Elsevier Editorial System(tm) for Journal of Biomedical Informatics

Manuscript Draft

Manuscript Number: JBI-18-819

Title: A machine learning approach to the prediction of drug side effects using repositioning candidates, known indications, and other drug properties

Article Type: Research paper

Keywords: Machine learning, drug repositioning, side effect prediction, bioinformatics, systems biology

Corresponding Author: Professor Youngmi Yoon, ph.D.

Corresponding Author's Institution: Gachon University

First Author: Sukyung Seo, Bachelor of Engineering

Order of Authors: Sukyung Seo, Bachelor of Engineering; Min Oh; Mi-hyun Kim; Youngmi Yoon, ph.D.

Abstract: In the past decades, side effects have been a primary cause of failure for drugs to be approved by the Food and Drug Administration (FDA) and to be commercialized. As side effects can pose major risks not only to pharmaceutical companies but also to patients, it is important to identify the potential side effects of drugs to prevent unexpected outcomes. As some drug repositioning studies have used side effects to predict candidate indications, this study set out to drug repositioning candidates, known indications, and other types of drug properties to identify candidate side effects. It aimed to propose a machine learning approach to the identification of potential side effects based on diverse information about drugs: (1) the candidate indications of repurposed drugs, (2) the known therapeutic indications of drugs from a variety of databases, and (3) the drugs' chemical structures and target proteins. First, the drug-drug similarities in indications and the properties mentioned above were calculated by taking into account the indications contained in each database. Second, a set of features for a drug-side effect pair was built by selecting the maximum drug-drug similarity for each feature using the known associations between the side effect and other drugs. Third, training sets were created with different combinations of features and were subjected to four classification algorithms, including a random forest, neural network, logistic regression, and naïve Bayesian algorithm. The findings revealed that the addition of indication features yielded better results than the use of chemical and target features only. The method developed in this study yielded 7% better results in the area under the curve (AUC) than when using target and chemical features only. Moreover, the random forest model showed the best results for all combinations of feature types. Finally, it allowed to find the candidate side effects of drugs. This study focused on the four following drugs: 1) dasatinib, 2) sitagliptin, 3) vorinostat and 4) clonidine.

Suggested Reviewers: Tae-Hyuk Ahn
Saint Louis University

ted.ahn@slu.edu

Jong-Il Kim
Seoul National University
jongil@snu.ac.kr

Tae Hyun Hwang
Lerner Research Institute, Cleveland Clinic
hwangt@ccf.org

Opposed Reviewers:

October 22th, 2018

Dear Editor-in-Chief

Attached please find my manuscript entitled “A machine learning approach to the prediction of drug side effects, using repositioning candidates, known indications, and other drug properties” by Seo, Oh, Kim, and Yoon, submitted to Journal of Biomedical Informatics for consideration of publication. We identified candidate side effects of drugs based known indications, repositioning candidates, and other drug properties by applying diverse machine learning algorithms. All authors are confident that the subject and the result of our study are appropriate for Journal of Biomedical Informatics, and we believe it will draw attention from the bioinformatics fields.

Thank you for your attention to my paper. I am looking forward to your reply.

Sincerely,

Youngmi Yoon, Ph.D.

Department of Computer Engineering, Gachon University

1342 Seongnamdaero, Sujeong-gu, Seongnam-si,

Gyeonggi-do, Korea

(Tel) +82 31 750 4755

(e-mail) ymyoon@gachon.ac.kr

A machine learning approach to the prediction of drug side effects using repositioning candidates, known indications, and other drug properties

Sukyung Seo[a], Min Oh[b], Mi-hyun Kim[c], Youngmi Yoon[a,*]

Affiliations

[a] Department of Computer Engineering, Gachon University, Seongnam, Republic of Korea

[b] Department of Computer Science, Virginia Tech, Blacksburg, VA 26061, USA

[c] Gachon Institute of Pharmaceutical Science and Department of Pharmacy, College of Pharmacy, Gachon University, Yeonsu-gu, Incheon, Republic of Korea

[*] Corresponding author Tel: +82 31 750 4755; E-mail address: ymyoon@gachon.ac.kr

Abstract

In the past decades, side effects have been a primary cause of failure for drugs to be approved by the Food and Drug Administration (FDA) and to be commercialized. As side effects can pose major risks not only to pharmaceutical companies but also to patients, it is important to identify the potential side effects of drugs to prevent unexpected outcomes. As some drug repositioning studies have used side effects to predict candidate indications, this study set out to drug repositioning candidates, known indications, and other types of drug properties to identify candidate side effects. It aimed to propose a machine learning approach to the identification of potential side effects based on diverse information about drugs: (1) the candidate indications of repurposed drugs, (2) the known therapeutic indications of drugs from a variety of databases, and (3) the drugs' chemical structures and target proteins. First, the drug-drug similarities in indications and the properties mentioned above were calculated by taking into account the indications contained in each database. Second, a set of features for a drug-side effect pair was built by selecting the maximum drug-drug similarity for each feature using the known associations between the side effect and other drugs. Third, training sets were created with different combinations of features and were subjected to four classification algorithms, including a random forest, neural network, logistic regression, and naïve Bayesian algorithm. The findings revealed that the addition of indication features yielded better results than the use of chemical and target features only. The method developed in this study yielded 7% better results in the area under the curve (AUC) than when using target and chemical features only. Moreover, the random forest model showed the best results for all combinations of feature types. Finally, it allowed to find the candidate side effects of drugs. This study focused on the four following drugs: 1) dasatinib, 2) sitagliptin, 3) vorinostat and 4) clonidine.

Keywords: Machine learning, drug repositioning, side effect prediction, bioinformatics, systems biology

1. Introduction

Over the past decades, the pharmaceutical industry has been facing low R&D productivity [1]. Because of side effects, most drug candidates fail to achieve US Food and Drug Administration (FDA) approval and commercialization. Identifying the undesirable off-target activities that may cause side effects leading to drug discovery failure is a challenging part of the drug developmental process. While most serious side effects are discovered in pre-clinical and clinical trials, some are only reported during post-approval monitoring. The uncertainty surrounding the potential side effects derived from new drugs is a great concern not only for pharmaceutical companies but also for patients, as it poses a risk to their health and can even cause deaths [2].

The existing computational methods used to predict side effects have assumed that similar drugs have similar properties in terms of chemical, and biological characteristics such as target and structure. Pauwels *et al.* predicted potential side effects through a sparse canonical correlation analysis model based on chemical structures [3], while Mizutani *et al.* developed a candidate method to predict the potential side effects of drugs based on their chemical structures and target proteins [4]. It has long been known that drugs with similar chemical structures exhibit similar biological activities [5]. Even in drug design studies (which predict the properties of chemical compounds), the screening of a large number of chemical databases containing the structure of available chemicals is a key process [6].

Meanwhile, drug repositioning–that is, the repurposing of approved drugs for new indications–has been highlighted for its time saving benefits for the development and approval of drugs. As repurposed drugs have already passed safety and toxicity tests, they carry a low-risk of failure. Many recent systematic and experimental studies have been focusing on the identification of new indications for existing drugs. Interestingly, recent successful computational approaches have leveraged the side effect information of drugs to predict drug repositioning candidates under the assumption that if drugs have similar side effects, they might also have similar therapeutic effects. For example, Ye *et al.* found new drug indications by building drug-drug network based on the side effect similarities between drugs [7]. Gottlieb *et al.* also used the side effects of drugs to measure drug-drug similarity, and calculated the disease-disease similarity to predict candidate indications [8].

Another possible assumption can be derived from the above theory by flipping it around: Drugs with similar therapeutic effects might have similar side effects. A key reason similar side effect profiles have a descriptive power in the prediction of drug indications is that the similar side effects of unrelated drugs can be caused by common drug targets [9]. Likewise, it is logical to expect that common drug targets that trigger similar therapeutic effects might induce similar signaling cascades, and therefore, similar side effects. A few studies have used drug indication data to inform machine learning models for the prediction of drug side effects. Liu *et al.* added drug-phenotypic information as one of the features for machine learning besides chemical and biological information and showed significant improvements in machine learning results when phenotypic features were integrated with other properties such as chemical and biological information [10]. Zhang *et al.* used Liu's dataset as a benchmark, and developed an ensemble learning model by combining individual feature selection-based models [11]. He also included indications as one of the features and proved that their use was effective. Muñoz *et al.* made a graph using drug indications as well as their other properties of a drug, and found possible side effects by propagating the existing knowledge on them [12]. However, these studies have tended to

rely on a single source to obtain drug indications. Furthermore, no studies have yet considered the new indications of drug repositioning candidates as features, although the latter may give a broader view of drugs' actions.

This study set out to propose a machine learning approach to identify potential side effects by leveraging various information resources for drug indications and properties: (1) the candidate indications of repurposed drugs, (2) the known therapeutic indications of drugs from diverse databases, and (3) drugs chemical structures and target proteins. Instead of merging the indication information found in different databases, each indication set was used separately to extract a feature, and diverse machine learning algorithms were adopted. The results showed that the use of new indications for repurposed drugs and of known therapeutic indications significantly improved the prediction of candidate side effects as compared with the use of chemical and target data alone. Furthermore, optimizing the machine learning model for maximal performance made it possible to obtain the unlabeled side effects of approved drugs. The four following drugs are discussed: 1) dasatinib, 2) sitagliptin, 3) vorinostat and 4) clonidine.

2. Datasets

Table 1 presents the number of drugs and indications–including both repositioned and known indications–in each database.

Table 1. Number of drugs, indications, and associations in databases

Type	Database name	Number of drugs	Number of indications	Number of associations
Repositioned drug indications	repoDB	1,519	1,229	6,677
Known drug indications	TTD	1,000	1,298	44,481
	CTD	1,510	5,709	150,175
	PREDICT from Gottlieb	590	304	1,912
	ClinicalTrials	772	7,883	40,638

2.1 Repositioned drug indications

Recently, the identification of new indications for approved drugs has been gaining growing attention in the pharmacology field. Drug repositioning helps to reduce the time and costs involved in the development of new compounds to treat diseases. repoDB is a web service that provides a collection of repurposed drugs, as drawn from DrugCentral [13] and ClinicalTrial.gov [14]. This study used 6,677 associations spanning 1,519 drugs and 1,229 diseases.

2.2 Known drug indications

Various drug indications were collected from four different databases, each of them corresponding to a single

feature. First, we used a therapeutic target database (TTD) containing indications manually curated from the literature. For each disease, we mapped the ICD-10 (International Classification of Diseases) code onto the unified medical language system (UMLS) provided by Disease Ontology (DO) [15], and obtained 44,481 associations spanning 1,000 drugs and 1,298 diseases. Second, we used comparative toxicogenomics database (CTD) consisting of curated and inferred chemical-disease associations [16]. However, among the curated associations, we only included those for which there was direct evidence of a chemical-disease association marked as a "marker/mechanism" or "therapeutic," and did not use the inferred associations. The "marker/mechanism" tag means that a chemical has an association with a disease or may play a role in its the etiology. As a result, 150,175 associations spanning 1,510 drugs and 5,709 diseases were processed from the CTD. Third, we secured the gold standard for 1,912 drug-disease associations spanning 590 drugs and 304 diseases from PREDICT [17], which compiles a comprehensive set of drug indications from DrugBank [18] and Online Mendelian Inheritance in Man (OMIM) [19]. Finally, the report-based database and registration system ClinicalTrials was used. We compiled 40,638 drug indications from phase 1 to 5 spanning 772 drugs and 4,181 diseases. The drug names and disease names were mapped using a drug vocabulary file provided by DrugBank and a disease vocabulary file provided by CTD, respectively. The data analyzed in this study totaled contains 40,638 associations spanning 772 drugs with DrugBank ID and 7,883 diseases with UMLS ID, respectively.

2.3 Chemical structures, drug targets, and known side effects

Chemical structures were used to represent the chemical traits of drugs. We obtained the SMILES format of 1,878 drugs from PubChem and DrugBank [20], and downloaded 6,076 drug-target associations for 1,366 drugs from DrugBank.

To use up-to-date known side effects, 107,878 drug-side effect associations were collected from the SIDER database (September 2017). For each drug, we mapped the PubChem compound ID to the DrugBank ID using the annotations provided by biodb.jp [21]. For validation purposes, the drug-side effect events were extracted using the FDA Adverse Event Reporting system (FAERS), which uploads event lists four times a year (2015.4 ~ 2017.9) [22]. The FAERS was compiled into 182,9465 drug-side effect associations for 2,921 drugs and 5,737 side effects. We compiled the data from Health Canada's adverse drug reaction reporting system MedEffect into 2,466 drug-side effect associations for 55 drugs and 902 side effects [23]. FAERS and MedEffect are both post-marketing surveillance systems.

3. Methods

Fig. 1 offers a system overview and a detailed description of the methods used.

Fig. 1-(A): The drug-drug similarities were calculated using indication databases, drug targets, and chemical structures. Fig. 1-(B): To build a set of features for a drug-side effect pair, we selected the maximum drug-drug similarity for each feature based on the known associations between the side effect and other drugs. Fig. 1-(C): Based on the values of the pair found in (B), training sets were created with different combinations of features, and diverse classification algorithms–including a random forest, neural network, logistic regression, and naïve

Bayesian model–were applied to predict the association between a side effect and a drug.

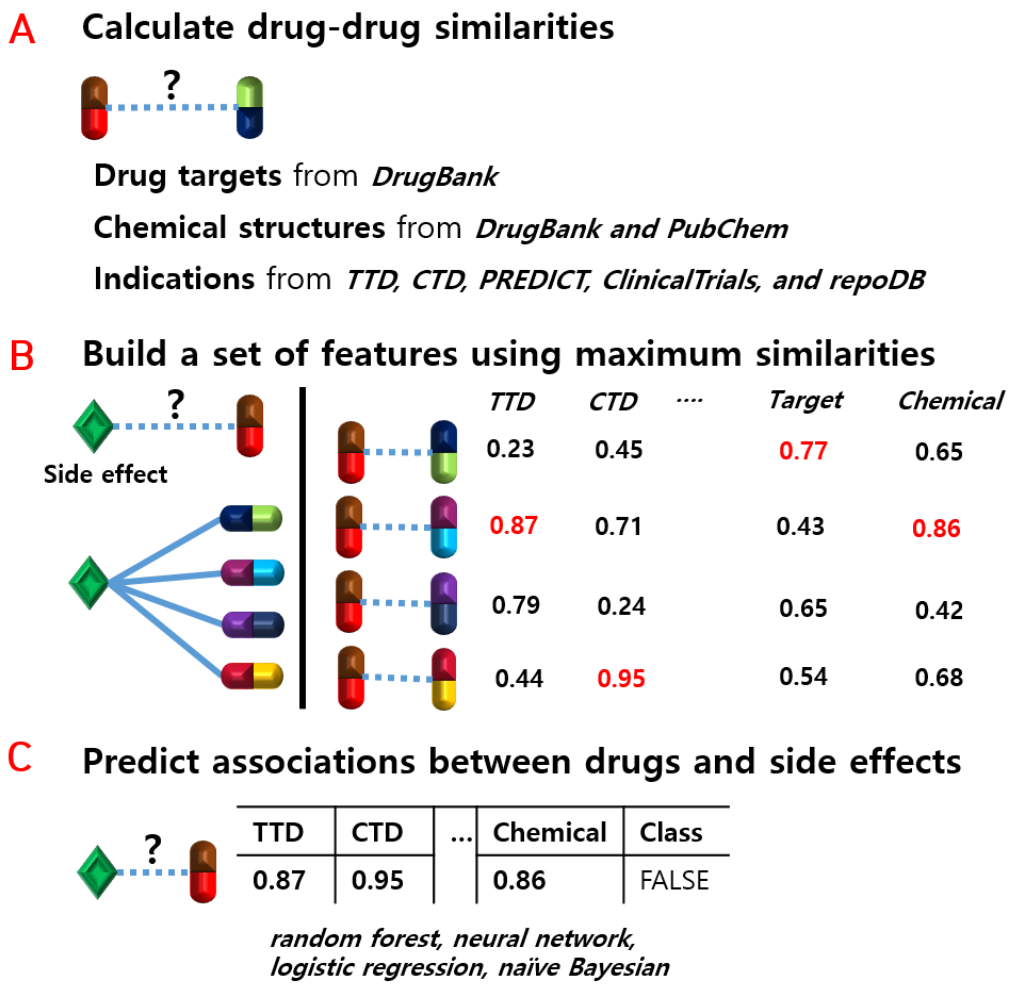


Fig. 1. System overview

3.1 Steps to predict drug side effect

We designed a method to predict the candidate side effects of drugs by leveraging the various similarities between them. While most existing studies had mainly concentrated on drugs' chemical and biological properties to identify their side effects, we added phenotypic information such as indications.

In step A, we assessed three types of similarities for all drug pairs using 1) indication databases individually, 2) targets, and 3) chemical structures. First, we computed the indication similarities between two drugs using the Jaccard coefficient, which is the number of traits in common divided by the union of the traits, as expressed by Eq. (1):

$$J(D_a, D_b) = \frac{|P_{Da} \cap P_{Db}|}{|P_{Da} \cup P_{Db}|} \quad (1)$$

For instance, let P_{Da} be the set of properties of drug a, which has five indications: I0001, I0002, I0003, I0004, and I0005. Let P_{Db} be the set of properties of drug b, which has five indications: I0003, I0005, I0006, I0007, and I0008. The number of unions and intersections of the two drugs are eight (I0001, I0002, I0003, I0004, I0005,

I0006, I0007, and I0008) and two (I0003 and I0005), respectively. Accordingly, the Jaccard coefficient of the two drugs is 0.25. Importantly, we did not merge the indications across databases, and calculated the drug-drug indication similarities for each database individually.

Moreover, the biological characteristics of drugs were considered using target information, and the target similarity between two drugs was computed using the coefficient mentioned above. For example, D_a had two drug targets: T0001 and T0002. Meanwhile, D_b had five drug targets: T0001, T0002, T0003, T0004, and T0005. The number of unions and intersections of the two drugs' targets were five (T0001, T0002, T0003, T0004, T0005) and two (T0001, T0002), respectively. Accordingly, the Jaccard coefficient for the two drugs' targets was 0.4.

Drugs with similar chemical structures traditionally exhibit similar biological activities [5]. Therefore, chemical properties play an important role in drug activity. OpenBabel calculates the number of bits in common divided by the union of the bits using drugs' fingerprints. We used OpenBabel to calculate the similarity for all drug pairs based on the Tanimoto coefficient [24].

In step B, we used the similarity values computed in the previous step to infer a drug D_a and side effect S_1 pair for each feature. First, we identified the drugs that had known associations with side effect S_1 from SIDER, and called these drugs D_b, D_c, D_d, D_e. For example, for TTD feature, we exploited the drug-drug similarities D_a-D_b (0.23), D_a-D_c (0.87), D_a-D_d (0.79) and D_a-D_e (0.44). Among them, we chose the maximum similarity value for TTD feature. This process was repeated for the remaining features. Moreover, we assigned a positive class if D_a and S_1 had an already known relationship, and a negative class if they did not have a known relationship.

In step C, we created a training set based on seven features using a variety of drugs similarities: the chemical structure, drug target, TTD, CTD, PREDICT, ClinicalTrials, and repoDB. Importantly, we created three different combinations of features to be able to distinguish the importance of including indication information: 1) Chemical+Target, 2) Indications only, and 3) Chemical+Target+Indications. There were some constraints to the creation of drug and side effect pairs for prediction. For a side effect and a drug to be considered (e.g., side effect S_1 and drug D_a), other drugs (e.g., drugs D_b, D_c, D_d, D_e) that had known associations with the side effect needed to have a similarity value with the D_a.

Finally, we applied diverse machine learning algorithms–including random forest, neural network, logistic regression, and naïve Bayesian models–to predict the candidate associations between drugs and side effects. This process was performed with the caret package in R set on default parameters [25].

3.2 Performance assessment

A 10-fold cross-validation was conducted to evaluate the performance of the method. A Positive set consisting of the known drug-side effect associations from SIDER was generated. Negative sets were also randomly generated from the drugs and side effects used in the positive set by excluding the known associations. The number of drug-side effect pairs was the same in the negative set as in the positive set. To avoid a negative samples bias, we extracted 100 negative sets by conducting replacement and repeat cross-validation 100 times. The average the area under the curve (AUC) scores were then calculated.

3.3 Validation of drugs' candidate side effects with FAERS and MedEffect

For each training set, we obtained average AUCs for each algorithm. Considering all the averaged AUCs, we chose the algorithm that showed the best results and applied a classifier to predict the candidate side effects of drugs. We used a remaining set that did not appear in the training set and predicted the candidate associations.

We used two different databases to validate the predictions: 1) FAERS and 2) MedEffect. These two databases contain the side effects of drugs reported by medical experts. There were two types of associations: 1) the predicted associations using our method and 2) the actual associations from FAERS and MedEffect. We created two contingency tables for FAERS and MedEffect with our predictions, as per Table 2. We applied Fisher's exact test (expressed as Eq. (2)) to check whether our candidate predictions significantly enriched the existing side effect databases. In this equation, the cells of the contingency table are given *x*, *y*, *z*, and *w*. Moreover, *n* is the grand total, and *P* represents the probability.

Table 2. Contingency table.

Predicted / ***Actual***	***True***	***False***	***Row total***
True	*x*	*y*	*x+y*
False	*z*	*w*	*z+w*
Column total	*x+z*	*y+w*	*n (=x+y+z+w)*

$$p = \frac{(x+y)!(z+w)!(x+z)!(y+w)!}{x!y!z!w!n!} \quad (2)$$

4. Results and discussion

4.1 Method performance

First, we verified that the use of drug indication information increased the predictive power to identify a drug's candidate side effects as compared with the use of target and chemical properties only. We calculated the similarities between drugs using the indications from the TTD, CTD, ClinicalTrials, PREDICT, and repoDB as well as the drug's target and chemical properties. We extracted a maximum of seven features for each drug-side effect pair.

Furthermore, we displayed the number of drugs, side effects, and pairs for each type of feature set, as shown in Table 3. For a drug and a side effect to be considered in the prediction process, there needed to be a similarity value between that drug and other drugs that had the side effect. Therefore, a different number of drugs and side effects was used for each type of feature set. Moreover, we built training sets by combining a positive set with negative sets. Positive set consisted of known associations from SIDER, while, negative sets were randomly generated from the drugs and side effects used in the positive set. We excluded the associations included in the

positive set from the negative sets. The positive and negative sets had to fulfill the constraints mentioned above. We randomly generated 100 negative sets and created training sets as described in Table 3. For each type of feature set, we performed 100 cross-validation runs and averaged the resulting AUC scores. In Tables 3 and 4, "+" represents an addition of features. For example, "Chemical+Target" means that both the chemical structures and targets of the drugs were used as features. Moreover, "Indications" means that five indication databases were used as individual features.

Table 3. Number of drugs, side effects, and pairs for each type of feature set.

Type of feature set	Number of drugs	Number of side effects	Number of pairs in positive set	Number of pairs in negative set	Number of pairs in training set
Chemical+Target	780	3,533	88,982	88,982	177,964
Indications	269	3,113	41,377	41,377	82,754
Chemical+Target+Indications	258	3,065	39,927	39,927	79,854

We adopted four machine learning algorithms, including a naïve Bayesian (NB), random forest (RF), logistic regression (LR), and neural network (NNET) model. We first tested the performance when using chemical and target features together, as had mainly been done in previous studies. We then used five indication features only, and added chemical and target features to them. The benefit of using indications is illustrated in Table 4.

Table 4. Averaged AUCs for 100 runs of 10-fold cross validation of four machine learning algorithms from our dataset

Type of feature set	NB	RF	LR	NNET
Chemical+Target	0.8611	0.8732	0.8584	0.8676
Indications	0.8808	0.9189	0.8733	0.8872
Chemical+Target+Indications	0.8950	**0.9344**	0.8921	0.9024

The results presented in Table 4 show that the use of indication information increased the AUCs for all machine learning algorithms. Among them, the random forest model increased the overall performance most. In the random forest model, the use of all feature types showed an approximately 7% increase in the AUC as compared with the use of chemical and target features only. Therefore, we concluded that the addition of indications improved the results.

4.2 Candidate predictions

Our method yielded the best results when using the random forest model with all the features included (chemical, target, and five indications), as shown in Table 4. Therefore, we chose the classifier that gave the highest AUCs from the random forest model to predict the candidate side effects of drugs. The training set for the model had 79,854 drug-side effect associations, which included 258 drugs and 3,065 side effects. We generated an entire set comprising 790,770 pairs from the above drugs and side effects. We excluded the pairs used in the training set and applied predicting algorithms to the remaining 710,916 pairs. Finally, we obtained the candidate

predictions for 55,653 associations, as shown in supplementary table 1 (S1). A drug could have multiple side effects. We compared the candidate predictions with the existing database of known associations to confirm that our predictions significantly enriched the existing databases. We performed Fisher's test using R, as mentioned in Eq. (2), and Table 2, and confirmed that the new predictions significantly enriched FAERS (one tailed p<2.2e-16 and odd ratio = 2.900271). The contingency table for FAERS and our predictions is reproduced in Table 5. Our predictions also significantly enriched MedEffect (p=1.939e-12 and odd ratio = 1.637783). Table 6 shows the contingency table for MedEffect and our predictions.

Table 5. Contingency table for FAERS and our predictions

Predictions / FAERS	True	False
True	31,289	189,090
False	14,471	253,638

Table 6. Contingency table for MedEffect and our predictions

Predictions / MedEffect	True	False
True	301	1,134
False	2,834	17,487

4.3 Comparison with previous studies

Liu used drug indications as one of several properties. His dataset consisted of six drug properties: substructures, targets, transporters, enzymes, pathways, and treatments. For a fair comparison, we changed Liu's dataset according to this study's method. First, we calculated the similarities for the six drug properties. Second, we selected the maximum drug-drug similarity values for each feature with regards to a drug and a side effect. Third, we created training sets with the six features and subjected them to the random forest algorithm. In this experiment, we only applied the random forest algorithm, as it yielded the best results (as shown in Table 4). After applying the same constraints as mentioned in the method section to Liu's dataset, we were able to use 159 drugs out of 832, and 961 side effects out of 1,385. The training set from Liu's dataset contained 14,049 positive and 14,049 negative associations. Liu had performed a 5-fold cross validation on his dataset, and we also applied a 5-fold cross-validation in this experiment. The results of our method and of others are presented in Table 7.

We compared our results with those reported by Liu's method and Zhang's method [26]. The latter had adopted a cost minimization method based on the linear neighborhood similarity methods (LNSMs). We applied Liu's 159 drugs and 961 side effects extracted using our method to an LNSM-CMI. The results of this experiment are given under the LNSM-CMI (159) column in Table 7. In particular, we adjusted one of two parameters of the LNSM-CMI. We empirically tested with 25, 50, 75, 100, and 125, and chose the N that showed the highest AUC: 100.

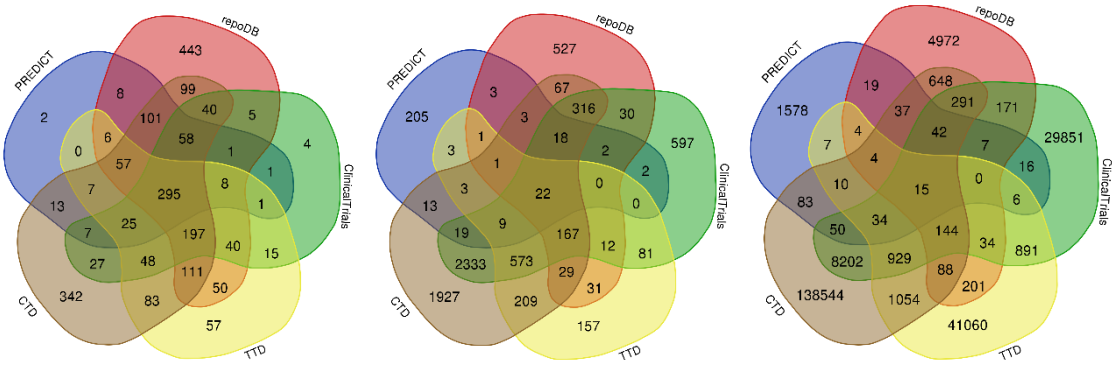
Table 7. Performances of the models using Liu's dataset.

Method	Our method (RF)	Liu's method	LNSM-CMI	LNSM-CMI (159)
AUC	**0.9187**	0.8850	0.9091	0.8967

Table 7 shows that our method yielded the best results for Liu's dataset. Even despite the smaller number of drugs and side effects, our method consistently outperformed others.

4.4 Display of five indication databases

We drew Venn diagrams for the indication databases used in this study to illustrate why we pulled out features from the indication databases individually. Fig. 2 shows that there were a lot of overlapping drugs, whereas there was a small number of overlapping indications between the databases. Moreover, the number of overlapping drug-indication associations was small. As it was more beneficial to leverage the indication traits from each database individually, we assumed that it would be more effective to use the databases as individual features rather than as a single indication feature.



a) Drugs b) Diseases c) Drug-indication associations

Fig. 2. Venn diagrams for the five indication databases

4.5 Case studies

Among 55,652 candidate predictions, 31,289 associations were verified by FAERS, and 301 associations were verified by MedEffect. To evaluate the practical benefits of our predictive classification models, we sampled diverse drugs according to the following criteria: 1) a recently approved drug based on a target-based rational drug design, 2) a long-term administrated drug whose side effects required more monitored, 3) a multi-targeted drug showing several differential indications, and 4) an old drug whose mechanism was imperfectly understood. We further examined them through a comparison with the academic literature and drug discovery materials.

First, dasatinib, a small-molecule multi-kinase inhibitor used to treat cancer, was selected as a recently approved drug based on a target-based rational drug design [27]. For this reason, there was relatively rich side effect and observational and interventional clinical data on the drug, and fast data updates for indication expansion were expected. Surprisingly, among the 181 predicted side effects of dasatinib, 11 matched the reported data in MedEffect, and 141 matched the reported data in FAERS. Accordingly, 54% of the predicted side effects of

dasatinib were found in the clinical data. Moreover, we also found literary evidence for some predicted side effects that were present in both MedEffect and FAERS. Angioedema is a serious skin adverse drug reaction that can be caused by dasatinib [28]. According to a study by Reyes-Habito *et al.*, dasatinib commonly causes skin reactions such as pruritus, acne, and xerosis [29]. A symptom of "blood alkaline phosphatase increased" can also be caused by dasatinib [30].

Second, sitagliptin, a diabetes therapeutics that act as a selective dipeptidyl peptidase-4 inhibitor, was chosen as a good representative of long-term care medication drugs [31]. As long-term administration is needed, it requires more side effect monitoring short-term administration drugs. In addition, sitagliptin also represents a very important new drug class for diabetes, i.e. with a clear mode of action and a rational drug design. In order to develop combination therapy drugs applicable to metabolic syndrome from other drugs and sitagliptin, the landscape of its side effects is essential. Among the 171 predicted side effects of sitagliptin, 145 matched the data reported in FAERS. Thus, 61% of the sitagliptin predictions were found in the clinical data.

Third, vorinostat, a multi-targeting drug was chosen for its very diverse indications. It is used to treat a range of diseases, from HIV infection to diverse types of cancers, including non-small cell lung cancer, ovarian cancer, breast cancer, and pancreatic cancer. As vorinostat is a target-based anti-cancer drug, its side effect mechanisms are more reasonably traceable than those cytotoxic agents. When considering the risk to benefit ratio, the side effects of anti-cancer drugs can be less important than those of other therapeutic drugs. However, recent cancer treatment tends to consider the quality of life as well as the survival rate. Among the 13 predicted side effects of vorinostat, 11 were identified in the FAERS clinical data. Thus, 85% of our predictions matched the clinically reported side effects of FAERS, proving that our method could powerfully be used to identify new side effects from the whole warning black box of drugs.

Finally, clonidine was chosen for having an unclear mechanism despite being an old drug. Although our insufficient understanding of its on-target effects could not be used to explain its side effects, the accumulated records of both diverse indications and side effects were sufficient for comparison with our predictions. Among the 228 side effects identified for clonidine, 173 side effects were confirmed by FAERS. Therefore, 75% of the predictions were relevant.

Thus, we showed that most of the predictions could be found in the clinically reported data. In addition, some unprecedented side effects that had not been reported but were suggested by in our predictions will need to be clinically validated in the near future.

5. Conclusion

This study inferred the candidate side effects of drugs using the existing knowledge on them. Unlike previous studies that had mainly focused on the biological and chemical characteristics of drugs, the present one used the indications from a variety of databases. As many drug repositioning studies had used side effects to explain drug activities, we assumed that drugs with similar indications would show similar side effects. As each database contained different indications for a given drug, it was more beneficial to leverage the indication traits contained in each database individually. As a result, the performance was 7% better than when using target and chemical features only.

Our method presented a limitation in that we could not identify the candidate side effects of drugs with insufficient indication information. For example, we could not predict the candidate side effects of a drug if at least one of the databases did not contain information about it. However, as the indication knowledge on drugs is growing fast, the potential indications of many drugs featured in drug repositioning studies will become readily applicable to our study.

Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT & Future Planning) (NRF-2018R1A2B6006223).

References

[1] Khanna, Ish. "Drug discovery in pharmaceutical industry: productivity challenges and trends." Drug discovery today 17.19-20 (2012): 1088-1102.

[2] Horrobin, David F. "Innovation in the pharmaceutical industry." Journal of the Royal Society of Medicine 93.7 (2000): 341-345.

[3] Pauwels, Edouard, Véronique Stoven, and Yoshihiro Yamanishi. "Predicting drug side effect profiles: a chemical fragment-based approach." BMC bioinformatics 12.1 (2011): 169.

[4] Mizutani, Sayaka, et al. "Relating drug–protein interaction network with drug side effects." Bioinformatics 28.18 (2012): i522-i528.

[5] Martin, Yvonne C., James L. Kofron, and Linda M. Traphagen. "Do structurally similar molecules have similar biological activity?." Journal of medicinal chemistry 45.19 (2002): 4350-4358.

[6] Johnson, A., and M. Maggiora Wiley-Interscience. "Concepts and Applications of Molecular Similarity. Edited." (1991): 3409.

[7] Ye, Hao, Qi Liu, and Jia Wei. "Construction of drug network based on side effects and its application for drug repositioning." PloS one 9.2 (2014): e87864.

[8] Gottlieb, Assaf, et al. "PREDICT: a method for inferring novel drug indications with application to personalized medicine." Molecular systems biology 7.1 (2011): 496.

[9] Campillos, Monica, et al. "Drug target identification using side effect similarity." Science 321.5886 (2008): 263-266.

[10] Liu, Mei, et al. "Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs." Journal of the American Medical Informatics Association 19.e1 (2012): e28-e35.

[11] Zhang, Wen, et al. "Predicting drug side effects by multi-label learning and ensemble learning." BMC bioinformatics 16.1 (2015): 365.

[12] Muñoz, Emir, Vít Nováček, and Pierre-Yves Vandenbussche. "Using drug similarities for discovery of possible adverse reactions." AMIA Annual Symposium Proceedings. Vol. 2016. American Medical Informatics Association, 2016.

[13] Sam, Elizabeth, and Prashanth Athri. "Web-based drug repurposing tools: a survey." Briefings in bioinformatics (2017).

[14] Brown, Adam S., and Chirag J. Patel. "A standard database for drug repositioning." Scientific data 4 (2017): 170029.

[15] Li, Ying Hong, et al. "Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics." Nucleic acids research 46.D1 (2017): D1121-D1127.

[16] Davis, Allan Peter, et al. "Chemical-induced phenotypes at CTD help inform the pre-disease state and construct adverse outcome pathways." Toxicological Sciences (2018): kfy131.

[17] Gottlieb, Assaf, et al. "PREDICT: a method for inferring novel drug indications with application to personalized medicine." Molecular systems biology 7.1 (2011): 496.

[18] Law, Vivian, et al. "DrugBank 4.0: shedding new light on drug metabolism." Nucleic acids research 42.D1 (2013): D1091-D1097.

[19] Hamosh, Ada, et al. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic acids research 33.suppl_1 (2005): D514-D517.

[20] Kim, Sunghwan, et al. "PubChem substance and compound databases." Nucleic acids research 44.D1 (2015): D1202-D1213.

[21] Imanishi, Tadashi, and Hajime Nakaoka. "Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases." Nucleic acids research 37.suppl_2 (2009): W17-W22.

[22] The food and drug administration adverse event reporting system (FAERS). http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects.

[23] Osborne, Carol-anne. “Adverse Drug Reactions: Investigating to Reporting.” Journal of the Canadian Academy of Child and Adolescent Psychiatry 19.1 (2010): 46–47. Print.

[24] O'Boyle, Noel M., et al. "Open Babel: An open chemical toolbox." Journal of cheminformatics 3.1 (2011): 33.

[25] Kuhn, Max. "Caret package." Journal of statistical software 28.5 (2008): 1-26.

[26] Zhang, Wen, et al. "Drug side effect prediction through linear neighborhoods and multiple data source integration." Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on. IEEE, 2016.

[27] Ling, Yuan, et al. "Protein kinase inhibitors for acute leukemia." Biomarker research 6.1 (2018): 8.

[28] Faye, Emmanuelle, et al. "Spontaneous reporting of serious cutaneous reactions with protein kinase inhibitors." European journal of clinical pharmacology 69.10 (2013): 1819-1826.

[29] Reyes-Habito, Claire Marie, and Ellen K. Roh. "Cutaneous reactions to chemotherapeutic drugs and targeted therapy for cancer: Part II. Targeted therapy." Journal of the American Academy of Dermatology 71.2 (2014): 217-e1.

[30] Tibullo, Daniele, et al. "Effects of second- generation tyrosine kinase inhibitors towards osteogenic differentiation of human mesenchymal cells of healthy donors." Hematological oncology 30.1 (2012): 27-33.

[31] Hanefeld, Markolf, et al. "Once-daily sitagliptin, a dipeptidyl peptidase-4 inhibitor, for the treatment of patients with type 2 diabetes." Current medical research and opinion 23.6 (2007): 1329-1339.

Supplementary Material

Click here to download Supplementary Material: supplementary_table1.xlsx