Improved Adverse Drug Event Prediction Through Information Component Guided Pharmacological Network Model (IC-PNM)

Xiangmin Ji, Lei Wang, Liyan Hua, Xueying Wang, Pengyue Zhang, Aditi Shendre, Weixing Feng, Jin Li, and Lang Li

Abstract—Improving adverse drug event (ADE) prediction is highly critical in pharmacovigilance research. We propose a novel information component guided pharmacological network model (IC-PNM) to predict drug-ADE signals. This new method combines the pharmacological network model and information component, a Bayes statistics method. We use 33,947 drug-ADE pairs from the FDA Adverse Event Reporting System (FAERS) 2010 data as the training data, and the new 21,065 drug-ADE pairs from FAERS 2011-2015 as the validations samples. The IC-PNM data analysis suggests that both large and small sample size drug-ADE pairs are needed in training the predictive model for its prediction performance to reach an area under the receiver operating characteristic curve $(\mathrm{AUROC}) = 0.82$. On the other hand, the IC-PNM prediction performance improved to $\mathrm{AUROC} = 0.91$ if we removed the small sample size drug-ADE pairs from the prediction model during validation.

Index Terms—Adverse drug event, information component, pharmacological network model, pharmacovigilance

1 Introduction

Improving adverse drug event (ADE) prediction has always been a primary focus of pharmacovigilance research. In the US ADE's account for about 3.5 million outpatient visits per year, with a third of these resulting in hospitalizations [1]. However, many ADEs cannot be detected during the pre-approval stages of a clinical trial. Clinical trials are also expensive and time consuming. Therefore, spontaneous reporting systems (SRS) are highly efficient approaches for collecting ADE cases during post-approval surveillance. The U.S Food and Drug Administration's (FDA) Adverse Event Reporting System (FAERS) [2] is one of the prominent SRS databases, which contains information on adverse drug events. This database was designed to support FDA's posting-marketing safety surveillance program for drug and biologic products. Drug-ADE signals can now be detected using the FAERS database by employing computational methods. However, not all drug-ADE signals detected through these methods are true positives. Some of the ADEs may be accidental or due to drugs other than those in the drug-ADE pairs. Therefore, it is critical to differentiate the true positive signals from the false positive ones [3].

Several computational approaches have been developed so far to detect drug-ADE signals from FAERS or similar SRS databases [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. Disproportionality analysis (DPA) is one of the notable computational methods, which compares the reported frequency to the baseline frequency under the assumption of no statistical association between a drug and an ADE, and is further designed to estimate the measure of drug-ADE association in the SRS. Proportional reporting ratio (PRR) [4], reporting odds ratio (ROR) [5] are frequentist DPA methods whereas Information component (IC) [6] and Empirical Bayesian Geometric Mean (EBGM) [7] belong to Bayesian DPA methods. Harpaz et al presented the performance of DPA, where PRR, ROR, and EBGM have similar performances, the AUROC being 0.71, 0.72, and 0.75, respectively [9]. In addition to ranking the drug-ADE signals, the DPA approach can also identify the true positive signals from the database based on the strength of their association. However, each DPA method is unique in ranking the drug-ADE associations, and the differences between them in ranking signals are well studied [10]. IC was proposed by Bate et al., and is also known as the Bayesian Confidence Propagation Neural Network (BCPNN) method. It measures the disproportional drug-ADE signals using a two-by-two table. Details of IC are further reviewed in section 2 below. The advantage of the IC approach is that it penalizes the drug-ADE signals when their sample sizes are small, because these small sample drug-ADE signals are prone to accidental findings.

- *X. Ji is with the College of Automation, Harbin Engineering University, Harbin 150001, China, and also with the Ordos Institute of Technology, Ordos 017000, China. E-mail: jixiangmin_hrbeu@foxmail.com.*
- *L. Wang is with the College of Automation, Harbin Engineering University, Harbin 150001, China, and also with the Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, OH 43210 USA. E-mail: lei.wang2@osumc.edu.*
- *L. Hua, X. Wang, W. Feng, and J. Li are with the College of Automation, Harbin Engineering University, Harbin 150001, China. E-mail: {hualiyan, fengweixing, lijin}@hrbeu.edu.cn, wangxueyinghrbeu@foxmail.com.*
- *P. Zhang, A. Shendre, and L. Li are with the Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, OH 43210 USA. E-mail: {pengyue.zhang, aditi.shendre, lang.li}@osumc.edu.*

Manuscript received 7 Nov. 2018; revised 5 May 2019; accepted 28 June 2019. Date of publication 20 Aug. 2019; date of current version 3 June 2021. (Corresponding authors: Jin Li and Lang Li.)
Digital Object Identifier no. 10.1109/TCBB.2019.2928305

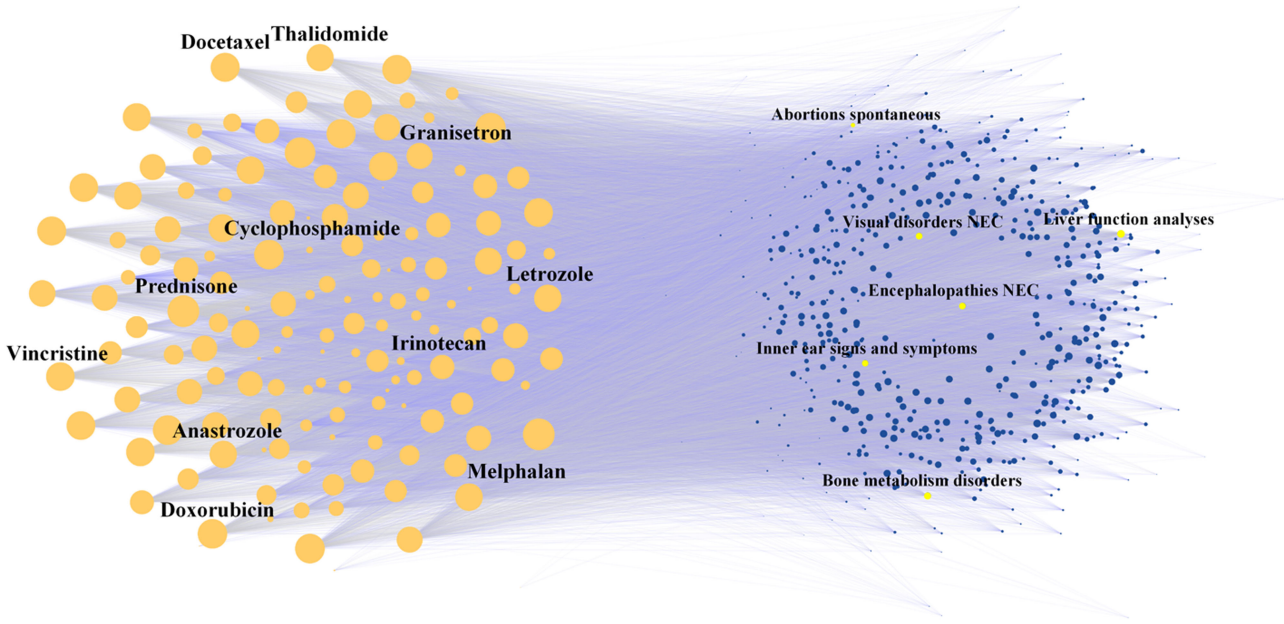


Fig. 1. Drug-ADE network produced with the software package Cytoscape (http://www.cytoscape.org). The nodes of drugs are on the left in yellow, and ADEs on the right in blue. The size of each node is proportional to its degree in the 2010 network. The edges of the 2010 network are shown in gray. The edges generated between 2011 and 2015 are shown in purple. Some drugs and ADEs have been labeled for illustration.

Currently, systems pharmacology-based methods are being developed to predict drug-related adverse events, to identify ADEs early on and reduce the drug-related morbidity and mortality [14]. Many systems pharmacology models have been proposed in predicting unknown ADEs [15], [16], [17], [18], [19], [20]. These systems pharmacology approaches use drug pharmacology features derived from chemical biology, biochemistry and structural biology data, obtained from various databases, and construct drug response predictive models using the network pharmacology framework [21], [22], [23]. Cheng et al developed a phenotypic network inference model for the prediction of ADEs using a database named MetaADEDB which includes 1,330 drugs, 13,200 ADEs, and more than 520,000 drug-ADE associations [24], and also developed a drug side effect similarity inference method (DSESI) for prediction of drug-target interaction using the same database [25]. These models were trained over known drug-ADE associations. For example, a pharmacological network model (PNM) integrated the network structure using three types of data, including safety data, taxonomic data and biological data, to predict the unknown ADEs using the Lexicomp database [23]. Details of PNM will also be reviewed in Section 2. However, all the existing systems pharmacology models treat the observed drug-ADE pairs as true signals and do not consider the frequency information reported in the current dataset. Moreover, none of these systems pharmacology models have investigated whether the sample size of the drug-ADE associations plays an important role in predicting true drug-ADEs associations.

In this paper, we propose a new approach, an information component guided pharmacology network model (IC-PNM) to determine drug-ADE associations. Our model is different from some of the previous one in that 1) it considers the frequency information between the drug and ADE from the FAERS data based on the PNM, 2) it combines the strengths of both the IC and PNM methods to optimize the drug-ADE prediction.

2 Methods

2.1 FAERS Data Processing

The FAERS database contains ADE reports submitted by either physicians, patients, or pharmaceutical companies to FDA. FAERS is updated quarterly. ADEs in FAERS are annotated using the Medical Dictionary for Regulatory Activities (MedDRA) [26]. In this paper, we investigated 238 cancer drugs. After integrating their Anatomical Therapeutic Chemical Classification System (ATC) codes and biochemical and biophysical drug properties from PubChem [27], 152 drugs had the full information, and were further investigated. On the other hand, 633 high level terms (HLTs) in MedDRA were chosen as ADEs. The FAERS data from 2010 to 2015 was used for data analysis. If the reports shared the same case identifiers (isr and primaryid), the latest reports were included in our dataset. The drug names were normalized using their Drug Bank IDs [28], and the preferred ADE terms (PTs) were mapped to their high-level terms (HLTs). FAERS 2010 data was selected as the training set, and had 33,947 distinctive drug-ADE combinations among 152 drugs and 633 ADEs. The validation set, i.e., FAERS data from 2011 to 2015, had 21,065 new drug-ADE combinations that were not part of the training set. In total, 55,012 drug-ADE combinations were used in building the predictive model. The ATC codes of 152 drugs from DrugBank, and 633 ADE terms from MedDRA are listed in Supplementary Table S1 and Table S2, respectively. The 17 biochemical features include molecular weight, XLogP3, hydrogen bond donor count, hydrogen bond acceptor count, rotatable bond count, exact mass, monoisotopic mass, topological polar surface area, heavy atom count, formal charge, complexity, isotope atom count, defined atom stereocenter count, undefined atom stereocenter count,

TABLE 1a
Definition of the Network Features

Feature Name	Feature Definition	Supplementary Information
degree-prod	$X_1(i,j) = degree(i) \times degree(j)$	
degree-sum	$X_2(i,j) = degree(i) + degree(j)$	
degree-ratio	$X_3(i,j) = degree(i)/degree(j)$	
degree-absdiff	$X_4(i,j) = degree(i) - degree(j)$	
jaccard-ADE-max	$X_5(i,j) = \max_{k \in N(i)-\{j\}} \{J(j,k)\}$	$J(j,k)$denotes jaccard coefficient between $N(j)$ and $N(k)$ $J(j,k) = \lvert N(j) \cap N(k) \rvert / \lvert N(j) \cup N(k) \rvert$
jaccard-ADE-KL	$X_6(i,j)$: Kullback–Leibler (KL) divergence between the distribution $D_{ADE}(i,j)$ of the variable $J(i,k), k \in N(j) - \{i\}$ and its reference distribution.	The reference distribution was computed as the mean of distributions $D_{ADE}(i,j)$ over the training edges (i, j)
jaccard-drug-max	$X_7(i,j) = \max_{k \in N(j)-\{i\}} \{J(i,k)\}$	
jaccard-drug-KL	$X_8(i,j)$: KL divergence between the distribution $D_{drug}(i,j)$ of the variable $J(j,k), k \in N(i) - \{j\}$ and its reference distribution.	The reference distribution was computed as the mean of distributions $D_{drug}(i,j)$ over the training edges (i, j)

TABLE 1b
Definition of the Taxonomic Features

Feature Name	Feature Definition	Supplementary Information
atc-min	$X_9(i,j) = \min_{k \in N(j)-\{i\}} \{d_{ATC}(i,k)\}$	
atc-KL	$X_{10}(i,j)$: KL divergence between the distribution $D_{ATC}(i,j)$ of the variable $D_{ATC}(i,k), k \in N(j) - \{i\}$ and its reference distribution.	The reference distribution was computed as the mean of distributions $D_{ATC}(i,j)$ over the training edges (i, j)
meddra-min	$X_{11}(i,j) = \min_{k \in N(i)-\{j\}} \{d_{MedDRA}(j,k)\}$	
meddra-KL	$X_{12}(i,j)$: KL divergence between the distribution $D_{Med\,\mathrm{DRA}}(i,j)$ of the variable $D_{MedDRA}(j,k), k \in N(i) - \{j\}$ and its reference distribution.	The reference distribution was computed as the mean of distributions $D_{Med\,\mathrm{DRA}}(i,j)$ over the training edges (i, j)

TABLE 1c
Definition of the Intrinsic Features

ICL Threshold in training data	The number of non-significant drug-ADE combinations filtered by ICL	The number of drug-ADE combinations to construct the PNM in training data	The number of drug-ADE combinations in validation data	AUROC of the IC-PNM i
Null	0	33,947	21,065	0.82
-4.79	3,395	30,552	21,065	0.75
-4.02	6,789	27,158	21,065	0.75
0.00	25,139	8,808	21,065	0.64

defined bond stereocenter count, undefined bond stereocenter count, and covalently-bonded unit count.

2.2 Model Speculation

2.2.1 *Pharmacological Network Model (PNM)*

A pharmacological network model (PNM) [23] is designed to predict new and unknown ADEs based on known drug-ADE signals. Using the training data, the network is constructed as follows: the nodes are drugs or ADEs, and edges are the known drug-ADE associations. The drug-ADE associations network is shown in Fig. 1. Using this network, eight network features are generated: degree-prod, degree-sum, degree-ratio, degree-absdiff, jaccard-ADE-max, jaccard-ADE-Kullback-Leiber(KL), jaccard-drug-max, and jaccard-drug-KL. In addition, there are four taxonomic features: atc-min, atc-KL, meddra-min, and meddra-KL; and two intrinsic features: Euclid-min and Euclid-KL. The definition of these three types of features are described in Tables 1a, 1b, and 1c respectively. Index i denotes the drug, and j denotes the ADE. $N(i)$ denotes the set of neighbors of node i, and $\mathrm{N}(j)$

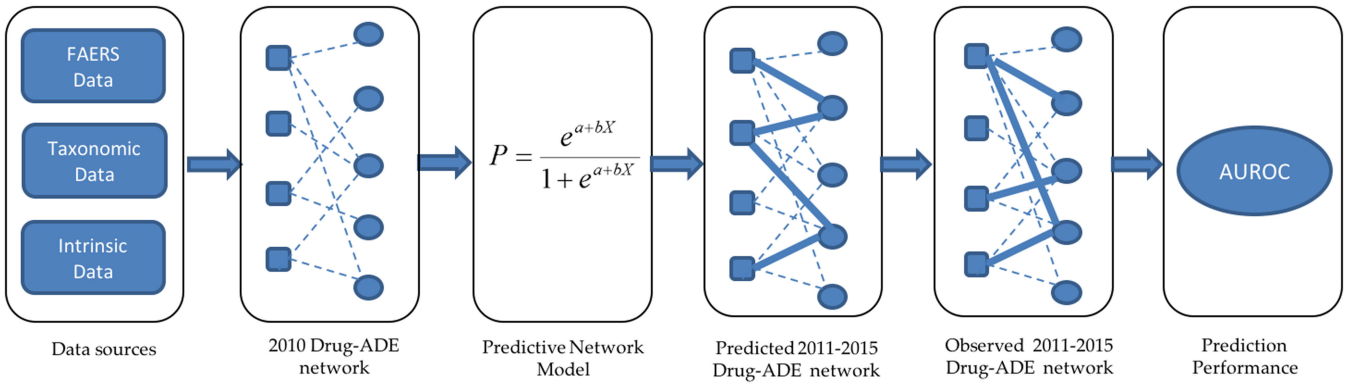


Fig. 2. Overview of the pharmacological network model (PNM). First, three types of data were integrated, including FAERS data, taxonomic data (ATC code and MedDRA code), and intrinsic data (biochemical features of the drugs). Second, the drug-ADE network was constructed based on the drug-ADE combinations from the 2010 data. Third, the network features, taxonomic features, and intrinsic features were generated based on the drug-ADE network built upon the training data. Next, a logistic regression (LR) model was trained using the training set (FAERS 2010 data). Finally, its prediction performance was evaluated using the new drug-ADEs reported in the validation data set, using 2011-2015 FAERS data.

denotes the set of neighbors of node j. d_{ATC} denotes the length of the shortest path between two drugs in the ATC taxonomy; d_{MedDRA} denotes the length of the shortest path between two ADEs in the MedDRA taxonomy, and d_{BIC} denotes the Euclidean distance between two drugs in the 17-dimensional biochemical feature space, subscript BIC is an abbreviation for the biochemical features.

Using these 14 network features, drug-ADE pairs are predicted using logistic regression. Denote Y as the outcome variable. In the training data, if a drug-ADE pair exist, then $\mathrm{Y} = 1$, otherwise $\mathrm{Y} = 0$. Let Y_{ij} be the outcome of drug i and ADE j, where $\mathrm{i} = 1, \ldots$ number of drugs, and $\mathrm{j} = 1, \ldots$ number of ADEs. In a logistic regression model, the probability of a drug-ADE pair being true is defined as in (1)

$$E(Y_{\mathrm{ij}}) = p_{ij} = exp\left(\sum_s q_s x_s(i,j)\right) \Big/ \left[1 + exp\left(\sum_S q_s x_s(i,j)\right)\right]. \quad (1)$$

Here, $q_s \mathrm{q_S}$ denotes the regression parameter and $x_s \mathrm{X_S}$ denotes the features. The model fit for the training data set is determined using the Akaike information criterion (AIC), where the optimal model has the lowest AIC.

Once we have the fully trained model, the probability of each drug-ADE pair in the validation set is predicted as in (2):

$$\mathrm{predict_{ij}} = 1 \Big/ \left[1 + exp\left(-\sum_s q_s x_s(i,j)\right)\right] \quad (2)$$

When $\mathrm{predict}_{ij}$ is greater than a specified threshold, the drug-ADE pair is predicted to be positive, otherwise, it is predicted to be negative. The prediction performance is evaluated using the new drug-ADE pairs reported in the validation data set. Using different prediction thresholds, an area under the receiver operating characteristic curve (AUROC) is calculated to evaluate the prediction performance. An overview of the PNM is illustrated in Fig. 2.

2.2.2 Information Component (IC)

Information component was proposed by Bate et al. [6], which measures the strength of association between a drug i and an ADE j. Let us define C_{ij} as the number of reports that contain a drug-ADE pair; let C_{i+} and $C_{+\mathrm{j}}$ be the number of reports containing drug i and ADE j respectively; and C_{++} be the total number of reports. IC is a Bayesian approach that assumes that C_{ij}, C_{i+} and $C_{+\mathrm{j}}$ follow binomial distributions [12], with the prior distributions of the parameters chosen to be beta and uniform distributions:

$$C_{ij} \sim Bin(C_{++}, p_{ij}) \text{ and } p_{ij} \sim Beta[1, (p_{i+} \times p_{+j})^{-1}]$$
$$C_{i+} \sim Bin(C_{++}, p_{i+}) \text{ and } p_{i+} \sim Uniform(0,1)$$
$$C_{+j} \sim Bin(C_{++}, p_{+j}) \text{ and } p_{+j} \sim Uniform(0,1).$$

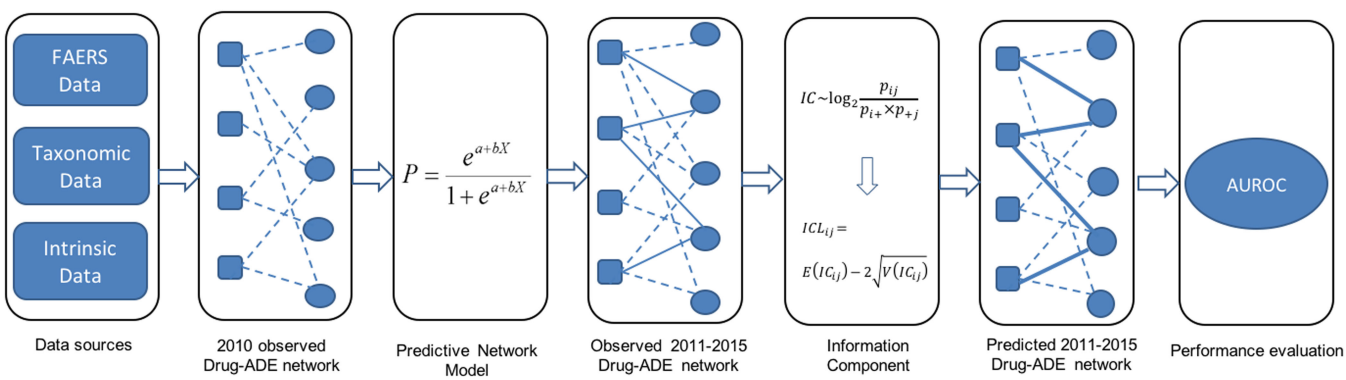


Fig. 3. Overview of an information component guided pharmacology network model (IC-PNM) during validation. The work flows can be referred to in the algorithm flow chart ii.

The IC of drug i and ADE j is defined as $IC \sim \log_2 \frac{p_{ij}}{p_{i+} \times p_{+j}}$. Its posterior expectation [8] is:

$$E(IC_{ij}) = \log_2 \left[\frac{(c_{ij}+1)(c_{++}+2)^2}{(c_{++}+2)^2 + c_{++}(c_{i+}+1)(c_{+j}+1)} \right]. \quad (3)$$

The variance of $E(IC_{ij})$ is obtained via the delta method:

$$V(IC_{ij}) = \frac{\frac{c_{++}-c_{ij}+\gamma-1}{(c_{ij}+1)(1+c_{++}+\gamma)} + \frac{c_{++}-c_{i+}+1}{(c_{i+}+1)(3+c_{++})} + \frac{c_{++}-c_{+j}+1}{(c_{+j}+1)(3+c_{++})}}{(\log 2)^2} \quad (4)$$

$$\gamma = \frac{(C_{++}+2)^2}{(C_{i+}+1)(C_{+j}+1)},$$

Finally, we use the approximated lower bound of IC's 95 percent confidence interval (5), denoted as ICL_{ij}, to rank drug-ADE combinations.

$$ICL_{\mathrm{ij}} = E(IC_{\mathrm{ij}}) - 2\sqrt{V(IC_{\mathrm{ij}})} \quad (5)$$

2.2.3 Information Component Guided Pharmacological Network Model (IC-PNM) for Drug-ADE Prediction Using the FAERS Data

PNM and IC are two strikingly different approaches in detecting drug-ADE associations. PNM utilize the existing drug-ADE association data to predict the future drug-ADE associations. PNM does not consider how frequently drug-ADE pairs are reported in the current data set, nor the effect size of the drug-ADE associations. On the other hand, the IC approach is designed to identify highly significant drug-ADE pairs. IC itself estimates the effect size of a drug-ADE association, and the standard deviation in the ICL is driven by the frequency of a drug-ADE association. Therefore, in the IC statistics, both the frequency and effect size of a drug-ADE association makes an enormous impact in determining the statistical significance of the drug-ADE associations. In this paper, we will determine whether using the IC and PNM together optimizes the predictive performance of the model. We have two situations in which the IC may improve the PNM prediction performance.

i. In the training dataset where, if we remove the low frequency or non-significant drug-ADE pairs determined by the IC, will IC-PNM has improved prediction performance?
ii. In the validation dataset where, if we remove the low frequency or non-significant drug-ADE pairs determined by the IC, will IC-PNM has improved prediction performance?

In order to answer these two different questions, two new IC-PNM algorithms are presented in the following flow charts i and ii. Moreover, the overview of IC-PNM is presented in Fig. 3 simultaneously, the descriptions of which are refered to in the algorithm flow chart ii.

3 Results

3.1 Training Data and Validation Data

A total of 152 drugs and 633 ADEs were used to determine the drug-ADE combinations in the training and validation datasets. The FAERS 2010 data was used as the training set and included 33,947 drug-ADE combinations. The frequency of drug-ADE combinations that were reported one, two, and three times was 5780, 3723, and 2608, respectively. These low frequency drug-ADE combinations accounted for 35.6 percent of all the combinations. The FAERS 2011-2015 data was used as the validation set and included a total of 21,065 new drug-ADE combinations, which were not part of the 2010 FAERS data. Among these 21,065 new drug-ADE combinations, the frequency of drug-ADE combinations reported one, two, and three times was 6547, 3430, and 2200, respectively. These low-frequency drug-ADE combinations accounted for 57.8 percent of all combinations in the validation data set.

Algorithm flow chart i. IC-PNM in training data

Input: a bipartite network G, nodes denote drugs or ADEs, edges denote known drug-ADE signals
Output: AUROC
Parameters: AIC, ICL_{ij}, threshold T, the number of cycles n

Construct a bipartite network G based on the known drug-ADE pairs which constitute the training set;
Calculate the covariates of each drug-ADE pair according to the definitions in the training data;
Calculate the ICL_{ij} of each drug-ADE pair in the training set;
Construct the validation set from the database, and calculate the ICL_{ij} of each drug-ADE pair in the validation set;
The logistic regression models of all possible combinations of covariates are fitted based on the training data;
Calculate the Akaike Information Criterion (AIC) for all possible models, select the model with the minimum AIC as the optimal model, here model parameters are obtained;
Calculate the AUROC of the validation set (**AUROC of PNM**) using the optimal model selected in the sixth step;
The drug-ADE pairs in the training set are ranked by the ICL_{ij}; using a threshold T, some drug-ADE pairs are removed from the training set, and a new training set is formed, sample size of which is smaller than before;
Construct a new bipartite network based on the new training data;
Repeat steps from 2 to 6, calculate the AUROC of the validation set (**AUROC of IC-PNM**) based on the new bipartite network Gi;
IF AUROC of IC-PNM > AUROC of PNM:
  AUROC = AUROC of IC-PNM;
End;
IF AUROC of IC-PNM <= AUROC of PNM:
  WHILE AUROC of IC-PNM <1:
    i = 1;
    Re-select the threshold T, repeat steps 9 and 10, and re-calculate the AUROC of IC-PNM;
    IF AUROC of IC-PNM > AUROC of PNM:
      AUROC = AUROC of IC-PNM;
      BREAK;
    End;
    IF i = n && AUROC of IC-PNM <= AUROC of PNM:
      AUROC = AUROC of PNM;
      BREAK;
    End;
    IF AUROC of IC-PNM <= AUROC of PNM:
      i = i + 1;
    End;
End;
RETURN AUROC;

Function AIC(k, L)

Input: the number of estimated parameters k, the likelihood of the LG model $\hat{L}$
Output: AIC score
1 AICscore = $2k - 2\ln \hat{L}$;
2 **RETURN** AIC score;
Where:
k is the number of estimated parameters;
L is the likelihood of LG model;
$\hat{L} = \log\left(\prod_{i=1}^{n} h_\theta(x)^{y(i)}\right)\ (1 - h_\theta(x)^{1-y(i)})$
$\hat{\mathrm{L}} = \log\left(\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{h}_\theta(\mathrm{x})^{\mathrm{y(i)}}(1 - \mathrm{h}_\theta(\mathrm{x}))^{1-\mathrm{y(i)}}\right)$;; $y(i)$ is the outcome variable of the LG model; $h_\theta\ (x) = \frac{1}{1+e^{-\theta Tx}}$, θ , θ is the estimated parameters;

Algorithm flow chart ii. IC-PNM in validation data

Input: a bipartite network G, nodes denote drugs or ADEs, edges denote known drug-ADE signals
Output: AUROC
Parameters: AIC, ICL_{ij}, threshold T, the number of cycles n
1 Construct a bipartite network G based on the known drug-ADE pairs which constitute the training set;
2 Calculate the covariates of each drug-ADE pair according to the definitions;
3 Calculate the ICL_{ij} of each drug-ADE pair in the training set;
4 Construct the validation set from the database, and calculate the ICL_{ij} of drug-ADE pair in the validation set;
5 The logistic regression models of all possible combinations of covariates are fitted based on the training data;
6 Calculate the Akaike Information Criterion (AIC) for all possible models, select the model with the minimum AIC as the optimal model, here parameters of the model q_s are obtained;
7 Calculate the AUROC of the validation set (**AUROC of PNM**) using the optimal model selected in the sixth step;
8 The drug-ADE pairs in the validation set are ranked by the ICL_{ij}; using a threshold T, some drug-ADE pairs are removed from the validation set, and then a new validation set is formed, sample size of which is small than before;
9 Calculate the AUROC of the new validation set (**AUROC of IC-PNM**) using the optimal model;
10 **IF** AUROC of IC-PNM > AUROC of PNM:
11 AUROC = AUROC of IC-PNM;
12 End
13 **IF** AUROC of IC-PNM <= AUROC of PNM:
14 **WHILE** AUROC of IC-PNM <1:
15 i = 1;
16 Re-select the threshold T, re-calculate the AUROC of IC-PNM;
17 **IF** AUROC of IC-PNM > AUROC of PNM:
18 AUROC = AUROC of IC-PNM;
19 **BREAK**;
20 End;
21 **IF** i = n && AUROC of IC-PNM <= AUROC of PNM:
22 AUROC = AUROC of PNM;
23 **BREAK**;
24 End;
25 **IF** AUROC of IC-PNM <= AUROC of PNM:
26 i = i + 1;
27 End;
28 End;
29 **RETURN** AUROC;

TABLE 2
Results of Features of the Multi-Variates Model

Feature Name	Regression Coefficients	Standard Bror	P-value
degree-prod	8.903e-05	6.850e-06	< 2e-16
degree-sum	−1.280e-03	1.032e-03	0.21513
degree-ratio	2.581 e-03	1.013e-03	0.01083
degree- absdiff	1.942e-03	6.382e-04	0.00234
jaccard-ADE-max	1.048e+01	2.094e-01	< 2e-16
jaccard-ADE-KL	−1.984e-01	1.615e-02	< 2e-16
jaccard-drug-max	8.073e+00	3.466e-01	< 2e-16
jaccard-drug-KL	−3.522e-01	3.095e-02	< 2e-16
atc-min	−8.341 e-02	1.654e-02	4.56e-07
atc-KL	−1.892e-01	2.839e-02	2.69e-11
meddra-min	−1.625e-02	1.714e-02	0.34322
meddra-KL	−5.152e-02	1.780e-02	0.00379
euclid-min	8.889e-05	6.949e-05	0.20082
euclid-KL	−1.308e-02	3.504e-03	0.00019

3.2 The Statistical Significance of Pharmacology Features in the PNM Training Model

Fourteen pharmacology and topological drug features were fitted into the multivariate logistic regression model. Their statistics are shown in Table 2. All but three features (i.e., degree-sum, meddra-min and euclid-min) were significant in predicting drug-ADE associations. In particular, features such as degree-prod, degree ratio, degree-absdiff, jaccard-ADE-max, and jaccard-drug-max, showed positive correlations. The other features, such as jaccard-ADE-KL, jaccard-drug-KL, atc-min, atc-KL, meddra-KL, and euclid-KL showed negative correlations.

3.3 IC-PNM Does Not Improve the Drug-ADE Prediction after Removing Non-Significant Drug-ADE Combinations in the Training Data

After applying IC to the training data, the 33,947 drug-ADE combinations were ranked using $ICLij$. We then investigated whether filtering out non-significant drug-ADE combinations through thresholding $ICLij$ can improve PNM's prediction performance.

Table 3 shows the prediction performance of our model under various conditions. If we include all the training data, i.e., no ICL filtering, in building up the PNM, its prediction performance for the drug-ADE combinations is $\mathrm{AUROC} = 0.82$, which is the performance of the existing PNM. If we set the ICL threshold at 0, 25,139 non-significant drug-ADE combinations are removed, and 8,808 drug-ADE combinations are left in the training data, the detailed description of which can be found in supplementary table S3. Using this reduced training data to build up the PNM, its prediction performance decreases to 0.64. Table 3 clearly illustrates that when ICL threshold increases, the more non-significant drug-ADE combinations are filtered out, and the less accurate the prediction performance of PNM model becomes.

3.4 IC-PNM Shows Higher Prediction Performance for Drug-ADE Combinations that Have Higher Statistical Significance Ranked by ICL

There were 21,065 new drug-ADE combinations in the validation set using the FAERS 2011-2015 data. IC was applied to the validation data, and 21,065 drug-ADE combinations were ranked based on their *ICLs*. Then we investigated

TABLE 3
The Prediction Performance of IC-PNM Algorithm I

ICL Threshold in training data	The number of non-significant drug-ADE combinations filtered by ICL	The number of drug-ADE combinations to construct the PNM in training data	The number of drug-ADE combinations in validation data	All ROC of the IC-PNM
Null	0	33,947	21,065	0.82
−4.79	3,395	30,552	21,065	0.75
−4.02	6,789	27,158	21,065	0.75
0.00	25,139	8,808	21,065	0.64

whether drug-ADE pairs with higher statistical significance can be predicted better by PNM than the other drug-ADE pairs. Table 4 presents our IC-PNM analysis results.

Table 4 shows that the prediction performance of the existing PNM is $\mathrm{AUROC} = 0.82$, when none of the new drug-ADE combinations are filtered. In comparison, when the ICL threshold changes from small to critical value 0, the AUROC of IC-PNM increases gradually. Specifically, when the filtering threshold of the ICL is set at 0.00, 19,329 out of 21,065 drug-ADE combinations are not statistically significant and filtered out, and the prediction performance of IC-PNM for the remaining 1,736 drug-ADE combinations becomes $\mathrm{AUROC} = 0.91$. The detailed description of the 1,736 new drug-ADE associations can be found in the supplementary table S4.

Table 4 demonstrates that as the ICL threshold increases, more non-significant drug-ADE combinations get filtered from the validation set, leading to an increase in the prediction performance of the IC-PNM.

4 Conclusion and Discussion

In this study, we propose a new method, named IC-PNM, that combines the strengths of network pharmacology and Bayesian statistics to determine drug-ADE associations. IC-PNM not only takes advantage of the network pharmacology features in predicting drug-ADE associations, but also penalizes small sample size drug-ADE pairs, removing false positive signals.

While building the training model, our IC-PNM analysis showed that the prediction performance of the model, with an $\mathrm{AUROC} = 0.82$, is highest when the low frequency and non-significant drug-ADE pairs are not removed from the training data set. We believe this is because the training model needs both positive and negative data. While the small sample size drug-ADE pairs have more negative data than large sample size drug-ADE pairs, they also contain positive data. Therefore, in training the PNM model, the training model should include both large and small sample size drug-ADE pairs when selecting the training data.

On the other hand, when validating the predictive performance of PNM, including small sample size drug-ADE pairs hurts the validation. This result clearly demonstrates that small sample size drug-ADE pairs contain more negative signals than the large sample size drug-ADE pairs. If we exclude drug-ADE pairs with ICL < 0, the IC-PNM prediction performance improves to $\mathrm{AUROC} = 0.91$. This improves upon the performance of the existing PNM, with an $\mathrm{AUROC} = 0.82$, where none of the drug-ADE combinations are excluded from the validation data.

In predicting the drug-ADE associations, we found that certain pharmacology and network features showed positive correlations, including degree-prod, degree ratio, degree-absdiff, jaccard-ADE-max, and jaccard-drug-max. Whereas, some other features, including jaccard-ADE-KL, jaccard-drug-KL, atc-min, atc-KL, meddra-KL, and euclid-KL showed negative correlations. The features in the PNM developed by Cami et al. [23] showed similar correlations, with degree-prod, degree-absdiff, jaccard-ADE-max, and jaccard-drug-max being positively correlated while degree-ratio, jaccard-ADE-KL, jaccard-drug-KL, atc-min, atc-KL, meddra-min, meddra-KL, Euclid-min and euclid-KL being negatively correlated with the outcome. The network feature, degree-ratio, was the only feature to have correlations that were opposite to each other in our PNM model and the one proposed by Cami et al.

In order to further investigate the connections between IC and PNM, significant amount of research needs to be

TABLE 4
The Prediction Performance of IC-PNM Algorithm ii

The number of drug-ADE combinations to construct the PNM in training data	ICL Threshold in the validation data	The number of new drug-ADE combinations filtered by ICL in validation data	The number of new drug-ADE combinations remained in validation data	AUROC of the IO PNM ii
33,947	null	0	21,065	0.82
33,947	−4.79	4,392	16,673	0.82
33,947	−4.02	8,214	12,851	0.83
33,947	−3.14	11,442	9,623	0.85
33,947	−2.78	12,640	8,425	0.85
33,947	−2,14	14,470	6,325	0.86
33,947	−1.33	16,960	4,105	0.86
33,947	0.00	19,329	1,736	0.91

conducted to determine how PNM can be used as a proper prior for IC. PNM can generate probabilities for any drug-ADE pair, but these probabilities have different interpretations from the population drug-ADE frequencieis estimated using SRS databases. To address this challenging issue, new integrated methods need to be developed in the future.

Acknowledgments

This work was supported in part by several National Natural Science Foundation of China (61773134, 61803117). Fundamental Research Funds for the Central Universities (HEUCFG201824, 3072019CFM0403) at Harbin Engineering University, the Natural Science Foundation of Heilongjiang Province of China (YQ2019F003), the US National Institutes of Health (NIH) grants (GM10448301-A1, LM011945, GM117206), and the US National Science Foundation (NSF) grant (NSF1622526).

References

[1] F. T. Bourgeois, M. W. Shannon, C. Valim, K. D. Mandl. "Adverse drug events in the outpatient setting: an 11-year national analysis," *Pharmacoepidemiol Drug Safety*, vol. 19, no. 9, pp. 901–910, 2010.

[2] FAERS, "US food and drug administration (FDA) adverse event reporting system (FAERS)," [Online]. Available: https://www.fda.gov/ Drugs/InformationOnDrugs/ucm135151.htm. 2018.

[3] L. Wang, G. Jiang, D. Li, and H. Liu, "Standardizing adverse drug event reporting data," *J. Biomed. Semantics*, vol. 5, no. 1, 2014, Art. no. 36.

[4] 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 Safty*, vol. 10, no. 6, pp. 483–486, 2001.

[5] E. P. V. Puijenbroek, et al., "A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions," *Pharmacoepidemiol Drug Safety*, vol. 11, no. 1, pp. 3–10, 2002.

[6] A. Bate, et al., "A Bayesian neural network method for adverse drug reaction signal generation", *Eur. J. Clin. Pharmacol.*, vol. 54, no. 4, pp. 315–321, 1998.

[7] W. Dumouchel, "Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting system," *Amer. Statistician*, vol. 53, no. 3, pp. 177–190, 1999.

[8] M. Lindquist, I. R. Edwards, A. Bate, I. R. Edwards, H. Fucik, and A. M. Nunes, "From association to alert-a revised approach to international signal analysis," *Pharmacoepidemiol Drug Safety*, vol. 8, no. s1, pp. 7–31, 1999.

[9] R. Harpaz, et al., "Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system," *Clin. Pharmacology Therapeutics*, vol. 93, no. 6, pp. 539–546. 2013.

[10] P. Zhang, et al., "Three-component mixture model-based adverse drug event signal detection for the adverse event reporting system," *CPT-Pharmacometrics Syst. Pharmacology*, vol. 7, no. 8, pp. 499–506, 2018.

[11] M. Lindquist, M. Stahl, A. Bate, I. R. Edwards, and R. H. B. Meyboom, "A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database," *Drug Safety*, vol. 23, no. 6, pp. 533–542, 2000.

[12] R. Orre, A. Lansner, A. Bate, and M. Lindquist, "Bayesian neural networks with confidence estimations applied to data mining," *Comput. Statist. Data Anal.*, vol. 34, no. 4, pp. 473–493, 2000.

[13] A. Bate and SJ. Evans, "Quantitative signal detection using spontaneous ADR reporting," *Pharmacoepidemiology Drug Safety*, vol. 18, no. 6, pp. 427–436, 2009.

[14] J. P. F. Bai and D. R. Abernethy, "Systems pharmacology to predict drug toxicity: Integration across levels of biological organization," *Annu. Rev. Pharmacology Toxicology*, vol. 53, no. 1, pp. 451–473, 2013.

[15] N. Atias and R, Sharan, "An algorithmic framework for predicting side effects of drugs," *J. Comput. Biol.*, vol, 18, no. 3, pp. 207–218, 2010.

[16] L. C. Huang, X. Wu, and J. Y. Chen, "Predicting adverse side effects of drugs," *BMC Genomics*, vol. 12, no. S5, pp. S11–S23, 2011.

[17] M. Liu, et al., "Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs," *J. Amer. Med. Inf. Assoc.*, vol. 19, no. e1, pp. e28–e35, 2012.

[18] N. P. Tatonetti, P. Y. Patrick, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Sci. Transl. Med.*, vol. 4, no. 125, pp. 125–131, 2012.

[19] M. Duran-Frigola and P. Aloy, "Analysis of chemical and biological features yields mechanistic insights into drug side effects," *Chem Biol.*, vol. 20, no. 4, pp. 594–603, 2013.

[20] J. Lin, Q. Kuang, Y, Li, Y, Zhang, J. Sun, Z. Ding, and M. Li, "Prediction of adverse drug reactions by a network based external link prediction method," *Anal. Methods*, vol. 5, no. 21, pp. 6120–6127, 2013.

[21] AL. Hopkins, "Network pharmacology: The next paradigm in drug discovery," *Nature Chemical Biol.*, vol. 4, no. 11, pp. 682–690, 2008.

[22] F. Azuaje, "Drug interaction networks: An introduction to translational and clinical applications," *Cardiovasc Res.*, vol. 97, no. 4, pp. 631–641, 2013.

[23] A. Cami, A. Arnold, S. Manzi, and B. Reis, "Predicting adverse drug events using pharmacological network models," *Sci. Transl. Med.*, vol. 3, no. 114, pp. 114–127, 2011.

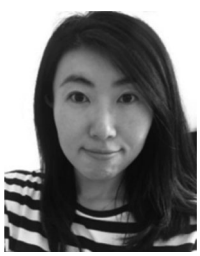
[24] F. Cheng, W. Li, X. Wang, Y. Zhou, Z, Wu, J. Shen, and Y. Tang, "Adverse drug events: database construction and in silico prediction," *J. Chemical Inf. Model*, vol. 53, no. 125, pp. 744–752, 2013.

[25] F. Cheng, et al., "Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space," *J. Chemical Inf. Model.*, vol. 53, no. 4, pp. 753–762, 2013

[26] MedDRA, "Medical Dictionary for Regulatory Activities," [Online]. Available: https://www.meddra.org, 2018.

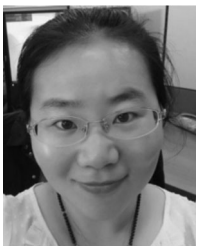
[27] PubChem, "Biochemical and biophysical drug properties," [Online]. Available: https://pubchem.ncbi.nlm.nih. gov, 2018.

[28] DrugBank, "Anatomical therapeutic chemical classification," [Online]. Available: https://www.drugbank.ca, 2018.

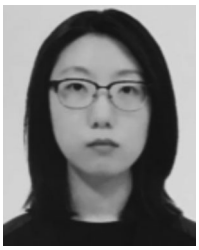


Xiangmin Ji received the MS degree from Harbin Engineering University, Harbin, China, in 2009. Currently, she is currently working toward the PhD degree in control science and engineering from Harbin Engineering University, Harbin, China. Her current research interests include bioinformatics, pharmacovigilance, data mining, and medical records analysis.

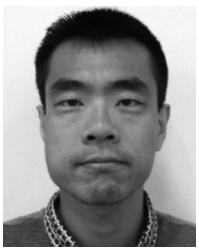
Lei Wang received the PhD degree in pattern recognition and intelligent systems from Harbin Engineering University, in 2013. His research focuses on translational biomedical informatics based systems pharmacology which integrate various biomedical data sources to explore high dimension drug-drug interactions and drug efficiency. He has published 20 scientific articles on several international academic journals such as *Clinical Pharmacology & Therapeutics*, *Pharmacometrics & Systems Pharmacology*. He is a member of the American Medical Informatics Association and the Chinese Computer Federation.



Liyan Hua received the MS degree from Harbin Normal University, Harbin, China, in 2012. Currently, she is working toward the PhD degree in control science and engineering from Harbin Engineering University, Harbin, China. Her current research interests include bioinformatics, data mining, and drug metabolic pathway analysis.



Xueying Wang received the BS degree from Harbin Engineering University, Harbin, China, in 2013. Currently, she is working toward the PhD degree in control science and engineering from Harbin Engineering University, Harbin, China. Her current research interests include bioinformatics, data mining, and medical records analysis.



Pengyue Zhang received the BS degree in mathematics, from Tianjin Normal University, in 2008, the MS degree in applied statistics, from Purdue University, in 2011, and the PhD degree biostatistics from Indiana University, in 2016. He is an instructor in the Departments of Biomedical Informatics within the Ohio State University College of Medicine. His research interests are in the fields of pharmacovigilance, pharmacometrics, and pharmacoepidemiology. His current focus lies in the development of study designs and statistical models for detecting adverse drug event (ADE) and drug interaction signals from spontaneous reporting system (SRS) and electronical medical record (EMR) and other databases.

Aditi Shendre received the MBBS degree in medicine, from B.J. Medical College, Pune, in 2006, and the MPH degree in public health and the PhD degree in epidemiology, from the University of Alabama at Birmingham, in 2010 and 2016, respectively. She is a postdoctoral researcher in the Department of Biomedical Informatics at the College of Medicine, the Ohio State University. Her research interests include pharmacoepidemiology and pharmacogenetics. Her current research focuses on developing and testing genetic hypotheses associated with adverse drug reactions (ADEs) determined from drug-drug interactions, using data from electronic health records and publicly available databases.

Weixing Feng received the BS degree in automation, and the MS degree and PhD degrees both in pattern recognition and intelligent system from Harbin Engineering University, China, in 1993, 2005, and 2007, respectively. He was a postdoctoral at Harbin Medical University from 2009 to 2011.He worked as a visiting scholar at Indiana University, US, from Nov. 2007 to Nov. 2008. He is currently a professor in Automation College, Harbin Engineering University. His research interests include quality control and calibration of high throughput sequencing data and mechanisms investigation on transcription process, where he has more than 120 peer reviewed publications.

Jin Li received the BS degree in computer science and technology, the MS degree in computer application, and the PhD degree in control theory and control engineering from Harbin Engineering University, China, in 1984, 1991, and 2001, respectively. She was a postdoctoral fellow in the Harbin Institute of Technology from 2002 to 2004. She worked as a senior visiting researcher at the Kitami Institute of Technology, Japan, from Dec. 1999 to Dec. 2000, and at Hong Kong Polytechnic University from Dec. 2001 to Feb. 2002, Jan. 2003 to Mar. 2003, respectively. She is currently a professor at Automation College, Harbin Engineering University. Her research interests include digital image processing and bioinformatics, where she has more than 100 peer reviewed publications.

Lang Li received the PhD degree in biostatistics from the University of Michigan, in 2001. He is a professor and the chair of biomedical informatics in the College of Medicine at the Ohio State University. He uses translational biomedical informatics approaches to identify drug targets, predict drug responses, and investigate drug interaction mechanisms. His lab employed population data (i.e., medical records, claims data, and patient case reporting system), literature data (i.e., PubMed), and omics profiling data to generate molecular signals, drug targets, and drugs for either developing therapeutics or identifying drug side effect and their molecular mechanisms.