Adverse Drug Effect Detection

Lian Duan, Mohammad Khoshneshin, W. Nick Street, and Mei Liu

***Abstract*—Large collections of electronic patient records provide abundant but under-explored information on the real-world use of medicines. Although they are maintained for patient administration, they provide a broad range of clinical information for data analysis. One growing interest is drug safety signal detection from these longitudinal observational data. In this paper, we proposed two novel algorithms—a likelihood ratio model and a Bayesian network model—for adverse drug effect discovery. Although the performance of these two algorithms is comparable to the state-of-the-art algorithm, Bayesian confidence propagation neural network, the combination of three works better due to their diversity in solutions. Since the actual adverse drug effects on a given dataset cannot be absolutely determined, we make use of the simulated OMOP dataset constructed with the predefined adverse drug effects to evaluate our methods. Experimental results show the usefulness of the proposed pattern discovery method on the simulated OMOP dataset by improving the standard baseline algorithm—chi-square—by 23.83%.**

***Index Terms*—adverse drug effect, correlation, BCPNN, likelihood ratio, Bayesian network.**

I. Introduction

Drug safety is a major public health concern in the world. Though America's drug-approval process has the most world-renowned rigorous standards on safety and effectiveness, it cannot possibly uncover everything about a drug's performance that may occur even with pre-market clinical trials involving thousands of people. One important issue related to drug safety is how to detect adverse drug reactions. Adverse drug reactions are defined as those unintended and undesired responses to drugs beyond their anticipated therapeutic effects during clinical use at normal doses [13]. In the US, people spend billions of dollars on prescription drugs each year. Most of them are safely used. However, the few adverse drug reactions can cause serious healthcare and financial burdens. It is estimated that 6-7% of hospitalized patients experience severe adverse drug reactions each year with a potential of 100,000 deaths, which makes it the fourth largest cause of death in the US [8]. In addition, adverse drug reactions require extra treatment and prolong hospitalization, which causes a big financial problem. Therefore, effective methods for determining the relationship between pharmaceutical drugs and conditions (potential adverse events) are highly in need.

Currently, spontaneous adverse event reporting systems record every specific self-report of a suspected causal association between a drug and an adverse event for post-marketing safety detection. For observational analysis, methods need to provide useful information about associations between drugs and outcomes across a population of interest. It is unnecessary to ascertain whether a specific person had a particular outcome due to a particular drug, but instead we need to infer whether a population of individuals exposed to a drug experiences more of the outcome than expected. This population-based approach differs from the spontaneous adverse event reporting systems currently used. Many methods have been developed for post-marketing signal detection such as Bayesian confidence propagation neural network (BCPNN) [1], χ^2-statistics, and proportional reporting ratios (PRR) [5]. These methods assign each drug and adverse event pair a score. If one pair has a high score and is not confirmed before, it is the promising hypothesis to check.

However, due to poorly characterized data, insufficiently recorded clinical observations, and confounding effects, the true causal relationships between drugs and adverse events cannot be absolutely detected. In order to test the performance of different methods, the true causal relationships are required. Because of these above issues, the Observational Medical Outcomes Partnership (OMOP) [11] designed and developed a procedure to construct simulated dataset for method evaluations. The procedure generates fictional persons with fictional drug exposure and fictional adverse event occurrences with predefined association between fictional drugs and fictional outcomes. Though the dataset is contrived, to the best of our knowledge, it is the best simulated dataset and close to real observational data.

The goal of this paper is to develop methods to identify the associations that OMOP predefined to simulate data from the observational simulated dataset. Section 2 introduces the simulation procedure of OMOP data. Different signal detection models are discussed in Section 3. Experiments are conducted and the characteristics of different models are analyzed in Section 4. Finally, we draw a conclusion in Section 5.

II. Data Simulation

The OMOP simulation is a project funded by Foundation for the National Institutes of Health. It involves pharmaceutical industry, academic institutions, non-profit organizations, the Food and Drug Administration (FDA), and other federal agencies. The whole simulation procedure is complicated and details can be found at [11]. There has been some existing research [6], [14] conducted on this dataset. In the following, we provide a brief introduction of the simulation procedure. The simulated dataset contains 10 million persons, 90 million drug exposures from 5000 different drugs and 300 million condition occurrences from 4500 different conditions over a span of 10 years. For only 1.8% of the 20 million possible drug-condition combinations, there exists a true causal association between the drug and the condition. For the remaining combinations, no causal association exists.

The nature of the temporal relation between drugs and outcomes can vary. For example, many site reactions from vaccinations or biological injections can be observed within one day of exposure. Other outcomes can only be observed after many years due to slow changes in biology or the need for cumulative dose. Some drugs have increased fracture risks from years of exposure due to gradual bone loss from calcium malabsorption. For most outcomes, the temporal relationship between the drug and outcome is not clear. The occurrence of the outcome can appear anytime after exposure. Insidious outcomes are also common for rare and serious events because of the small number of observed cases, which makes it difficult to infer the temporal relationship.

Due to the above complicated temporal relationship of a true causal association between the drug and the condition, OMOP categories associations into constant risk onset or constant rate onset types. Constant risk onset types have a 50% chance to be acute, 40% chance to be insidious, and 10% chance to be delayed. Constant rate onset types have a 90% chance to be insidious, and 10% chance to be delayed. For the acute type, the outcome appears within the first week after drug exposure. For the insidious type, outcome appears at any time after the drug exposure. For the delayed type, outcome appears between one year and ten years after drug exposure.

III. Signal Detection Models

We adopted an ensemble of three different methods, two disproportionality analysis techniques and a Bayesian network model, to discover the association between drugs and potential adverse events (conditions). The disproportionality analysis methods calculate the pair correlations by comparing the expected and observed co-occurrences of a drug and a condition. We explored multiple counting methods and correlation measures and chose two that were both accurate and complementary. The Bayesian network model estimates the pair risk factors by using a Bayesian network. Finally a weighted combination of the raw scores from the three models was computed for each drug-condition pair to give the final ranking of possible associations.

A. Disproportionality Analysis

In order to apply disproportionality analysis, we need to generate a good two-by-two contingency table from the raw data first, and then use correlation measures to infer the ranking of possible associations.

1) Two-by-two Contingency Table: We generated the two-by-two contingency table based on the Modified SRSs (Spontaneous Report Systems) method [12]. The foundation of correlation measures was a collection of 2-dimensional tables of the form in Table I.

		Condition C1		
		Yes	No	
Drug D1	Yes	a	b	a+b
	No	c	d	c+d
		a+c	b+d	a+b+c+d=n

TABLE I
Two-by-two Contingency Table

The counts for drugs and conditions were given different weights, based on empirical observation. For example, conditions and drugs that start on the same day have an unclear causal relationship due to the characteristics of the simulated dataset. We don't know whether the condition is caused by the drug or the drug is dispensed for the condition. Therefore, the weight of conditions and drugs on the same day was set to be small. However, if time is more precisely recorded and we know whether a given drug is used before a given condition or not, we might set a large weight for the condition after drug exposure, and 0 or negative weight for the condition before drug exposure. In this paper, we designed a weighting scheme according to the OMOP data generation mechanism in Section 2. Such weighting scheme needs to be changed according to the empirical observation of given datasets. A condition occurring on the first day of drug usage was assigned weight w_1. A condition occurring during a drug period of less than 7 days was given weight w_2. A condition occurring

within 7 days of the beginning of the drug use but outside the drug period was given weight w_3. If the condition happened within 7 days of the beginning of the drug use and the drug period is greater than 7, we still use weight w_3. If the condition occurred later than 7 days from the beginning of the drug use and during the drug period, we use weight w_4. w_1, w_3, and w_4 are constants. w_1 was set to be small. The drug could be used w_2 might be a fixed value C, or dynamically calculated as $(C - w_3) * (7 - DrugPeriod)/7 + w_3$, giving interactions within a shorter drug period a higher weight. Since conditions caused by drugs were assumed to be caused by a single drug, we modified the counts for conditions occurring during a period with multiple drugs. Specifically, if the condition is strongly correlated with one of the co-occurring drugs, we reduce the count value for the co-occurrence with all the other drugs being taken during that period.

2) BCPNN: Probability Ratio, $Pr = tp/ep$, is straightforward and means how many times the combination happens more than expected. However, the Probability Ratio is very volatile when the expected value is small, which makes it favor the rare combinations rather than significant trends in the data. In order to solve the problem, people use shrinkage [1], [3], [10] to regularize and reduce the volatility of a measure by trading a bias to no correlation for decreased variance. Specifically, we add a continuity correction number to both nominator and denominator. Suppose the continuity correction is cc, the formula of BCPNN is $BCPNN = ln(tp + cc)/(ep + cc)$. Normally, we set $cc = 0.5/n$; however, it could be any positive number. This shrinkage strength has been successfully applied to pattern discovery in the analysis of large collections of individual case safety reports. Noren et al. [10] used it for drug signal detection and claimed that it precludes highlighting any pattern based on less than three events but is still able to find strongly correlated rare patterns. From a frequency perspective, BCPNN is a conservative version of Probability Ratio, tending towards 0 for rare events and with better variance properties. As tp and ep increase, the impact of the shrinkage diminishes.

B. Likelihood Ratio

The Likelihood Ratio (LR) is similar to a statistical test based on the loglikelihood ratio described by Dunning [4]. The concept of a likelihood measure can be used to statistically test a given hypothesis, by applying the likelihood ratio test. Essentially, we take the ratio of the highest likelihood possible given our hypothesis to the likelihood of the best "explanation" overall. The greater the value of the ratio, the stronger our hypothesis will be.

Given the counts, we calculate the true probability of $D1$ and $C1$, $tp = a/n$, the probability of $D1$, $p_d = (a + b)/n$, the probability of $C1$, $p_c = (a + c)/n$, and the expected probability of $D1$ and $C1$, $ep = p_d \cdot p_c$. To apply the likelihood ratio test as a correlation measure, it is useful to consider the binomial distribution. This is a function of three variables: $Pr(p, k, n) \rightarrow [0 : 1]$. Given our assumption of independence of drug and outcome, we predict that each trial has a probability of success ep. Then the binomial likelihood of observing k out of n records is $Pr(ep, k, n)$. However, the best possible explanation of each trial probability is tp instead of ep. Therefore, we perform the Likelihood Ratio test, comparing the binomial likelihood of observing k out of n records under the assumption of independence with the best possible binomial explanation. Formally, the Likelihood Ratio in this case is $LikelihoodRatio(S) = Pr(tp, k, n)/Pr(ep, k, n)$.

1) Region Bias: In this section, we study the different upper bounds of LR and BCPNN to discuss the different region bias with respect to pair support. Given a pair S with the actual probability tp, we have $tp \leq p_c \leq 1$ and $tp \leq p_d \leq 1$. When p_c and p_d reach their lower bound tp and the drug always occur together with the condition, the expected probability ep reaches its lower bound tp^2.

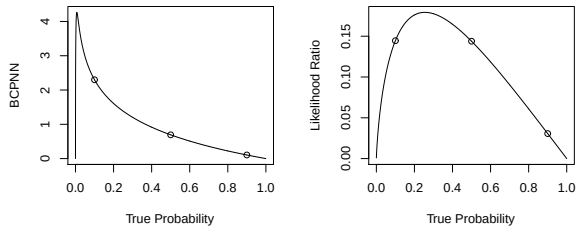
Lemma 1. *Both LR and BCPNN decreases with the increase of ep when tp remain unchanged.*

Proof: (1) When $tp > ep$,

$$\begin{aligned} & LikelihoodRatio(S) \\ = \; & n \cdot tp \cdot (ln(tp) - ln(ep)) \\ & +n \cdot (1 - tp) \cdot (ln(1 - tp) - ln(1 - ep)) \\ = \; & n \cdot tp \cdot ln(tp) - n \cdot tp \cdot ln(ep) \\ & +n \cdot ln(1 - tp) - n \cdot ln(1 - ep) \\ & -n \cdot tp \cdot ln(1 - tp) + n \cdot tp \cdot ln(1 - ep) \\ = \; & n \cdot tp \cdot ln\frac{tp}{1 - tp} + n \cdot ln(1 - tp) \\ & -n \cdot ln(1 - ep) + n \cdot tp \cdot ln\frac{1 - ep}{ep}. \end{aligned}$$

If we consider $LikelihoodRatio(S)$ as a function of ep, then

$$\begin{aligned} & LikelihoodRatio(S)' \\ = \; & \frac{n}{1 - ep} - \frac{n \cdot tp}{(1 - ep) \cdot ep} \\ = \; & \frac{n \cdot (ep - tp)}{(1 - ep) \cdot ep}. \end{aligned}$$



(a) BCPNN (b) LR

Fig. 1. Upper and lower bounds of BCPNN and LR

Since $tp > ep$, then $LikelihoodRatio(S)' < 0$. In other words, Likelihood Ratio decreases with the increase of ep when $tp > ep$. Similarly, when $tp < ep$, we can prove Likelihood Ratio decreases with the increase of ep. In all, Likelihood Ratio decreases with the increase of ep.

(2) When tp is fixed, BCPNN decreases with the increase of ep according to the formula. ■

According to Lemma 1, given the actual probability tp for a pair S, both LR and BCPNN reach their upper bounds when ep reaches the lower bound tp^2. We draw the upper bound curve of LR and BCPNN with respect to the change of tp in Figure 1. Though both LR and BCPNN reach their highest upper bound when tp is between 0 and 1, pairs on very low tp region have more chance to get higher BCPNN while pairs on relatively high tp region have more chance to get higher LR.

C. Bayesian Network Model

In this section we describe the proposed Bayesian network model for discovering adverse drug effects. Bayesain networks [7] are directed graphical models [9] which are useful in representing complex probability distributions. In Bayesian networks, random variables are shown with circles while dependencies are represented with directed edges. The joint probability over variables is given by a product over conditional probability of each variable given its parents:

$$P(x_1, x_2, ..., x_n) = \prod_i P(x_i|Pa(x_i))$$

where Pa denotes the parents of a random variable.

Figure 2 represents the the Bayesian network in plate notation. Each plate denotes an enumeration over the random variables. Here we have four distinct enumerations; N_I individuals, N_D drugs, N_C conditions, and T_i intervals for individual i. Intervals are defined based on the change in the status of individuals depending on the drugs they use. That is, upon each drug use, a new interval begins. By the end of the time window of the used drug, another interval starts. Therefore, each interval is associated with a specific number of drugs. Note that it is possible that an interval is not associated with any drugs.

In Figure 2, two different modeling paradigms are presented with regard to time. In the time-independent model, we assume that the length of the interval does not influence the number of occurred conditions. In the time-dependent model, increasing the length of an interval, conditions are more likely to happen.

Gray circles denote observed random variables while white circles represent hidden random variables. Generally, we observe x_{cit}: the number of occurrences of condition c in interval t for every individual i, and y_{dit} is 1 if the interval t for individual i is within the effective drug d's side-effect time window and 0 otherwise. Additionally, for the time-dependent case, we know δ_{it} which is the length of the interval t for individual i.

There is a deterministic relation between observed variable x_{cit} and hidden variables z_{cit} and z_{dcit} as follows:

$$x_{cit} = z_{cit} + \sum_d y_{dit} z_{dcit} \tag{1}$$

where z_{cit} is the number of conditions happened not because of any drug in interval t for individual i, and z_{dcit} is the number of conditions occurred caused by drug d in interval t for individual i.

The probability distribution of the hidden variable z_{cit} given its parents follows a Poisson distribution:

$$P(z_{cit} = k|\alpha_c) = \frac{e^{-\alpha_c} \cdot (\alpha_c)^k}{k!} \tag{2}$$

and the probability of α_c (for each condition c) given its parents ρ_1 and ρ_2 follows a gamma distribution:

$$P(\alpha_c|\rho_1, \rho_2) = \frac{\rho_2^{\rho_1}}{\Gamma(\rho_1)} \alpha_c^{\rho_1 - 1} e^{-\rho_2 \alpha_c} \tag{3}$$

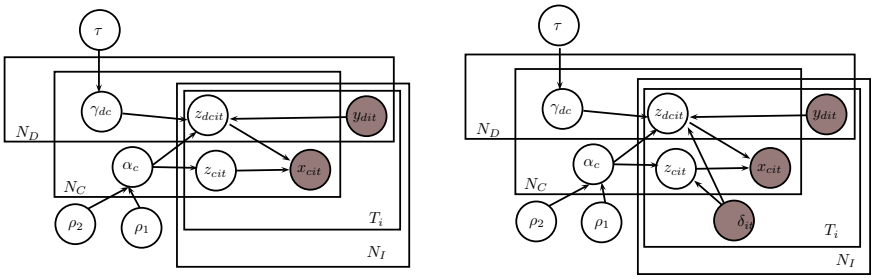
For the time-dependent case, Formula 2 can be written as follows:

$$P(z_{cit} = k|\alpha_c, \delta_{cit}) = \frac{e^{-\alpha_c \delta_{cit}} \cdot (\alpha_c \delta_{cit})^k}{k!} \tag{4}$$

The probability distribution of the hidden variable z_{dcit} given its parents follows a Poisson distribution:

$$P(z_{dcit} = k|\alpha_c, \gamma_{dc}) = \frac{e^{-\gamma_{dc}\alpha_c} \cdot (\gamma_{dc}\alpha_c)^k}{k!}. \tag{5}$$

More accurately, α_c is the background prevalence of



(a) time-independent model (b) time-dependent model

Fig. 2. The graphical model of the Bayesian network model. Gray circles denote observed variables while white circles represent hidden variables.

condition c while γ_{dc} is the impact of drug d on background prevalence. If $\gamma_{dc} > 1$, drug may be considered responsible. The probability of γ_{dc} given its parent τ follows a gamma distribution with equal parameters:

$$P(\gamma_{dc}|\tau) = \frac{\tau^{\tau}}{\Gamma(\tau)}\gamma_{dc}^{\tau-1}e^{-\tau\gamma_{dc}} \quad (6)$$

We considered gamma distribution with equal parameters (causes the prior mean of one), since we believe that the mean of drug effects is one apriori. For the time-dependent case, Formula 5 can be written as follows:

$$P(z_{dcit} = k|\alpha_c, \gamma_{dc}) = \frac{e^{-\delta_{cit}\gamma_{dc}\alpha_c} \cdot (\delta_{cit}\gamma_{dc}\alpha_c)^k}{k!}. \quad (7)$$

Given the description of the Baysian network, we are interested to infer the value of query variables, γ_{dc} given the observed variables. We assume prior parameters τ, ρ_1 and ρ_2 are given as algorithmic inputs. The remained hidden variables are z_{dcit}, z_{cit}, γ_{dc} and α_c. To infer the values of query variables, we use expectation-maximization (EM)) algorithm [2]. EM algorithm is way to learn parameters when there are some missing data. Here, we consider hidden variables z_{dcit} and z_{cit} as missing data, and γ_{dc} and α_c as distribution parameters. EM algorithm is an iterative algorithm where in the expectation step, the expectation of missing values given the last estimation of parameters is computed. In the maximization step, the parameters are maximized given the estimation of missing values.

In the M-step of our algorithm, we maximize the parameter α_c by maximizing the log likelihood function as follows:

$$\alpha_c = \frac{\sum_{it} z_{cit} + \rho_1 - 1}{\sum_i T_i + \rho_2}. \quad (8)$$

Then given the updated value of α_c, the value of γ_{dc} is updated as follows:

$$\gamma_{dc} = \frac{\sum_{it} z_{dcit} + \tau}{\alpha_c \sum_{it} y_{dit} + \tau}. \quad (9)$$

In the E-step of the algorithm, we compute the expectation of hidden variables z given the parameter α and γ. For solving this problem we used a lemma regarding the Poisson distribution. If Poisson distributions with rates $\lambda_1, \lambda_2, ..., \lambda_n$ created events when only the sum of the events N is observed, then the probability of the number of events follows a multinomial distribution with parameter $\frac{\lambda_j}{\sum_i \lambda_i}$ for the jth Poisson distribution. Therefore, The expectation value of z_{cit} is as follows:

$$z_{cit} = \frac{1}{1 + \sum_{d'} y_{d'it}\gamma_{d'c}}, \quad (10)$$

and similarly the the expectation of z_{dcit} is given by

$$z_{dcit} = \frac{y_{dit}\gamma_{dc}}{1 + \sum_{d'} y_{d'it}\gamma_{d'c}}. \quad (11)$$

Deriving the relevant equations for the time-dependent case is straightforward as we add δ_{it} to the related Poisson distribution in above equations. Finally, since we do not know that which scenario—time-dependent or time-independent—holds, we run the algorithm based on both and average the result:

$$\gamma_{dc} = \psi\gamma_{dc}^{(dep)} + (1 - \psi)\gamma_{dc}^{(ind)} \quad (12)$$

where $0 \leq \psi \leq 1$ is the mixing weight, $\gamma_{dc}^{(dep)}$ is the drug-condition pair for the time-dependent case, and $\gamma_{dc}^{(ind)}$ is the drug-condition pair for the time-independent case.

In comparison to LR and BCPNN which work with contingency table of each drug-condition pair, the Bayesian network model (BN) aims intervals. This capability helps resolve the confusion between different drugs as well as background prevalence. LR and BCPNN only

consider the confusion between a specific drug and a condition. Therefore, BN introduces another level of decomposition and more diversity as a result. This diversity causes the ensemble of these algorithms perform well.

IV. Experiment

The average precision (AP), a commonly-used metric in the field of information retrieval, is used to evaluate each method. It measures how well a system ranks items, and emphasizes ranking true positive items higher. Let y_{dc} is equal to 1 if the dth drug causes the cth condition, and 0 otherwise. Let $M = \sum_{d,c} y_{dc}$ denote the number of causal combinations and $N = D \times C$ the total number of combinations. Let z_{dc} denote the estimated value for the dth drug causing the cth condition. For a given set of estimated values $\overrightarrow{z} = (z_{11}, ..., z_{DC})$, we define "precision-at-K" denoted $P^K(\overrightarrow{z})$ as the fraction of causal combinations among the K largest predicted values in $\overrightarrow{z}$. Specifically, let $z_1 > ... > z_N$ denote the ordered value of $\overrightarrow{z}$. Then, $P^K(\overrightarrow{z}) = \frac{1}{K}\sum_{i=1}^{K} y_i$, where y_i is the true status of combination corresponding to z_i. The AP is calculated as $\frac{1}{M}\sum_{K=1}^{N}(P^K(\overrightarrow{z}) \cdot y_K)$. The AP is very similar to the area under the precision-recall curve, which penalizes both type of misclassification: identifying a correlation when no relationship exists (false positive) and failing to identify true correlations (false negative).

For only a small subset of the 20 million possible drug-condition combinations in the dataset, there exists a true causal association between the drug and the condition. However, OMOP does not provide all the ground truth. Instead, they provides a testing set with 4000 true associations and 4000 false associations. Therefore, it is impossible to calculate the true AP score across the whole dataset as OMOP does. We used the following bootstrapping method to mimic the AP score from OMOP. Given a ranking list, we randomly select 3000 true association and 3000 false association from the testing set to calculate the mimic AP score. Since the true associations and the false associations are unbalanced in the whole dataset, but balanced in the test data, we treat each false association in the ranking list of the testing data as sixty false associations in that ranking list, and then calculate the AP score for the transformed list. We conduct this procedure 100 times and calculate the mean of the AP scores as the mimic AP score. We tried different methods during the competition period and compared the mimic AP score with the true AP score from OMOP. There is a roughly linear relationship between these two scores shown in Figure 3.

For the two-by-two contingency table calculation, we get the best result when $w_1 = 1$, $w_2 = 5$, $w_3 = 2$, $w_4 = 1$. We also plot the occurrence of top-1000 pairs got from different methods in Figure 4. For Likelihood Ratio, the occurrence of pairs need to be large in order to get larger value. For BCPNN, the occurrence of high ranking pairs is very small. The shape of Bayesian network method is similar to that of Likelihood Ratio; however, high ranking pairs of Bayesian network model has relatively lower bias to large occurrence than those of Likelihood Ratio. Since some frequently occurred conditions are background noise and the Bayesian network model can remove such noise, frequently occurred pairs might not get high Poisson score and the tailor part of Bayesian network model is more dispersed than that of Likelihood Ratio. Since the three models work on different aspects of data, we simply add raw values of these three methods for ranking, and it generates the best AP score. The mimic AP score of different methods is shown in Table II. Likelihood Ratio, BCPNN, and the Bayesian network model achieve the similar AP score. They performance much better than the traditional χ^2 method. The AP score of the ensemble method has roughly 5% increase from three separate models. The true AP score for our ensemble method is 0.2569 which lands us on the third place on the OMOP competition.

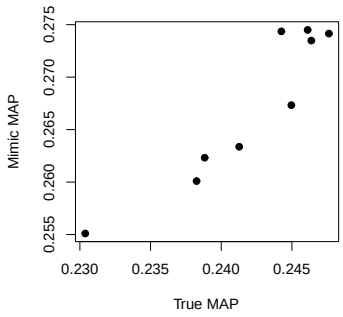
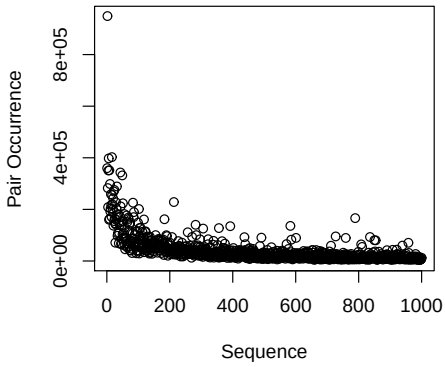


Fig. 3. Mimic AP score vs. the true AP score

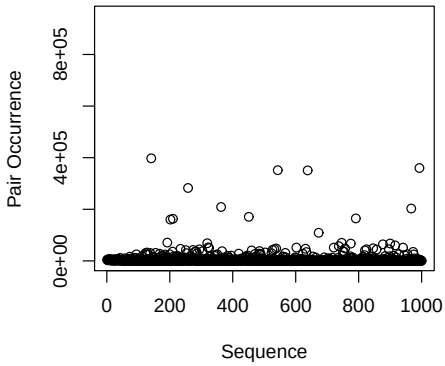
V. Conclusion

In the practical environment, there are two advantages of using our method. First, it helps to propose promising hypotheses to test. Most existing post-market drug surveillance methods require statistical analysis of the voluntarily submitted adverse event reports to filter out the many false positives. For example, 673,259 records were submitted to the US FDA Adverse Event Reporting System in 2010[1]. Although the spontaneous reports have

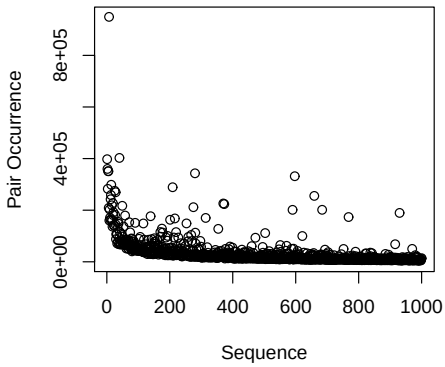
[1]Data from http://www.fda.gov/



(a) LR



(b) BCPNN



(c) Bayesian network model

Fig. 4. Occurrence of top-1000 pairs

Method	Mimic AP	% over chi-square
PRR	0.1084	-51.76%
chi-square	0.2247	0%
BCPNN	0.2662	18.43%
LR	0.2649	17.87%
BN	0.2670	18.81%
Ensemble	0.2783	23.83%

TABLE II
MIMIC AP SCORE OF DIFFERENT METHODS. % OVER CHI-SQUARE DENOTES THE IMPROVEMENT OVER THE STANDARD BASELINE ALGORIHTM—CHI-SQUARE

provided valuable information for clinical decisions, the spontaneous adverse event reporting system has a serious limitation of under-reporting. Therefore, we want to explore alternative data sources for adverse drug reaction signal discovery such as electronic medical records. An objective of the OMOP initiative is to provide simulated data that resemble electronic medical record for method development and evaluation. Using the simulated data, our method was shown to be able to return a reliable rank list of possible associations. If there are some high ranking associations that are not confirmed by pre-market clinical trials, we need to pursue further randomized control studies on them. Second, our method helps to identify some regional associations. Associations introduced in the textbook are general. However, the association ranking list generated by our method can be from the regional data we have. Some prevalent associations in the textbook might not be true for a specific region, while the other associations might be true for a specific region but not true in the national level because of the race or environmental difference.

In order to use our method, we need to make use of the hospital electronic health records, set up the algorithm to find the ranking list from the data, and check the difference between the ranking list and our existing knowledge. Nowadays, many hospitals have already digitalized their health records, and have staff to generate adverse event reports. Therefore, hiring a data mining technical staff is the overhead cost of deploying our system, which roughly costs $150,000 each year according to the salary information from indeed.com. In the national level, our method can help to identify the promising unknown adverse drug reactions to allocate our limited funds to set up the controlled test. The newly identified adverse drug reactions will change the guideline for drug prescription. In the regional (hospital) level, our method can help to identify the suspicious adverse drug reactions, which can guide doctors to

use the related drugs in a conservative way. By using our method, we can reduce the number of unnecessary treatments and even deaths caused by the related adverse drug reactions, which is hard to be measured financially. Apparently, the gain of using our method can easily surpass the overhead cost of deploying our system.

In this paper, likelihood ratio, BCPNN, and the Bayesian network model are introduced for drug safety signal detection. Evaluation on different methods in the OMOP dataset indicates the ensemble model works better than each individual model. Among all the methods we tried for the OMOP competition, the best ensemble is from these three models which work well and have diversity with each other. In the future, we are going to investigate more sophisticated way to ensemble these three models.

REFERENCES

[1] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54(4):315–321, 1998.

[2] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[3] W. Dumouchel. Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician*, 53(3):177–202, 1999.

[4] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[5] S. Evans, P. Waller, and S. Davis. Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10(6):483–486, 2001.

[6] R. Harpaz, W. DuMouchel, N. Shah, D. Madigan, P. Ryan, and C. Friedman. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology and Therapeutics*, May 2012.

[7] D. Heckerman. A tutorial on learning with bayesian networks. *Innovations in Bayesian Networks*, pages 33–82, 2008.

[8] L. J, P. BH, and C. PN. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *The Journal of the American Medical Association*, 279(15):1200–1205, 1998.

[9] S. Lauritzen. *Graphical models*, volume 17. Oxford University Press, USA, 1996.

[10] G. N. Norén, A. Bate, J. Hopstadius, K. Star, and I. R. Edwards. Temporal pattern discovery for trends and transient effects: its application to patient records. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 963–971, New York, NY, USA, 2008. ACM.

[11] OMOP. Omop common data model specifications, http://omop.fnih.org/CDMandTerminologies, 2011.

[12] OMOP. Disproportionality analysis, http://omop.fnih.org/MethodsLibrary, 2011.

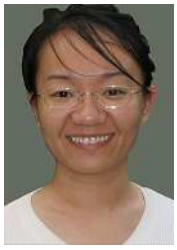
[13] M. Pirmohamed, A. M. Breckenridge, N. R. Kitteringham, and B. K. Park. Adverse drug reactions. *British Medical Journal*, 316(7140):1295C1298, 1998.

[14] M. J. Schuemie. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiology and Drug Safety*, 20(3):292–299, Mar. 2011.

Lian Duan received the Ph.D in Management Sciences from University of Iowa in 2012, and the Ph.D in computer science from Chinese Academy of Sciences in 2007. He is currently an assistant professor in the Department of Information Systems at New Jersey Institute of Technology. His main research interests include correlation analysis, community detection, recommender systems, and clustering.

Mohammad Khoshneshin received his PhD in Management Sciences from the University of Iowa in 2012. Currently, he is an Assistant Professor of Business Analytics in Bowling Green State University. His research interests includes machine learning and data mining, statistical relational learning, probabilistic graphical models, and approximate Bayesian inference and learning, especially variational inference and Markov chain Monte Carlo.

W. Nick Street received a Ph.D. in Computer Sciences from the University of Wisconsin-Madison in 1994. His current position is Professor, Departmental Executive Officer, and Henry B. Tippie Research Fellow in the Management Sciences Department at the University of Iowa. He also serves as director of the interdisciplinary graduate programs in Health Informatics and Information Science. His research interests are in machine learning and data mining, particularly the use of mathematical optimization in inductive learning techniques. His recent work has focused on personalized health care, correlation analysis, statistical relational learning, ensemble learning, and knowledge transfer. He has received an NSF CAREER award and an NIH INRSA postdoctoral fellowship. He is a member of IEEE, ACM, INFORMS and AAAI.



Mei Liu received her PhD from the University of Kansas in computer science in 2009. She completed her postdoctoral training in biomedical informatics as a National Library of Medicine fellow at Vanderbilt University in 2012. Currently, she is an assistant professor in the Department of Computer Science at New Jersey Institute of Technology. Her research interests include medical informatics, bioinformatics, data mining, and machine learning.