal negligence. These times, added up over all subjects, are denoted t_d and $t_{\bar{d}}$.

Given a drug, the LRC method divides the observation period of every patient into the time periods within which the patient was taking the drug (time period with length t_d), and the time within which the drug was not taken (time period with length $t_{\bar{d}}$). These time windows can be extended to include surveillance periods.

For each condition, its count is calculated in the drug-taking time windows, yielding the condition count value x_1, and in the non-drug-taking time windows, yielding x_2. We also consider the total time taking the drug, t_d, and the total time not taking the drug, $t_{\bar{d}}$.

4.2 Interestingness and Threshold

Using the condition occurrence rates $\lambda_1 = x_1/t_d$ and $\lambda_2 = x_2/t_{\bar{d}}$. We define two interestingness measures, one based on the rate ratio (LRC-ratio), and another based on the rate standardized difference (LRC-SD), and a method for extracting significance values from them.

For both measures, we signal a potential ADR for a given condition / drug pair when the condition occurrence rates λ_1 and λ_2 differ with p-value less than 0.05.

4.2.1 LRC-ratio

One possible interestingness value is the ratio of the two rates, which we call LRC-ratio (or LRC-R for short):

$$\text{LRC-ratio} = \frac{\lambda_1}{\lambda_2} \tag{7}$$

The p-value for LRC-R can be calculated using the C-test, which determines the significance of two Poisson rates [23]. The method uses the fact that x_1 is distributed binomially, when conditioned on $n = x_1 + x_2$ and $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$:

$$P(X_1 \leq x_1 | n, p) = \sum_{k=0}^{n} B(k, p)$$

For our two sided C-test, with the null hypothesis $\lambda_1 = \lambda_2$, the p-value is given by the formula:

$$2 * \min(P(X_1 \geq x_1 | n, p), 1 - P(X_1 \geq x_1 | n, p)) \tag{8}$$

When the condition occurrence rates are very similar, then we do not signal a potential ADR, and LRC-R should be around 1.

4.3 LRC-standard deviation

LRC-SD uses the standardized difference of two condition counts as an interestingness measure. The benefit of LRC-SD over LRC-R is that it uses an estimate of the standardized variance; thus, it shrinks towards 0 as variance increases.

If the null hypothesis is that the two observed condition occurrence rates λ_1 and λ_2 are equal, then the standardized difference

$$SD = \frac{\lambda_1 - \lambda_2}{\sqrt{\hat{V}}} \tag{9}$$

should be 0. In this formula, $\hat{V}$ is estimated variance, given by

$$\hat{V} = \frac{x_1}{t_d^2} + \frac{x_2}{t_{\bar{d}}^2}$$

To calculate the p-value for Equation 9, we add together all probabilities in the joint space where the standardized differences T_{k_1,k_2} and T_{x_1,x_2} are the same [12]:

$$\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\frac{(t_d\hat{\lambda}_2)^{x_1}\exp(-t_d\hat{\lambda}_2)}{x_1!}\frac{(t_{\bar{d}}\hat{\lambda}_2)^{x_2}\exp(-t_{\bar{d}}\hat{\lambda}_2)}{x_2!}[I(T_{k_1,k_2}=T_{x_1,x_2})] \qquad (10)$$

where I is the indicator function and where: $\hat{\lambda}_2=\frac{k_1+k_2}{t_d+t_{\bar{d}}}$.

When the condition occurrence rates are very similar, then we do not signal a potential ADR, and LRC-SD should be around 0.

4.4 Example

As an example, consider a condition c and drug d, experienced and used by several subjects. Suppose that all the subjects in our sample have collectively used d for a total of 300 days, and within this time, they suffered c 18 times. The total observation time minus the time spent under d is 500 days, and in this period, subjects suffered c only twice.

From this, we get $x_1=18$, $x_2=2$, $t_d=300$, $t_{\bar{d}}=500$, and:

- $\lambda_1=\frac{18}{300}=0.06$
- $\lambda_2=\frac{2}{500}=0.004$
- LRC-ratio $=15$
- LRC-SD $=3.9598,$ since $\hat{V}=0.0002$

Since LRC-ratio is much greater than 1, and LRC-SD is larger than 0, it is likely that d and c is a potential ADR. However, we cannot trust these ratios unless their p-values, which we can calculate using Equations 8 and 10, are significant. In our experiments, we use a significance threshold of 0.05.

5 Experiments

5.1 Data

Longitudinal patient data was generated using the Observational Medical Dataset Simulator Generation 1 (OSIM) [22].

5.1.1 Observational Medical Dataset Simulator

OSIM was created by the Observational Medical Outcomes Partnership, a non-profit agency with close ties to the FDA, in order to identify reliable methods for mining huge volumes of data for ADRs. The advent of OSIM has circumvented several difficult problems, namely:

- The difficulty of obtaining patient data, and thus evaluating ADR mining methods, due to privacy laws such as HIPAA (Health Insurance Portability and Accountability Act).
- Providing a method for injecting drug/condition reactions into the data, thereby providing a ground truth and facilitating method evaluation.

– Obtaining a high-quality dataset which is scalable, modifiable, and realistic, yet anonymous and posing no threat to patient privacy.

The OSIM methodology accomplishes the latter goal by first obtaining probabilities about patient demographics, drug usage, condition occurrences, and other variables. With this information, the software generates datasets which preserve these distributions, completely de-anonymizing the data. For more details about the OSIM process, see [20, 22].

5.1.2 Datasets

Fifteen datasets were generated using OSIM. All datasets contained 25,000 simulated persons; the differences in their parameters are described in Table 2.

Dataset ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Drugs	30	30	30	30	30	35	35	35	35	35	40	40	40	40	40
Conditions	30	35	40	45	50	30	35	40	45	50	30	35	40	45	50

Table 2 Run parameters for OSIM data generation. "Drugs" refers to number of drugs generated for that specific dataset; likewise for "Conditions".

With the baseline probabilities, very large datasets would have to be generated in order to observe an adequate number of occurrences of drug/condition reactions, including those found in the ground truth. Thus, condition prevalence and drug prevalence were increased so as to fall within the 1% - 10% category.

5.2 Methodology

For each drug/condition pair which occurs in the data, we extract the four contingency values and the time values $t_d, t_{\bar{d}}$. From these, all other calculations can be done.

We use four different surveillance period lengths ($t = 0, t = 7, t = 30, t = 365$) in order to detect delayed ADR effects. The value of t specifies the length of time (in days) after drug usage has subsided in which we still consider a condition to be possibly linked to that drug. For example, if $t = 7$, a condition appearing 6 days after stopping drug usage is still considered a co-occurrence, but one appearing 8 days after is not.

We tested several different detection methodologies: ROR, PRR, RFET, IC temporal pattern detection (TimeIC), BCPNN (IC), GPS, LRC-R, and LRC-SD.

The typical use case for these interestingness values is to use some sort of cutoff to detect potential ADR signals, then rank these signal with the interestingness measure.

Thus, in our experiments, we cull potential ADRs using methods specific to each proportionality-ratio method, as done in their original papers. In particular, for the following methods, drug/condition pairs are culled as follows:

– PRR: pairs are culled if PRR is less than 2.
– ROR: pairs are culled if the lower boundary of the two-sided 95% confidence interval is less than or equal to 1.

- RFET: pairs are culled if their false discovery rate is less than 0.05.
- BCPNN: pairs are culled if the lower boundary of the two-sided 95% confidence interval is less than or equal to 0.
- TimeIC: pairs are culled if the lower boundary of the two-sided 95% confidence interval is less than or equal to 0.
- GPS: pairs are culled if the one-sided 95% confidence interval is less than 1.
- LRC-R and LRC-SD: pairs are culled if their p-values are not less than 0.05.

In addition, pairs are removed if they do not contain at least 3 co-occurrences.

Methods are evaluated using a receiver operator characteristic curve (ROC) after culling. The ROC curve is a standard way of comparing a method's true positive rate with its false positive rate, as the discrimination threshold for the interestingness value is varied. Area under the curve (AUC) is also calculated. Larger area means that the method better separates the ADR signals from the non-signals.

5.3 Results

AUCs for all experiments are shown in Table 3. There seems to be a wide range of difficulty for each dataset - for example, dataset 2 is relatively diffficult for all methods, and dataset 1 is relatively easy. There does not seem to be an obvious relationship between the value of t, or the number of conditions and drugs in a dataset, with the dataset's difficulty.

LRC-SD achieves the highest AUC for 48 out of the 60 testing conditions, with one of these being a tie. The performance of the GPS, TimeIC, and IC methods is close to LRC-SD, probably due to their usage of statistical shrinkage.

Calculating the significance of these AUC results is complicated by the large range of difficulties between datasets. Figure 4 shows the means and standard errors of all ADR measures tested, averaged over all datasets. Note that between different values of t, error regions overlap, suggesting that differences between values of t are not significant. Also, the standard deviations actually increase for some methods (PRR, ROR, GPS) when averaged across all t (see darkest columns in Figure 4). This suggests that a simple comparison of means will not suffice.

The large differences between datasets suggests a pairwise test approach. Justifying this is a preliminary ANOVA with a significant p-value of less than 2×10^{-16}, showing that there is indeed a difference between methods. The results of the pairwise t-tests with a Holm-Bonferroni adjustment are shown in Table 4. The extremely small p-values suggest a significant different between the AUC values achieved by LRC-SD, and the AUC values of all others.

6 Conclusion

This paper explores mining ADRs from longitudinal data, introduces the ideas of conceptual and temporal sensitivity, and proposes two sensitive LRC measures for mining ADRs in such environment. Temporal and contexual sensitivity are important properties not only for ADR mining, but also for the general problem of extracting temporal event patterns, a common topic in data mining. Through

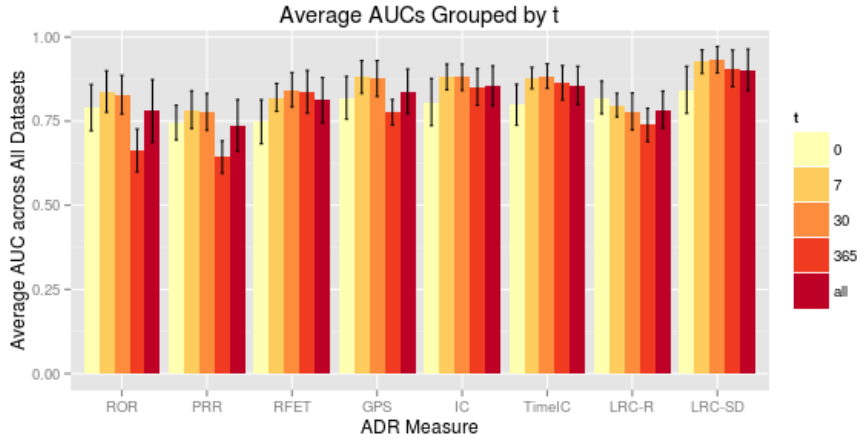


Fig. 4 The results of all experiments; each bar represents AUCs averaged across all datasets for each measure and value of t. Error bars represent one standard deviation about and below the mean.

extensive experiments, we demonstrate that the LRC measures, which are contextually and temporally sensitive, significantly improve over the state of the art methods in the longitudinal context.

7 Acknowledgements

The authors would like to thank MITRE corporation for its support of this project in 2013.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93, pp. 207 – 216. ACM, New York, NY, US (1993). DOI 10.1145/170035.170072. URL http://doi.acm.org/10.1145/170035.170072
2. Ahmed, I., Dalmasso, C., Haramburu, F., Thiessard, F., Brot, P., Tubert-Bitter, P.: False discovery rate estimation for frequentist pharmacovigilance signal detection methods. Biometrics **66**(1), 301–309 (2010). DOI 10.1111/j.1541-0420.2009.01262.x. URL http://dx.doi.org/10.1111/j.1541-0420.2009.01262.x
3. Bate, A.: The use of bayesian confidence propagation neural network in pharmacovigilance. Ph.D. thesis, Ume University, Pharmacology and Clinical Neuroscience (2003)
4. Bate, A., Lindquist, M., Edwards, I.R., Olsson, S., Orre, R., Lansner, A., De Freitas, R.M.: A Bayesian neural network method for adverse drug reaction signal generation. Eur. J. Clin. Pharmacol. **54**(4), 315–321 (1998)
5. DS, B., DA, P., KN, W., AB, M., TJ, S., JL, A.: National surveillance of emergency department visits for outpatient adverse drug events. JAMA **296**(15), 1858–1866 (2006). DOI 10.1001/jama.296.15.1858. URL + http://dx.doi.org/10.1001/jama.296.15.1858
6. Dumouchel, W.: Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. The American Statistician **53**(3), 177–190 (1999). DOI 10.1080/00031305.1999.10474456

7. DuMouchel, W., Pregibon, D.: Empirical bayes screening for multi-item associations. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining pp. 67–76 (2001). DOI 10.1145/502512.502526. URL http://doi.acm.org/10.1145/502512.502526
8. Ernst, F.R., Grizzle, A.J.: Drug-related morbidity and mortality: updating the cost-of-illness model. J Am Pharm Assoc (Wash) **41**(2), 192–199 (2001)
9. Evans, S.: Statistical Methods of Signal Detection, pp. 273–279. John Wiley & Sons, Ltd (2002). DOI 10.1002/0470853093.ch20. URL http://dx.doi.org/10.1002/0470853093.ch20
10. Evans, S.J., Waller, P.C., Davis, S.: Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and drug safety **10**(6), 483–486 (2001). DOI 10.1002/pds.677. URL http://dx.doi.org/10.1002/pds.677
11. Hauben, M., Reich, L.: Potential utility of data-mining algorithms for early detection of potentially fatal/disabling adverse drug reactions: a retrospective evaluation. J Clin Pharmacol **45**(4), 378–384 (2005)
12. Krishnamoorthy, K., Thomson, J.: A more powerful test for comparing two poisson means. Journal of Statistical Planning and Inference **119**(1), 23 – 35 (2004)
13. Li, Y., Ning, P., Wang, X., Jajodia, S.: Discovering calendar-based temporal association rules. Data and Knowledge Engineering **44**(2), 193 – 218 (2003). DOI http://dx.doi.org/10.1016/S0169-023X(02)00135-0. URL http://www.sciencedirect.com/science/article/pii/S0169023X02001350
14. Lin, S., Qiao, J., Wang, Y.: Frequent episode mining within the latest time windows over event streams. Applied Intelligence **40**(1), 13–28 (2014). DOI 10.1007/s10489-013-0442-8. URL http://dx.doi.org/10.1007/s10489-013-0442-8
15. Lindquist, M., Stahl, M., Bate, A., Edwards, I.R., Meyboom, R.H.B.: A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Saf **23**(6), 533–542 (2000)
16. Liu, W., Zheng, Y., Chawla, S., Yuan, J., Xing, X.: Discovering spatio-temporal causal interactions in traffic data streams. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, pp. 1010–1018. ACM, New York, NY, USA (2011). DOI 10.1145/2020408.2020571. URL http://doi.acm.org/10.1145/2020408.2020571
17. Mannila, H., Toivonen, H.: Discovering generalized episodes using minimal occurrences. In: In KDD 96: Proc. 2nd International Conference on Knowledge Discovery and Data Mining, pp. 146–151. AAAI Press (1996)
18. Mohan, P., Shekhar, S., Shine, J., Rogers, J.: Cascading spatio-temporal pattern discovery. Knowledge and Data Engineering, IEEE Transactions on **24**(11), 1977–1992 (2012). DOI 10.1109/TKDE.2011.146
19. Moore, N., Kreft-Jais, C., Haramburu, F., Noblet, C., Andrejak, M., Ollagnier, M., Bgaud, B.: Reports of hypoglycaemia associated with the use of ACE inhibitors and other drugs: a case/non-case study in the French pharmacovigilance system database. Br J Clin Pharmacol **44**(5), 513–518 (1997)
20. Murray, R.E., Ryan, P.B., Reisinger, S.J.: Design and validation of a data simulation model for longitudinal healthcare data. AMIA Annu Symp Proc **2011**, 1176–1185 (2011)
21. Noren, G., Hopstadius, J., Bate, A., Star, K., Edwards, I.: Temporal pattern discovery in longitudinal electronic patient records. Data Mining and Knowledge Discovery **20**(3), 361–387 (2010). DOI 10.1007/s10618-009-0152-3. URL http://dx.doi.org/10.1007/s10618-009-0152-3
22. (OMOP), O.M.O.P.: Observational medical dataset simulator generation 1 (2009). Available from OMOP at http://www.omop.org
23. Przyborowski, J., Wilenski, H.: Homogeneity of results in testing samples from poisson series: With an application to testing clover seed for dodder. Biometrika **31**(3-4), 313–323 (1940). DOI 10.1093/biomet/31.3-4.313. URL http://biomet.oxfordjournals.org/content/31/3-4/313.short
24. Sakaeda, T., Tamon, A., Kadoyama, K., Okuno, Y.: Data mining of the public version of the FDA adverse event reporting system. Int J Med Sci **10**(7), 796–803 (2013)
25. U.S. Food and Drug Administration: FAERS Patient Outcomes by Year. U.S. Food and Drug Administration (2012). URL http://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance
26. Yen, S.J., Lee, Y.S.: Mining non-redundant time-gap sequential patterns. Applied Intelligence **39**(4), 727–738 (2013). DOI 10.1007/s10489-013-0426-8. URL http://dx.doi.org/10.1007/s10489-013-0426-8

27. Zorych, I., Madigan, D., Ryan, P., Bate, A.: Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res **22**(1), 39–56 (2013)

Data	t	ROR	PRR	RFET	IC	GPS	TimeIC	LRC-SD	LRC-R
1	0	0.984	0.815	0.835	0.941	**0.989**	0.922	0.988	0.843
	7	0.988	0.817	0.859	0.942	0.990	0.927	**0.993**	0.845
	30	0.990	0.820	0.917	0.942	**0.992**	0.937	**0.992**	0.868
	365	0.591	0.592	0.773	0.737	**0.790**	0.726	0.760	0.742
2	0	0.766	0.743	0.745	0.796	0.765	0.789	**0.825**	0.726
	7	0.815	0.759	0.823	0.867	0.815	0.861	**0.905**	0.746
	30	0.837	0.776	0.878	0.901	0.837	0.897	**0.945**	0.746
	365	0.666	0.667	0.841	0.859	0.731	0.891	**0.903**	0.662
3	0	0.763	0.766	0.749	0.784	0.763	0.814	**0.875**	0.806
	7	0.803	0.782	0.799	0.837	0.803	0.871	**0.918**	0.792
	30	0.768	0.748	0.805	0.819	0.789	0.850	**0.896**	0.760
	365	0.622	0.601	0.742	0.786	0.724	0.839	**0.879**	0.726
4	0	0.794	0.746	0.710	0.760	0.793	0.757	**0.795**	0.742
	7	0.838	0.808	0.800	0.844	0.868	0.841	**0.885**	0.776
	30	0.819	0.791	0.789	0.820	0.818	0.847	**0.892**	0.693
	365	0.747	0.716	0.795	0.844	0.778	0.844	**0.885**	0.683
5	0	0.901	0.835	0.816	0.871	0.880	0.861	**0.905**	0.846
	7	0.921	0.827	0.827	0.892	0.898	0.882	**0.926**	0.825
	30	0.842	0.818	0.848	0.913	0.918	0.902	**0.949**	0.834
	365	0.723	0.673	0.859	0.915	0.772	0.911	**0.958**	0.758
6	0	0.771	0.751	0.787	0.811	**0.823**	0.773	0.814	0.797
	7	0.840	0.814	0.876	0.905	0.902	0.866	**0.913**	0.768
	30	0.844	0.817	0.918	0.910	0.905	0.874	**0.921**	0.757
	365	0.716	0.654	0.900	0.909	0.870	0.905	**0.952**	0.718
7	0	0.729	0.677	0.673	0.767	0.812	0.786	0.827	**0.837**
	7	0.775	0.745	0.752	0.842	0.871	0.866	**0.910**	0.787
	30	0.853	0.818	0.792	0.907	0.925	0.938	**0.988**	0.779
	365	0.592	0.593	0.817	0.848	0.780	0.844	**0.887**	0.732
8	0	0.736	0.717	0.671	0.753	0.768	0.761	**0.802**	0.793
	7	0.847	0.744	0.794	0.907	0.889	0.932	**0.984**	0.791
	30	0.849	0.744	0.802	0.905	0.889	0.933	**0.985**	0.717
	365	0.632	0.632	0.799	0.838	0.766	0.869	**0.912**	0.725
9	0	0.755	0.666	0.790	0.822	0.849	0.841	**0.906**	0.833
	7	0.725	0.636	0.788	0.823	0.849	0.839	**0.906**	0.792
	30	0.774	0.648	0.820	0.846	0.871	0.866	**0.933**	0.805
	365	0.560	0.592	0.873	0.852	0.749	0.871	**0.941**	0.823
10	0	0.745	0.729	0.693	0.761	0.760	0.756	0.792	**0.797**
	7	0.824	0.782	0.784	0.867	0.823	0.864	**0.907**	0.767
	30	0.779	0.742	0.771	0.834	0.798	0.828	**0.871**	0.714
	365	0.703	0.680	0.825	0.888	0.815	0.888	**0.932**	0.763
11	0	0.744	0.661	0.652	0.729	0.776	0.727	0.764	**0.792**
	7	0.788	0.747	0.817	0.902	0.915	0.902	**0.949**	0.838
	30	0.767	0.727	0.834	0.865	0.883	0.863	**0.909**	0.793
	365	0.666	0.633	0.729	0.770	0.732	0.834	**0.876**	0.686
12	0	0.801	0.770	0.776	0.831	0.843	0.816	0.859	**0.886**
	7	0.864	0.826	0.854	0.929	0.912	0.896	**0.945**	0.830
	30	0.833	0.816	0.902	0.890	0.858	0.900	**0.949**	0.828
	365	0.596	0.597	0.880	0.874	0.782	0.864	**0.906**	0.828
13	0	0.760	0.742	0.684	0.701	0.761	0.687	0.715	**0.897**
	7	0.874	0.875	0.806	0.850	**0.905**	0.845	0.886	0.745
	30	0.875	0.846	0.815	0.850	**0.905**	0.846	0.887	0.730
	365	0.707	0.655	0.851	0.832	0.812	0.824	**0.864**	0.685
14	0	0.778	0.766	0.777	0.825	0.823	0.804	0.846	**0.871**
	7	0.816	0.767	0.809	0.876	0.851	0.851	**0.896**	0.803
	30	0.770	0.720	0.840	0.883	0.855	0.858	**0.904**	0.787
	365	0.650	0.611	0.929	0.917	0.765	0.934	**0.982**	0.757
15	0	0.824	0.802	0.865	**0.943**	0.883	0.884	0.931	0.838
	7	0.851	0.827	0.917	0.932	0.924	0.927	**0.977**	0.858
	30	0.831	0.831	0.918	0.921	0.902	0.917	**0.966**	0.870
	365	0.771	0.748	0.943	0.906	0.783	0.917	**0.963**	0.788

Table 3 AUC measures for fifteen OSIM datasets, with four different surveillance period lengths, and various ADR mining methods. For each dataset and t-value, the highest AUCs are marked in bold.

	ROR	PRR	RFET	GPS	IC	TimeIC	LRC-R
PRR	1.5×10^{-9}	-	-	-	-	-	-
RFET	0.07587	3.0×10^{-7}	-	-	-	-	-
GPS	5.1×10^{-10}	$< 2 \times 10^{-16}$	0.04359	-	-	-	-
IC	3.9×10^{-08}	3.5×10^{-16}	8.3×10^{-12}	0.07338	-	-	-
TimeIC	1.9×10^{-07}	7.9×10^{-15}	8.1×10^{-10}	0.07587	1.00000	-	-
LRC-R	1.00000	0.00017	0.06192	2.3×10^{-7}	5.7×10^{-9}	1.6×10^{-8}	-
LRC-SD	1.1×10^{-13}	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	3.2×10^{-10}	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	1.6×10^{-15}

Table 4 p-values for pairwise t-tests between various method AUCs.