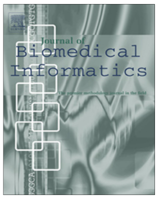
Journal of Biomedical Informatics 51 (2014) 15–23

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Exploring the associations between drug side-effects and therapeutic indications

Fei Wang [*,1], Ping Zhang [1], Nan Cao, Jianying Hu, Robert Sorrentino

IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

ARTICLE INFO

Article history:
Received 7 November 2013
Accepted 27 March 2014
Available online 13 April 2014

Keywords:
Theraputic indications
Side-effects
Associations

ABSTRACT

Drug therapeutic indications and side-effects are both measurable patient phenotype changes in response to the treatment. Inferring potential drug therapeutic indications and identifying clinically interesting drug side-effects are both important and challenging tasks. Previous studies have utilized either chemical structures or protein targets to predict indications and side-effects. In this study, we compared drug therapeutic indication prediction using various information including chemical structures, protein targets and side-effects. We also compared drug side-effect prediction with various information sources including chemical structures, protein targets and therapeutic indication. Prediction performance based on 10-fold cross-validation demonstrates that drug side-effects and therapeutic indications are the most predictive information source for each other. In addition, we extracted 6706 statistically significant indication-side-effect associations from all known drug-disease and drug-side-effect relationships. We further developed a novel user interface that allows the user to interactively explore these associations in the form of a dynamic bipartitie graph. Many relationship pairs provide explicit repositioning hypotheses (e.g., drugs causing postural hypotension are potential candidates for hypertension) and clear adverse-reaction watch lists (e.g., drugs for heart failure possibly cause impotence). All data sets and highly correlated disease-side-effect relationships are available at http://astro.temple.edu/~tua87106/druganalysis.html.

1. Introduction

Drug discovery is a slow and expensive process. By conservative estimation, it takes at least 10–15 years and USD 500 million to USD 2 billion to bring a single drug to market [1]. Although the research on drug development has increased significantly in recent years, the number of new therapeutic chemical and biological entities approved by the United States Food and Drug Administration (US FDA) has been declining since the late 1990s. There are two most important reasons for drugs fail clinical trials: (1) lack of efficacy; (2) adverse side-effect. And each of these two reasons accounts for around 30% of clinical trials failures [2]. Therefore it is highly desirable to develop tools that can predict drug therapeutic indications and side-effects accurately.

Therapeutic indication is a valid reason to use a certain medication. Inferring potential novel therapeutic indications for new or approved drugs is one important problem in drug development. Accurate indication prediction can drastically reduce the risk of attrition in clinical phases. In recent years, a number of computational methods have been developed to predict drug indications including.

- Inferring novel drug usage based on shared treatment profile using a network-based, guilt-by-association method [3].
- Predicting drug indications using their chemical structures [4].
- Inferring drug indications from protein targets interaction networks [5,6].
- Identifying relationships between drugs based on the similarity of their phenotypic profiles (e.g., side-effects [7,8] and connective map gene expression [9,10]).
- Integrating multiple information (e.g., chemical, biological, or phenotypic information) of drugs and diseases to predict drug indications [11–13].

With the exception of Yang et al. [8] which used side-effects, these strategies focus primarily on using preclinical information. However, clinical therapeutic effects are not always consistent with preclinical outcomes.

Drug side effect is a secondary, typically undesirable effect of a drug or medical treatment. Predicting drug side-effects, or adverse

[*] Corresponding author.
E-mail address: feiwang03@gmail.com (F. Wang).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.jbi.2014.03.014
1532-0464/© 2014 Elsevier Inc. All rights reserved.

drug reactions, is another important aspect of drug development. According to the statistics, serious drug side-effects has been the fourth leading cause of death in US, resulting in 100,000 deaths per year [14]. One approach for identifying potential adverse drug side-effects is preclinical in vitro safety profiling, which tests compounds with biomedical and cellular assays. However this experimental methodology is very expensive and labor intensive. Therefore, developing effective computational methods for accurate drug side-effect prediction is of vital importance. There have been some prior studies on this topic, which can be categorized into three classes:

- Linking drug side-effects to their chemical structures [15–17], following the spirit of QSAR (quantitative structure–activity relationship).
- Relating drug side-effects to its protein targets [18,19] because drugs with similar in vitro protein-binding profiles tend to exhibit similar side-effects.
- Predicting drug side-effects by integrating multiple data sources (e.g., chemical, biological, or phenotypic properties) [20–22].

From these existing studies we can see that, although therapeutic indications and side-effects are both measurable behavioral or physiological changes in response to the treatment, they have mostly been researched independently in the past. Intuitively, if drugs treating a disease share some common side-effects, this could suggest some underlying mechanism-of-action (MOA) linkage between the indicated disease and the side-effects. Moreover, many side effects are extensions of a drug's intended phenotypic effect (e.g., hyper- and hypo-tension), so it is logical that there is a correlation between indication and side effect. However, there is a lack of systematic study on exploring the associations between drug therapeutic indications and side-effects, which could be of broad interests in drug development and repositioning.

In this paper, we conducted a comprehensive investigation on building effective computational models for predicting drug therapeutic-indications and drug side-effects. We compared the predictive power of different sources of information (drug chemical structure, protein target, as well as disease indication and side-effects themselves), which shows that, indeed, drug side-effects and therapeutic indications are strong predictors of each other. This confirms the hypothesis that there exist strong associations between drug indications and side-effects. To quantize the strength of those associations, we performed Fisher's exact test with the prediction results [23]. Note that some preliminary evaluations on known associations between drug indications and side-effects are presented in our conference paper [24]. In this paper, we did a much more thorough investigation on all possible (both known and unknown) drug indication and side-effects associations. We also built a visualization tool to facilitate the user's exploration of those detected associations, which can be used to provide repositioning hypotheses (e.g., drugs causing postural hypotension are potential candidates for hypertension), as well as adverse-effect watch lists (e.g., drugs for heart failure possibly cause impotence).

The key differences between this paper and prior studies are:

- We evaluate effectiveness of both drug therapeutic indications and side-effects when predicting each other. Most prior work does not explicitly leverage the relationship between indications and side-effects, in combination with other drug properties. The prior work that is most closely associated with ours is Yang et al. [8]. However they used side-effects alone to predict drug indications. Moreover, their approach was only evaluated on a small data set (145 diseases and 584 side-effects). The data set we used in this paper is much larger, which includes 719 diseases and 1385 side-effects.
- We build disease-side-effect profiles to elucidate interesting relationships between drug side-effects and therapeutic indications with clinical implications, which provides a systematic way to generate drug indication hypotheses and adverse-effect watch lists. To the best of our knowledge, there is no prior work on this topic.
- We propose a novel visualization approach to support the interactive exploration of indication and side-effect associations in the form of a dynamic bi-partite graph.

The rest of this paper is organized as follows. In Section 2 we will introduce the details of the data set we used for this study. The methodology is presented in Section 3, followed by the experimental results in Section 4. Finally we will conclude in Section 5.

2. Data set

We performed our study on approved drugs from DrugBank [25], which is a widely used public drug information database. From DrugBank, we collected 1447 FDA-approved small-molecule drugs, and mapped them to PubChem [26] to get their chemical structure information. After matching by the DrugBank provided PubChem Compound ID for the drugs, we extracted chemical structures of the 1103 drugs. To encode the drug chemical structure, we used a fingerprint corresponding to the 881 chemical substructures defined in the PubChem. Each drug was represented by an 881-dimensional binary profile, within which the entry is 1 if the corresponding PubChem substructure is present, otherwise it is 0. Take the drug calcium as an example, its chemical formula is just Ca, which only meets the requirement of the bit 52 (>=1 Ca). Thus drug calcium only has 1 association with chemical substructures >=1 Ca. Similarly, aspirin has 115 associations with chemical substructures, ibuprofen has 84 associations with chemical substructures. A description of the 881 chemical substructures can be found at the website of PubChem (http://pubchem.ncbi.nlm.nih.gov/). Adding up together, we identified 132,092 associations between drugs and chemical substructures in the dataset, i.e., each drug has 119.8 substructures on average.

From DrugBank, we can also obtain the protein target information for each drug. To facilitate collecting such information, we mapped those target proteins to UniProt Knowledgebase [27], a central knowledgebase including the most comprehensive and complete information on proteins. After matching with the DrugBank provided UniProt ID for the drugs, we extracted 3152 relationships between 1007 drugs and 775 protein targets, so each drug has 3.1 protein targets on average. Similar to the chemical structure representation, each drug was represented by a 775-dimensional binary profile whose elements encode the presence or absence of their corresponding target proteins.

The third type of information we are interested in is drug side-effects. We extract side-effect keywords from the SIDER database [28], which contains information about medicines that are in market and their recorded adverse drug reactions. SIDER uses STITCH compound ids as its drug id, but can be easily matched to PubChem Compound ID via this rule (ftp://sideeffects.embl.de/SIDER/2012-10-17/README). This dataset contains 888 small-molecule drugs and 1385 side-effect keywords. Similar to the representations we mentioned above, each drug can be represented by a 1385-dimensional binary profile whose elements encode the presence or absence of each side-effect keyword. We plotted the cumulative counts of side-effect data in Fig. 1, from which we can observe that 1.69% of drugs have between 10 and 100 different side effects; 22% of drugs have more than 100 side-effects; only 9% of drugs have

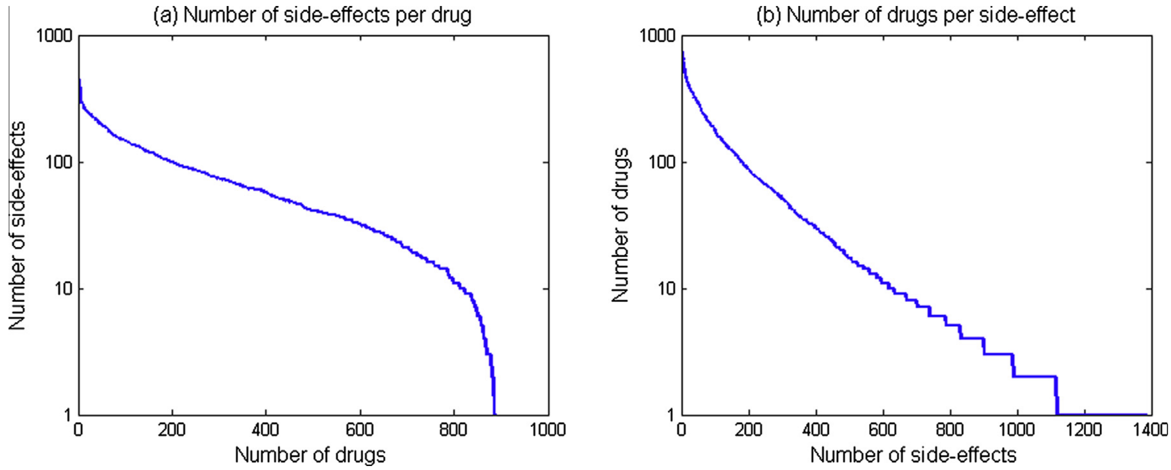


Fig. 1. Cumulative count information of the side-effect dataset. (a) The number of side-effects per drug. (b) The number of drugs per side-effect.

less than 10 side-effects (Fig. 1(a)). Also, 56% of all side-effects occur for less than 10 drugs; 32% of all side-effects occur for 10–100 drugs; 12% of all side-effects occur in more than 100 drugs (Fig. 1(b)). Altogether, there are 61,102 associations between drugs and side-effect terms in the dataset, so each drug has 68.8 side-effects on average.

The final piece of information we need to collect is drugs' therapeutic indications. These were obtained on extracting treatment relationships between drugs and diseases from the National Drug File-Reference Terminology (NDF-RT), which is part of the Unified Medical Language System (UMLS) [29]. This drug-disease treatment relationship list is also used by Li et al. [12] as the golden standard set of a drug repositioning study. The drug names in this list were DrugBank generic names, thus can directly be matched to the drugs from DrugBank. From the drug-disease treatment relationship list, we extracted 3250 treatment relationships between 799 drugs and 719 diseases. Thus each drug was represented by a 719-dimensional binary profile whose elements encode the presence or absence of each of the therapeutic indications. We plotted the cumulative count statistics of therapeutic indications data in Fig. 2. Most of drugs (75%) treat less than 5 indicated disease; 18% of drugs treat 5–10 diseases; only 7% of drugs treat more than 10 diseases (Fig. 2(a)). 80% Of the diseases have less than 5 drugs; 10% of the diseases have 5–10 related drugs; and the remaining 10% of diseases have more than 10 drugs (Fig. 2(b)).

3. Methodology

In this section, we introduce the details of the methodology we used in our study. Our method follows a very natural path on association study. Basically we first test whether there exists associations between drug therapeutic indications and side-effects by predictive modeling. If they two are both predictive to each other, there could exist associations between them. Then we apply statistical testing to obtain the significance of the existence of those associations. Finally we will show those significant associations on a user interface to facilitate the user's exploration. Fig. 3 provides a graphical illustration on the overall method flow, which include three phases: (I) predictive modeling for drug therapeutic indications, side-effects and various information sources; (II) Association analysis between drug therapeutic indications and side-effects with Fisher's exact test based on the prediction results from phase I; (III) Visualization of the discovered associations to facilitate interactive exploration. In the following we describe each phase in detail.

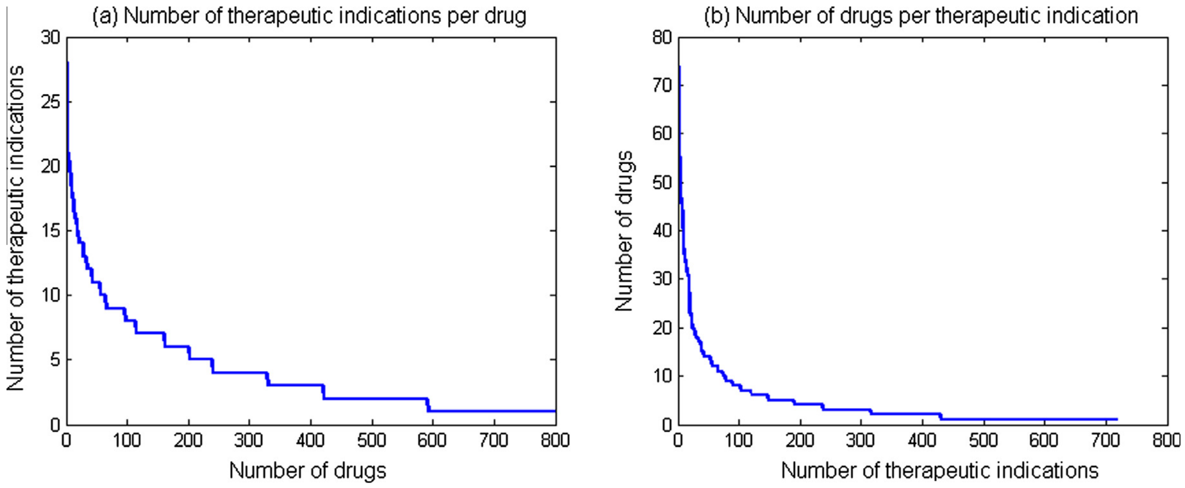


Fig. 2. Cumulative count information of the therapeutic indication dataset. (a) The number of therapeutic indications per drug. (b) The number of drugs per therapeutic indication.

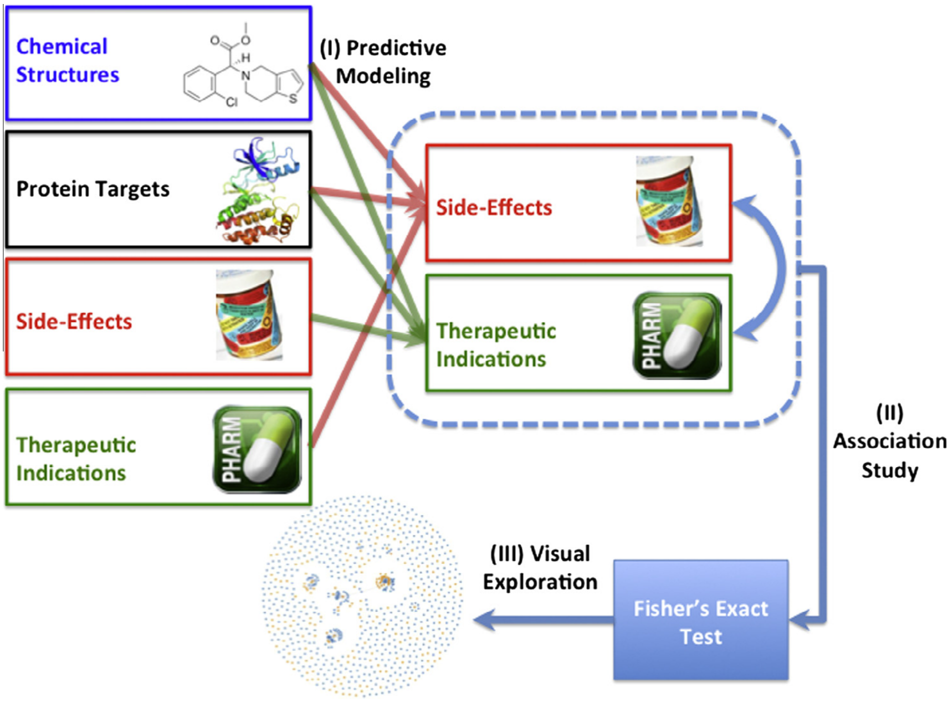


Fig. 3. The complete flowchart of our study, which include three phases: (I) predictive modeling for drug therapeutic indications and side-effects; (II) indication/side-effects association study from the predictive modeling results with Fisher's exact test; (III) visual exploration of the detected associations.

3.1. Predictive modeling

In the predictive modeling phase, we treat both drug therapeutic indication and side-effect prediction tasks as binary classification problems. For disease indication, we constructed a binary classifier using logistic regression for each of the 719 diseases. The input of every classifier is the drug related information. The output is the probability that the drug will have the corresponding disease therapeutic indication. Similarly, for side-effect prediction, we constructed a binary classifier with logistic regression for each of the 1385 side-effects. Our implementation uses Python 2.7 and the source codes of each of the four classifiers are available in the Scikit-Learn package [30] (http://scikit-learn.org/stable/). The model parameters are tuned with 10-fold cross validation.

We measure the final classification performance using three criteria: sensitivity, specificity, and area under the Receiver Operating Characteristic curve (AUC). In order to define those criteria, we construct the classification confusion matrix for binary classification problems as in Table 1, where the two classes are indicated as positive or negative. Then sensitivity is the true positive rate computed as TP/(TP + FN). Specificity is calculated as TN/(TN + FP), which is equal to one minus False Positive Rate. AUC score is the area under the ROC curve, which is a graphical plot of true positive rate vs. false positive rate. The whole ROC curve can be plotted by varying the threshold value for prediction score, above which the output is predicted as positive, and negative otherwise. The AUC score has been widely used as a classification performance measure in biomedical informatics [31]. After obtaining the AUC scores, we can get the optimal cut-off point by maximizing the corresponding sum of sensitivity and specificity scores.[2]

Table 1
Confusion matrix.

Predicted value\actual value	True	False
Positive	TP	FP
Negative	TN	FN

Table 2
Contingency table.

	Indication A	No indication A
Side-effect B	a	b
No side-effect B	c	d

3.2. Association analysis

After the predictive modeling procedure in phase I, association analysis is performed to capture the correlations between drug therapeutic indications and side-effects. In order to achieve this goal, we adopted Fishers exact test [23], which is a widely used approach for measuring the significance of the association between two nominal variables. For example, to test the significance of the association between drug therapeutic indication A and drug side-effect B, we first construct a 2 × 2 contingency table shown in Table 2, where *a* indicates the number of drugs that has indication A and side-effect B simultaneously according to our prediction from phase I; *b* indicates the number of drugs that does not have indication A but has side-effect B; *c* the number of drugs that has indication A but does not have side-effect B; *d* indicates the number of drugs that does not have either indication A or side-effect B. Then we can use the fisher_exact function in the statistics package in Scipy[3] to perform Fisher's exact test and get the *p*-value of the statistical testing of the association between indication A and side-effect B, the smaller the *p*-value, the stronger the association is. This test is repeated for all possible pairs of drug indication and side-effect.

[2] http://www.medicalbiostatistics.com/roccurve.pdf.

[3] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html#scipy.stats.fisher_exact.

3.3. Visual exploration

The strengths of all potential associations between drug indications and side-effects are then fed to the visual interface we developed to facilitate the user's interactive exploration. The visual interface is based on a dynamic bipartite graph layout strategy. It is bipartite because there are two types of nodes, drug therapeutic indications and side-effects, and the edges correspond to the discovered associations between them. Attached to each association there is a strength score (which is represented by the *p*-value of Fisher's exact test, The smaller the *p*-value the stronger the association). These strength scores can be used to support dynamic rendering of the graph. A sliding bar is provided which the user can use to adjust the association strength threshold: only the associations whose *p*-values are below the threshold will be shown. As the user drags that bar, the bipartite graph will change accordingly. We developed a novel methodology to make that change as smooth as possible for visual consistency. Fig. 4 gives an overview of the interface.

The dynamic bi-partite graph layout is based on an optimization approach designed to ensure smooth transition between graph renderings resulting from different association strength thresholds. Specifically, in order to help maintain a user's mental map, successive layouts of similar graphs when we tune the threshold should have minimal changes (*stability*). Furthermore, each of such layouts should still effectively convey the characteristics of the underlying graph (*readability*). Thus, our goal is to produce a sequence of graph layouts that optimize both the stability and readability of the resulting visualization. To achieve this goal, we developed a spectral layout algorithm.

Given a dynamic graph $\mathcal{G}_t = \langle \mathcal{V}_t, \mathcal{E}_t \rangle$ at time *t*, consisting of a set of nodes $\mathcal{V}_t$ and links $\mathcal{E}_t$, we define an energy function to model the desired graph layout as follows:

$$\min \left[\sum_{i<j} \omega_{ij} \alpha (||X_i - X_j|| - d_{ij})^2 + \sum_{i \in C_k} (1-\alpha)(X_i - X'_i)^2 \right] \tag{1}$$

where X'_i and X_i represent the previous and new position of node $v_i \in \mathcal{V}_t$, respectively. The first term of the objective in Eq. (1) is from the Kamada and Kawai method [32], which maximizes the readability of a graph visualization by preserving the pairwise distances, where d_{ij} is the shortest distance between two nodes v_i and v_j. The second item, which we have added, attempts to minimize the changes in successive layouts.

Instead of stabilizing all the unchanged nodes (which consists of a set $\mathcal{U}$), we extract a representative set of unchanged nodes $\mathcal{C}_k \in \mathcal{U}$ to improve the performance of the algorithm using the method we proposed in [33]. The final layout model is constructed by optimizing Eq. (1) with a spectral method. Here $\alpha \in (0, 1)$ is the weight that is dynamically computed to achieve the desired balance between readability and stability. An online demo of such user interface with our data can be found on http://nancao.org/demos/druggraph/.

4. Experimental results and discussion

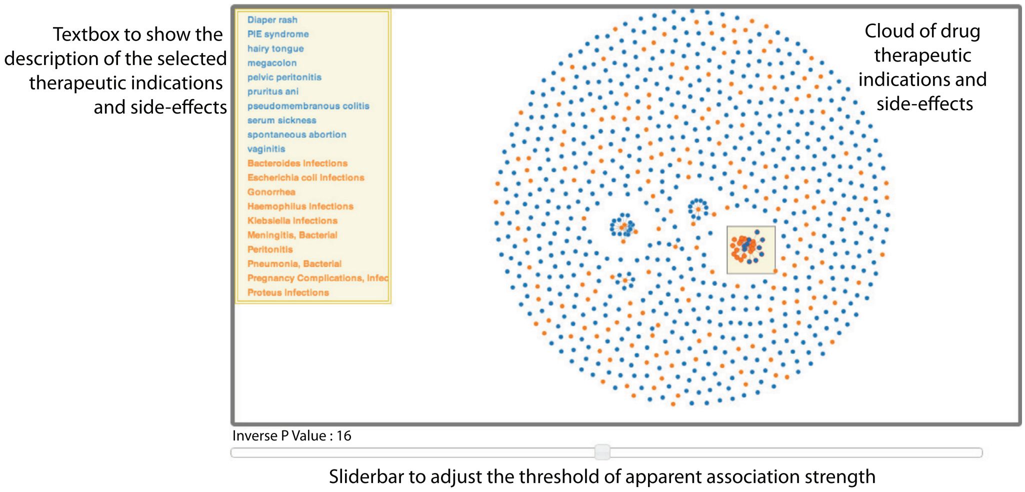
In this section we present details of the experimental results on our data set introduced in Section 2. All data sets used in our experiments are available at http://astro.temple.edu/~tua87106/druganalysis.html.

4.1. Prediction results

As introduced in Section 3, we have two tasks in the predictive modeling phase. For therapeutic indication prediction, we tested the following information combination as input: (1) chemical structures (881 dimensional); (2) protein targets (775 dimensional); (3) side-effects (1385 dimensional); (4) chemical structures + protein targets (881 + 775 dimensional); (5) chemical structures + side-effect (881 + 1385 dimensional); (6) protein targets + side-effect (775 + 1385 dimensional); (7) chemical structures + protein targets + side-effect (881 + 775 + 1385 dimensional).

For side-effect prediction, we tested the following combination as input: (1) chemical structures (881 dimensional); (2) protein targets (775 dimensional); (3) therapeutic indications (719 dimensional); (4) chemical structures + protein targets (881 + 775 dimensional); (5) chemical structures + therapeutic indications (881 + 719 dimensional); (6) protein targets + therapeutic indications (775 + 719 dimensional); (7) chemical structures + protein targets + therapeutic indications (881 + 775 + 719 dimensional).

Besides those different feature combinations, we also applied a random assignment procedure as baseline, where we used the 0/1


Textbox to show the description of the selected therapeutic indications and side-effects
Cloud of drug therapeutic indications and side-effects
Inverse P Value : 16
Sliderbar to adjust the threshold of apparent association strength

Diaper rash
PIE syndrome
hairy tongue
megacolon
pelvic peritonitis
pruritus ani
pseudomembranous colitis
serum sickness
spontaneous abortion
vaginitis
Bacteroides Infections
Escherichia coli Infections
Gonorrhea
Haemophilus Infections
Klebsiella Infections
Meningitis, Bacterial
Peritonitis
Pneumonia, Bacterial
Pregnancy Complications, Infec
Proteus Infections

Fig. 4. An overview of our visualization system, which consists of three parts: (I) the generated bipartite graph of the drug therapeutic indications (orange) and side-effects (blue); (II) a sliding bar for adjusting the significance threshold (*p*-value) for showing the discovered associations; (III) a text box displaying the descriptions of the selected nodes (the user can select a group of nodes with mouse and their descriptions will be depicted in this box). The clusters in the graph are connected drug therapeutic indications and side-effects under the current significance level, where any edge indicates an association whose significance is larger than the threshold specified sliding bar. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ratio to generate a binary label to each test drug randomly. For example, if the ratio in the given training data is 90%, we can assign zero to 90% of examples in test, and the rest to 1. This baseline is implemented for both indication prediction and side-effect prediction tasks.

We used 10-fold cross validation to evaluate the performance of all methods. For each fold, we held out all the associations involved with 10% of the drugs. For both indication prediction and side-effect prediction tasks, the sample sizes of output classes are highly imbalanced (negative class dominates). Consequently, the performance of the prediction could be overestimated. To alleviate this problem, we incorporated a sample balancing strategy in the 10-fold partitioning procedure, where all drugs were split into 10 equal-sized subsets, and each subset was used in turn as the testing set. For constructing the training set at each round of cross validation, we used all the positive drug-indication or drug-side-effect pairs from the remaining nine subsets, and randomly selected negative pairs from the same nine subsets, whose amount is twice as large as the positive pair number. The same strategy was also used in Gottlieb et al. [11]. To obtain robust results, we performed 10 independent cross-validation runs, in each of which a different random partition of the data set to 10 parts was used; we then computed the mean and the standard deviation of the evaluation scores over the entire 10 repetitions. To conduct a fair and accurate comparison across different data sources, we only considered the drugs which have all available sources for each task.

To evaluate the global performance across 719 diseases (for drug indication prediction) and 1385 side-effects (for drug side-effect prediction), we concatenate the prediction scores of all drugs over all diseases for drug indication prediction and draw a global ROC curve based on those scores. Then we compute the AUC value based on this overall ROC curve. Similar for side-effects prediction. This strategy has also widely been used in the past for both drug indication prediction tasks [11,12] and side-effect prediction tasks [16,22]. The reported sensitivity and specificity were obtained from the operating points of the global ROC curve, so that it gives the best tradeoff between false positives and negatives.

Fig. 5 summarizes the average ROC curves of 10 runs of the cross validation for different information sources for therapeutic indication prediction, and Table 3 summarizes the evaluation results. From those results we can see that when the information sources were compared independently, side-effect is the most informative (AUC of 0.8408), chemical structure ranks as the second (AUC of 0.8148), followed by target protein information (AUC of 0.8011). Overall, combining any two data sources improves the AUC, and adding side-effects works better than without it. The highest AUC score (AUC of 0.8640) is obtained by combing all three data sources.

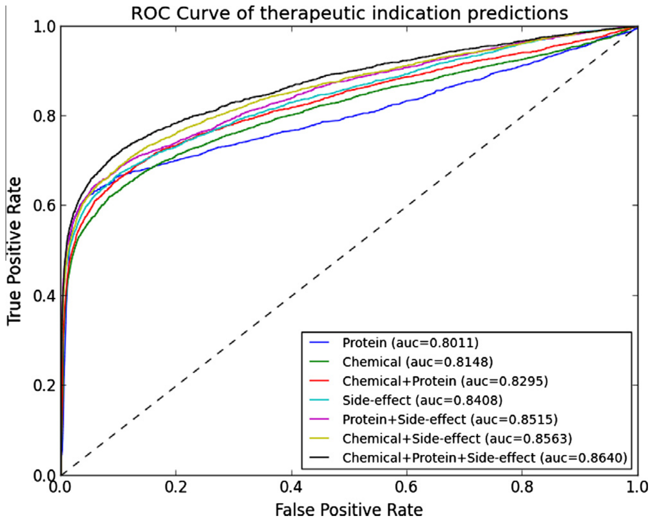


Fig. 5. The averaged ROC comparison of therapeutic indication predictions for various information source combinations using in 10-fold cross validation. Information sources are sorted in legend of the figure according to their AUC score.

Table 3
Performance comparison of drug therapeutic-indication prediction with different information sources.

Information source	AUC	Sensitivity	Specificity
Random	0.5000 ± 0.0010	0.0072 ± 0.0021	0.9929 ± 0.0002
Chemical	0.8148 ± 0.0019	0.5321 ± 0.0046	0.9647 ± 0.0004
Protein	0.8011 ± 0.0021	0.5387 ± 0.0038	0.9841 ± 0.0002
Side-effect	0.8408 ± 0.0036	0.5575 ± 0.0046	0.9737 ± 0.0004
Chemical + Protein	0.8295 ± 0.0021	0.4014 ± 0.0041	0.9921 ± 0.0001
Chemical + Side-effect	0.8563 ± 0.0022	0.6228 ± 0.0071	0.9516 ± 0.0006
Protein + Side-effect	0.8515 ± 0.0053	0.5625 ± 0.0070	0.9793 ± 0.0003
Chemical + Protein + Side-effect	**0.8640 ± 0.0035**	**0.6195** ± **0.0067**	**0.9650 ± 0.0004**

Similarly, Fig. 6 shows the average ROC curves of 10 runs of the cross validation for different information sources for side effect prediction, and Table 4 summarizes the results. When the information sources were compared independently, therapeutic indication is the most informative (AUC of 0.7058), target protein information is also highly informative (AUC of 0.6993), but chemical structure performed much worse (AUC of 0.6379). This could be partially explained with the following reasons. Both therapeutic indications

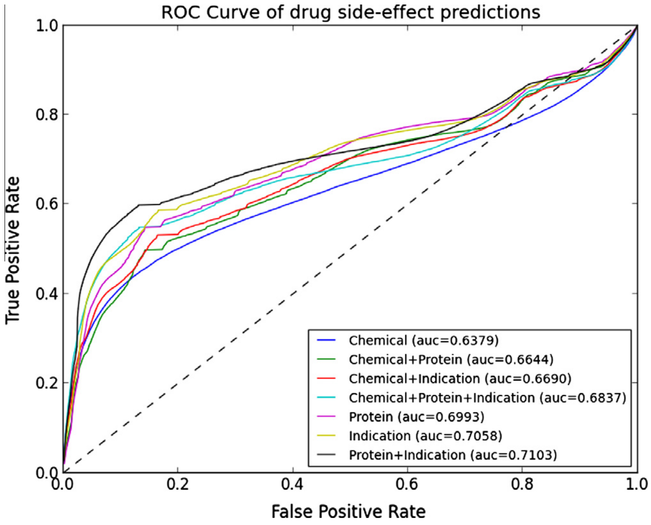


Fig. 6. The averaged ROC comparison of therapeutic indication predictions for various information source combinations using in 10-fold cross validation. Information sources are sorted in legend of the figure according to their AUC score.

Table 4
Performance comparison of drug side-effects prediction with different information source.

Information source	AUC	Sensitivity	Specificity
Random	0.5001 ± 0.0004	0.0599 ± 0.0007	0.9403 ± 0.0004
Chemical	0.6379 ± 0.0008	0.2436 ± 0.0012	0.9401 ± 0.0003
Protein	0.6993 ± 0.0014	0.4746 ± 0.0010	0.9128 ± 0.0006
Indication	0.7058 ± 0.0014	0.5207 ± 0.0017	0.8995 ± 0.0005
Chemical + Protein	0.6644 ± 0.0009	0.2843 ± 0.0016	0.9468 ± 0.0003
Chemical + Indication	0.6690 ± 0.0012	0.2881 ± 0.0016	0.9494 ± 0.0004
Protein + Indication	0.7103 ± 0.0011	0.4689 ± 0.0018	0.9319 ± 0.0002
Chemical + Protein + Indication	**0.6837** ± **0.0010**	**0.3035** ± **0.0015**	**0.9542** ± **0.0003**

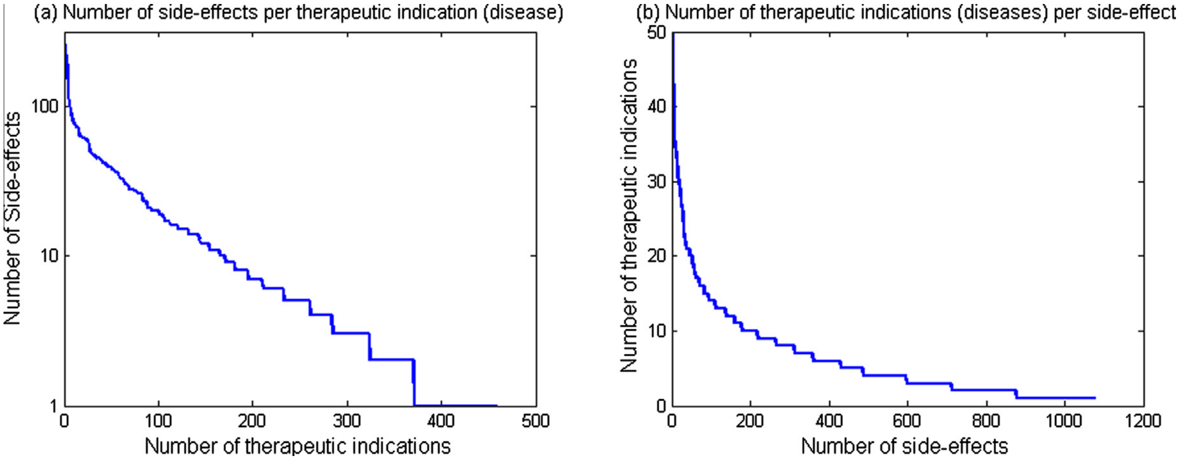


Fig. 7. Cumulative counts of the highly correlated disease-side-effect pairs result (p-value less than 0.01). (a) The number of side effects per disease (therapeutic indication). (b) The number of diseases (therapeutic indications) per side effect.

and side-effects are complex phenomenological observations that attributed to chemical structures (i.e., drugs) interact with primary or additional targets (off-targets hereafter). Expected activities derived from on-targets result in therapeutic effects. Unexpected (usually unwanted and harmful) activities derived from off-targets lead to side-effects. Minor differences in chemical structure of a drug may not affect the primary targets, therefore chemical structure could be very useful for predicting drug indications. However, even minor differences in chemical structure of a drug may cause a dramatic impact on how it interacts with off-targets, thus could result in significant differences in side-effect profiles of the drug. Therefore, drugs with similar chemical structures may not have similar side-effects, i.e., the performance could be bad if we use chemical structure to predict side-effects. While combing therapeutic indication and target protein results in the highest AUC score (AUC of 0.7103), combining chemical structure and any other information sources actually makes the prediction performance worse.

4.2. Association study results

From those predictive modeling results we can observe that drug side-effects is the most predictive information for drug therapeutic indications (with 0.8408 prediction AUC), and drug therapeutic indication is the most predictive information for drug side-effects (with 0.7058 prediction AUC). This suggests there indeed exists strong hidden correlations between them. To explore those correlations, we used Fisher's exact test as described in Section 3.2. We built indication-side-effect profiles, i.e., the most likely side-effects by the drugs which treat a specific disease, based on known and predicted drug-disease and drug-side-effect relationships. Among all 995,815 (719 diseases by 1385 side-effects) disease-side-effect pairs, there are 17,386 (1.75%) pairs with p-value less than 0.05, which is a typical threshold indicating whether a statistical testing is significant or not.

At a p-value cutoff of 0.01, we found 6706 highly correlated disease-side-effect pairs between 458 disease and 1077 side-effects. On average, each disease's drugs very likely to cause 14.6 side-effects and each side-effect highly associates with 6.2 types of diseases. We plotted the cumulative counts of highly correlated disease-side-effect pairs (p-value less than 0.01) in Fig. 7, from which we can observe that 63% of the diseases highly correlate with less than 10 side-effects; 36% of the diseases highly correlate with 10–100 drugs; only 4 diseases highly correlate with more than 100 side-effects (Fig. 7(a)). For example, disease Obsessive–Compulsive Disorder is highly correlated with 260 side-effect keywords in our analysis, but only 7 drugs treat this disease in our drug-disease dataset. 60% Of side-effects are highly associated with less than 5 diseases; 24% of side-effects are highly associated with 5–10 diseases; 16% of side-effects are highly associated with more than10 diseases (Fig. 7(b)).

To better illustrate the associations we found, we provided two concrete examples in Table 5, which are 10 most closely correlated side-effects for diseases **Hypertension** and **Pain**. From the table we can observe that for hypertension, some of the side-effects are physiologically related and the mechanism of action (MOA) can be explained. For example, some hypertension drugs may result in a sudden drop in blood pressure when a person stands up, thus the side-effect postural hypotension happens. Some hypertension drugs (e.g., β-blockers) hits α-adrenergic receptors protein target in penile tissue, which will cause side-effect impotence. The decreased blood pressure caused by some hypertension drugs (e.g., β-blockers) also cause side-effects syncope, vertigo, and weakness. Side-effect pemphigus is related to ACE inhibitors, which is also one kind of hypertension drug. Some hypertension treatments (e.g., Diuretics) cause human body to lose salt and water, potentially precipating side-effects gout and hyperuricemia. Similarly, for drugs that treat pain, nonsteroidal anti-inflammatory pain medicines (e.g., Ibuprofen and Celecoxib) increase risk of heart attack and stroke, and may cause tachycardia and heart block

Table 5
10 Most correlated side-effects for disease hypertension and pain.

Disease	Side-effects	*P*-value
Hypertension	Claudication	1.36E−23
	Impotence	5.20E−17
	Postural hypotension	4.33E−14
	Cold extremities	6.25E−14
	Gout	4.72E−12
	Pemphigus	4.94E−12
	Syncope	5.17E−08
	Weakness	4.52E−07
	Hyperuricemia	7.20E−07
	Vertigo	2.05E−06
Pain	Hallucinations	3.78E−05
	Heart block	5.66E−05
	Tachycardia	1.08E−04
	Apnea	7.57E−04
	Forgetful	1.19E−03
	Ventricular extrasystoles	1.85E−03
	Somnolence	2.79E−03
	Urinary retention	4.56E−03
	Blindness	8.11E−03
	Tinnitus	9.05E−03

as side-effects. Low doses of tricyclic or tetracyclic antidepressant drugs increase the level of certain brain chemicals, which affect how the brain perceives pain. But they cause side-effect urinary retention. Other types of antidepressants (e.g., SSRI and SNRI) also cause somnolence, and delayed ejaculation.

As another example, Table 6 shows the top 10 therapeutic indications (diseases) with strongest correlation to side-effects **weight loss** and **impotence**. For side-effect weight loss, many related diseases in the list are mood disorders (e.g., bipolar disorder, depressive disorder, panic disorder). The most widely prescribed mood control drugs come from a class of medications known as selective serotonin reuptake inhibitors (SSRIs, such as Prozac, Zoloft). SSRIs act on serotonin, a chemical in the brain that helps regulate mood. However, serotonin also plays a role in digestion, sleep and other bodily functions. Thus mood control drugs result in dizziness, nausea, loss of appetite, and finally cause weight loss. Similarly, the drugs for Alzheimer disease (e.g., Aricept, Cognex, Exelon) cause vomiting, nausea, loss of appetite, thus result in weight loss. For side-effect impotence, drugs for cardiovascular diseases (e.g., hypertension, heart failure) appear on the list, because they can lower the pressure inside blood vessels, so the heart does not have to work as hard as usual to pump blood throughout the body. However, the decreased blood flow can reduce desire and interfere with erections and ejaculation, thus cause impotence. Some cardiovascular drugs limit the availability of cholesterol and likely interfere with the production of testosterone, estrogen and other sex hormones, also cause impotence. Drugs for mood disorders (e.g., depressive disorder, bipolar disorder) are on the list as well because they will block the action of brain chemicals that relay signals between nerve cells, thus decrease sex drive, causing impotence as a side-effect.

Table 6
10 Most correlated indicated diseases for side-effect weight loss and impotence.

Side-effect	Disease	*P*-value
Weight loss	Bipolar disorder	6.54E−07
	Breast neoplasms	1.66E−06
	Alzheimer disease	8.23E−06
	Panic disorder	5.16E−05
	Epilepsies, partial	1.02E−04
	Colorectal neoplasms	1.09E−04
	Attention deficit disorder with hyperactivity	1.17E−04
	Diarrhea	1.59E−04
	Depressive disorder	6.06E−04
	Asthma	3.91E−03
Impotence	Hypertension	5.20E−17
	Heart failure	1.01E−08
	Diabetic nephropathies	4.61E−08
	Depressive disorder	2.22E−07
	Urinary tract infections	5.16E−06
	Bipolar disorder	3.56E−05
	Schizophrenia	3.56E−04
	Angina pectoris	6.81E−04
	Asthma	2.14E−03
	Myocardial infarction	7.99E−03

Both therapeutic indications and clinical side-effects are human phenotypic data obviating translation issues. Therefore, those strongly correlated disease-side-effect pairs are beneficial for drug discovery in the following sense.

- We could use the side-effects information to generate hypotheses for repurposing existing treatments. For example, based on the information of Table 5, we may consider drugs with side-effect postural hypotension as candidates for treating hypertension. Also based on the information in Table 6, we may consider and evaluate some mood-disorder drugs for the usage of weight loss (i.e., as weight-loss pills).
- If a new treatment is designed for a specific disease, all health care stakeholders (e.g., regulators, providers, patients and pharmaceutical companies) should pay more attention to adverse reactions in the associated side-effect list of the disease (e.g., Table 5 for hypertension and pain), and control the formulation and dosing of drugs in the clinical trials to prevent serious safety issues.

4.3. Visual exploration example

Although the discovered drug therapeutic indication and side-effect associations are informative, it would be very difficult to check all of them one by one on spreadsheets like Tables 5 and 6. The visual interface introduced in Section 3.3 can be used to explore these associations in a much more efficient and effective

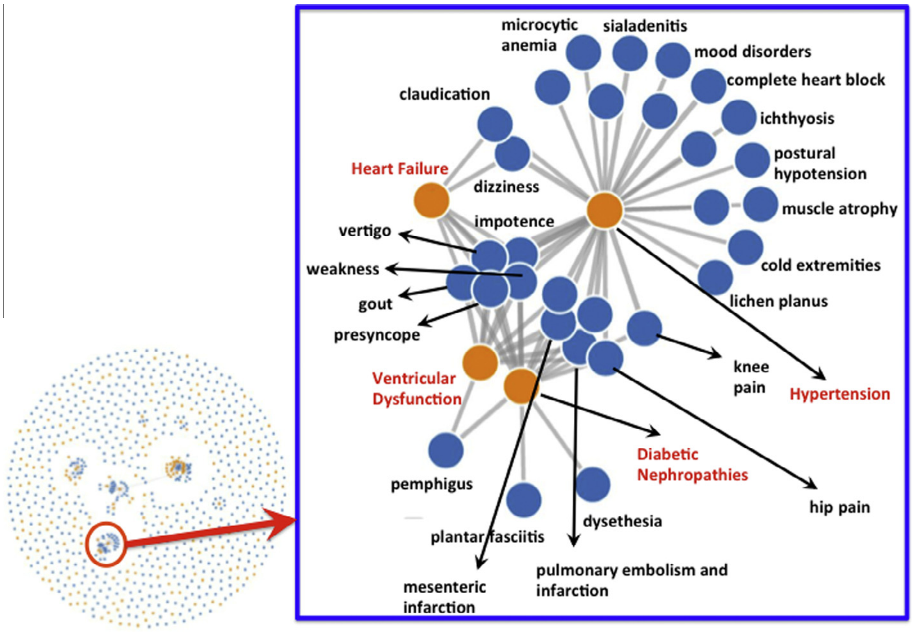


Fig. 8. An example clique of drug therapeutic indications and side-effects.

mannar. Based on the predictive modeling results, we obtained 167,392 predicted associations between 567 drugs and 1262 side-effect terms, and 22,639 predicted associations between 567 drugs and 612 indications. Then we combine both predicted and ground truth indication-side-effects associations and use Fisher's exact test to obtain all their *p*-values.

We demonstrate those discovered associations in our visualization system and show an example clique in Fig. 8, where there are four diseases: *Diabetic Nephropathies, Heart Failure, Hypertension, and Ventricular Dysfunction*. The last three are cardiovascular diseases, and Diabetic Nephropathies is a common comorbidity and one of the causes of cardiovascular diseases. This clique also contains 30 highly correlated side-effects. Some of the side-effects are physiologically linked to the cardiovascular diseases and the mechanism of action (MOA) can be explained. For example, some hypertension drugs may result in a sudden drop in blood pressure when a person stands up, thus the side-effect postural hypotension happens. Some cardiac drugs (e.g., β-blockers) hits α-adrenergic receptors protein target in penile tissue, which will cause side-effect impotence. The decreased blood pressure caused by some cardiac drugs (e.g., β-blockers) also cause side-effects cold extremities, dizziness, vertigo, and weakness. Side-effect pemphigus is related to ACE inhibitors, which may induce an autoimmune response to skin proteins. Some popular cardiac medications (e.g., Diuretics) cause human body to lose salt and water, potentially precipitating side-effect gout.

5. Conclusion

In this paper, we described a systematic study on the exploration of multiple sources of information (and their combinations) for therapeutic-indication and drug side-effect predictions. We found that side-effect and therapeutic indication are most predictive factors for each other, thus confirming that there exist strong association between drug indications and side effects. Furthermore, we performed statistical testing to obtain the strength of the discovered associations and developed a novel visual interface to facilitate the interactive exploration of these associations in a dynamic and comprehensive manner. These findings and tools could provide a powerful mechanism for hypothesis generation, which can be used to improve the drug development process via better targeted trial designs.

References

[1] DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ 2003;22:151–85.

[2] Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 2008;4:682–90.

[3] Chiang AP, Butte AJ. Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. Clin Pharmacol Therapeut 2009;86:507–10.

[4] Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature 2009;462:175–81.

[5] Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts. PLoS Comput Biol 2009;5:e1000450.

[6] Kotelnikova E, Yuryev A, Mazo I, Daraselia N. Computational approaches for drug repositioning and combination therapy design. J Bioinform Comput Biol 2010;8:593–606.

[7] Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science 2008;321:263–6.

[8] Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. PloS One 2011;6:e28025.

[9] Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. PLoS One 2009;4. e6536.

[10] Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Translat Med 2011;3. 96ra77.

[11] Gottlieb A, Stein GY, Ruppin E, Sharan R. Predict: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 2011;7.

[12] Li J, Lu Z. A new method for computational drug repositioning using drug pairwise similarity. In: 2012 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE; 2012. p. 1–4.

[13] Zhang P, Agarwal P, Obradovic Z. Computational drug repositioning by ranking and integrating multiple data sources. In: Machine learning and knowledge discovery in databases. Springer; 2013. p. 579–94.

[14] Giacomini KM, Krauss RM, Roden DM, Eichelbaum M, Hayden MR, Nakamura Y. When good drugs go bad. Nature 2007;446:975–7.

[15] Atias N, Sharan R. An algorithmic framework for predicting side effects of drugs. J Comput Biol 2011;18:207–18.

[16] Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinform 2011;12:169.

[17] Scheiber J, Jenkins JL, Sukuru SCK, Bender A, Mikhailov D, Milik M, et al. Mapping adverse drug reactions in chemical space. J Med Chem 2009;52:3103–7.

[18] Scheiber J, Chen B, Milik M, Sukuru SCK, Bender A, Mikhailov D, et al. Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. J Chem Inform Model 2009;49:308–17.

[19] Xie L, Li J, Xie L, Bourne PE. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of cetp inhibitors. PLoS Comput Biol 2009;5:e1000387.

[20] Yamanishi Y, Pauwels E, Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces. J Chem Inform Model 2012;52:3284–92.

[21] Huang L-C, Wu X, Chen JY. Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. Proteomics 2013;13:313–24.

[22] Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen X-W, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc 2012;19:e28–35.

[23] Upton GJ. Fisher's exact test. J Roy Stat Soc. Ser A (Stat Soc) 1992:395–402.

[24] Zhang P, Wang F, Hu J, Sorrentino R. Exploring the relationship between drug side-effects and therapeutic indications. In: Proceedings of Annual Symposium of American Medical Informatics Association (AMIA); 2013 p. 1568–1577.

[25] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. Drug actions and drug targets. Nucl Acids Res 2008;36:D901–6.

[26] Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. Pubchem: a public information system for analyzing bioactivities of small molecules. Nucl Acids Res 2009;37:W623–33.

[27] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. Uniprot: the universal protein knowledgebase. Nucl Acids Res 2004;32:D115–9.

[28] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6.

[29] Bodenreider O. The unified medical language system (umls): integrating biomedical terminology. Nucl Acids Res 2004;32:D267–70.

[30] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.

[31] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005;38:404–15.

[32] Kamada T, Kawai S. An algorithm for drawing general undirected graphs. Inform Process Lett 1989;31:7–15.

[33] Cao N, Liu S, Tan L, Zhou X. Interactive poster: context-preserving dynamic graph visualization. In: IEEE symposium on information visualization; 2008.