Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs

Mei Liu,[1] Yonghui Wu,[1] Yukun Chen,[1] Jingchun Sun,[1] Zhongming Zhao,[1] Xue-wen Chen,[2,3] Michael Edwin Matheny,[1,4,5,6] Hua Xu[1]

► An additional table is published online only. To view this file please visit the journal online (www.jamia.bmj.com/content/19/e1.toc).

[1]Department of Biomedical Informatics, Vanderbilt University, School of Medicine, Nashville, Tennessee, USA
[2]Bioinformatics and Computational Life Sciences Laboratory, Information and Telecommunication Technology Center, University of Kansas, Lawrence, Kansas, USA
[3]Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, Kansas, USA
[4]Department of Biostatistics, Vanderbilt University, School of Medicine, Nashville, Tennessee, USA
[5]Division of General Internal Medicine, Vanderbilt University, School of Medicine, Nashville, Tennessee, USA
[6]Geriatric Research Education and Clinical Care, Veterans Health Administration, Nashville, Tennessee, USA

Correspondence to
Dr Hua Xu, Department of Biomedical Informatics, Vanderbilt University, School of Medicine, 2209 Garland Ave, EBL 412, Nashville, TN 37232, USA; hua.xu@vanderbilt.edu

ML and YW contributed equally to this study.

Received 14 November 2011
Accepted 30 March 2012

ABSTRACT

Objective Adverse drug reaction (ADR) is one of the major causes of failure in drug development. Severe ADRs that go undetected until the post-marketing phase of a drug often lead to patient morbidity. Accurate prediction of potential ADRs is required in the entire life cycle of a drug, including early stages of drug design, different phases of clinical trials, and post-marketing surveillance.

Methods Many studies have utilized either chemical structures or molecular pathways of the drugs to predict ADRs. Here, the authors propose a machine-learning-based approach for ADR prediction by integrating the phenotypic characteristics of a drug, including indications and other known ADRs, with the drug's chemical structures and biological properties, including protein targets and pathway information. A large-scale study was conducted to predict 1385 known ADRs of 832 approved drugs, and five machine-learning algorithms for this task were compared.

Results This evaluation, based on a fivefold cross-validation, showed that the support vector machine algorithm outperformed the others. Of the three types of information, phenotypic data were the most informative for ADR prediction. When biological and phenotypic features were added to the baseline chemical information, the ADR prediction model achieved significant improvements in area under the curve (from 0.9054 to 0.9524), precision (from 43.37% to 66.17%), and recall (from 49.25% to 63.06%). Most importantly, the proposed model successfully predicted the ADRs associated with withdrawal of rofecoxib and cerivastatin.

Conclusion The results suggest that phenotypic information on drugs is valuable for ADR prediction. Moreover, they demonstrate that different models that combine chemical, biological, or phenotypic information can be built from approved drugs, and they have the potential to detect clinically important ADRs in both preclinical and post-marketing phases.

INTRODUCTION

The US public spends billions of dollars on prescription drugs every year, resulting in a significant healthcare burden from adverse drug reactions (ADRs). ADRs are defined as those unintended and undesired responses to drugs beyond their anticipated therapeutic effects during clinical use at normal doses.[1] It is estimated that 6–7% of hospitalized patients experience severe ADRs each year with a potential of 100 000 deaths, which makes it the fourth largest cause of death in the USA.[2] Within the past 10 years, both reported ADRs and related deaths have increased ~2.6 times and led to a number of drug withdrawals, with rofecoxib (Vioxx) and cerivastatin (Baycol) among the most prominent examples.[3 4] Therefore, it is extremely important to predict and monitor a drug's ADRs throughout its life cycle, from preclinical screening phase to post-market surveillance.

The fundamental method for predicting or assessing potential ADRs early in the drug development pipeline is the application of preclinical in vitro safety profiling by testing compounds with biochemical and cellular assays.[5] However, experimental detection of ADRs using extensive in vitro safety pharmacology profiling remains challenging in terms of cost and efficiency.[5] For post-market surveillance, it often relies on public databases containing ADR reports voluntarily submitted by physicians,[6–15] which take time to accumulate before a signal can be detected. Recently, a large amount of effort has been devoted to developing in silico approaches to predict ADRs using available large public datasets of drugs, at both preclinical[16] and post-market[17] stages. Most of these methods have used either chemical structure or protein target information on drugs to build the prediction models, and some have shown promising results.[18–27]

In this study, we proposed a new drug surveillance framework by investigating three types of information for ADR prediction: (1) chemical properties such as compound fingerprints or substructures; (2) biological properties including protein targets and pathways; and (3) phenotypic properties including indications and other known ADRs if available. Our evaluation showed that the phenotypic information (when available) largely improved the performance of ADR prediction models. The framework suggests an efficient way to optimize ADR prediction by combining different types of information at the different stages of drug surveillance (eg, 'chemical + biological' for preclinical drug screening and 'chemical + biological + phenotypic' for post-market surveillance).

Background

A number of computational methods have been developed to predict potential ADRs from preclinical characteristics of the compounds or screening data and post-marketing evidence. Existing efforts to predict ADRs from preclinical data can be categorized into protein-target-based

and chemical-structure-based approaches. The underlying principle of the protein-target-based approach is that drugs with similar in vitro protein-binding profiles tend to exhibit similar side effects.[18] Scheiber *et al*[20] demonstrated the concept by comparing pathways affected by toxic compounds versus those affected by non-toxic compounds. Fukuzaki *et al*[21] proposed a method to predict ADRs using sub-pathways that share correlated modifications of gene-expression profiles in the presence of the drug of interest. However, their work depends on the availability of gene-expression data observed under chemical perturbations by the drug. Xie *et al*[22] developed a chemical systems biology approach to identify off-targets of a drug by docking the drug into binding pockets of proteins that are similar to its primary target. Then the drug–protein interaction pair with the best docking score was mapped to known biological pathways to identify potential off-target binding networks of the drug. However, scalability of the method is hindered by its requirement for protein three-dimensional structures and known biological pathways.

Alternatively, the chemical-structure-based approach attempts to link ADRs to their chemical structures. As a proof-of-concept, Bender *et al*[23] explored the chemical space of drugs and established its correlation for ADR prediction. Scheiber *et al*[24] presented a global analysis that identified chemical substructures associated with ADRs, but the method was not designed to predict ADRs for any specific drug molecule. Yamanishi *et al*[25] proposed a method that predicted pharmacological effects from chemical structures and then used the effect similarity to infer drug–target interactions. Hammann *et al*[26] employed decision tree modeling to determine the chemical, physical, and structural properties of compounds that predispose them to causing ADRs. Notably, ADR-predictive models developed on preclinical characteristics could provide additional evidence to support potential signals from post-marketing surveillance. For example, a recent study by Pouliot *et al*[17] utilized screening data from the PubChem BioAssay[28] database to determine the correlation of post-marketing ADRs with drug bioactivity across vast BioAssay screens. However, most of these methods were not designed to predict high-dimensional side-effect profiles for drugs. In order to accomplish this goal, Pauwels *et al*[27] developed a sparse canonical correlation analysis method to predict high-dimensional side-effect profiles of drug molecules based on their chemical structures.

Despite the success of using chemical and biological information of drugs for ADR prediction, few studies have investigated the use of phenotypic information (eg, indication and other known ADRs). Existing resources, such as the SIDER[29] (Side Effect Resource) database, contain comprehensive drug phenotypic information such as indications and known ADRs. Such phenotypic information has been demonstrated to be useful for other drug-related studies. For example, Campillos *et al*[19] identified new drug targets by comparing the similarity of side effects of drugs. Here, we propose to investigate the use of phenotypic information on drugs, together with chemical and biological properties, to predict ADRs. Similarly to the work by Pauwels *et al*,[27] we conducted a large-scale study to develop and validate the ADR prediction model using 1385 known ADRs for 832 FDA (US Food and Drug Administration)-approved drugs in SIDER[29] using various machine learning (ML) algorithms. In addition, we comprehensively evaluated different combinations of features to see how each feature set contributes to prediction accuracy. Our experimental results show that integration of chemical, biological, and phenotypic properties outperformed the chemical-structured-based method and has the potential to detect clinically important ADRs at both preclinical and post-market phases for drug surveillance.

METHODS

Data description

To build and evaluate the proposed ADR-prediction model, we used data from SIDER.[29] SIDER presents an aggregate of dispersed public information on drug side effects and indications. SIDER extracted information on marketed medicines and their recorded ADRs from public documents and package inserts, which resulted in a collection of 888 drugs and 1385 side-effect keywords. There are a total of 61 102 associations between drugs and side-effect terms in SIDER, and each drug has an average of 68.8 side effects.

The chemical structures of drugs were collected from PubChem,[30 31] biological properties were obtained from the DrugBank[32–34] and KEGG,[35–37] and phenotypic data were from SIDER.[29] To link these databases, we mapped drugs in SIDER to DrugBank.[32–34] Fifty-six drug names from SIDER could not be mapped to their respective DrugBank IDs, resulting in a final dataset of 832 drugs, each of which has a 'Yes' or 'No' label for each of the 1385 side effects, indicating whether a drug has a specific side effect or not.

The PubChem, DrugBank, and KEGG databases comprise data that are available during chemical and animal trials, and are available before or during phase I clinical trials. However, the phenotypic data from SIDER are collected from phase I all the way through phase IV post-marketing surveillance. As such, this work describes a surveillance framework that allows pre-human association detection all the way through pre-marketing clinical trial phases to post-marketing surveillance. Figure 1 provides a visualization of the proposed ADR-prediction framework at different phases of drug surveillance.

Features

Each drug is associated with a 1385 dimensional binary side-effect profile, y, whose elements correspond to the presence or absence of each of the side-effect concepts with 1 or 0, respectively. Each drug is also associated with three types of feature: chemical, biological, and phenotypic properties. Table 1 shows the subgroups of each feature type, its source, and dimension. To encode the drug's chemical structure, we used fingerprints corresponding to 881 chemical substructures defined in PubChem.[30 31] The biological properties consisted of drug

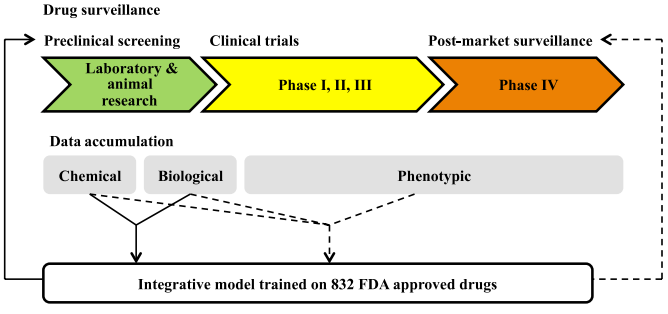

Figure 1 Overview of the proposed framework for drug surveillance. Different combinations of features can be used for different phases of drug surveillance. Chemical structures and relevant proteins of drugs can be combined to predict potential adverse drug reactions (ADRs) in the early phase of drug development. As drug indication and other ADRs become available, they can be integrated with chemical and biological information for post-market surveillance.

Table 1 Data features integrated in this study

Feature type	Specific feature	Source	Dimension
Chemical	Substructures	PubChem	881
Biological	Targets	DrugBank	786
	Transporters	DrugBank	72
	Enzymes	DrugBank	111
	Pathways	KEGG	173
Phenotypic	Treatment indications	SIDER	869
	Other side effects	SIDER	1384

protein targets, transporters (for drug transportation), enzymes (for drug metabolism), and derived pathway information from the protein targets. Information on the protein targets, transporters, and enzymes of a given drug was directly obtained from DrugBank.[32–34] Each drug target was then mapped to the corresponding KEGG pathway[35–37] through its protein-coding gene symbol. The phenotypic information included indications and other known ADRs of drugs. Both sets of data were obtained directly from SIDER. Therefore, for a particular ADR y_i, each drug is represented by its chemical, biological, and phenotypic properties as a 4276 (881+1142+2253) dimensional vector in which each element is either 1 or 0, respectively, for the presence or absence of each PubChem substructure, drug target, transporter, enzyme, KEGG pathway, indication, and remaining known ADRs.

Experimental design

In this study, we treat the ADR-prediction task as a classic binary classification problem where each drug either causes or does not cause a particular ADR. For each ADR, we built a classifier and evaluated its performance using 832 drugs as samples. We then repeated the process for each of the 1385 ADRs and summarized performance across all ADRs.

Evaluation was designed from different angles. First, we assessed the contributions of each feature type and their combinations to ADR prediction, using a fixed algorithm, support vector machine (SVM). Next, we compared the performance of five ML algorithms in predicting ADRs using an optimized feature set. Owing to the abundant variance of ADRs, we suspected that common ADRs (with more positive samples) might behave differently. Therefore we defined a subset of common ADRs, which were ADRs associated with more than 50 of the 832 drugs (denoted as 'ADR_50+'). We evaluated the performance of these ADRs separately and compared it with the performance of all ADRs.

ML algorithms

Five ML algorithms—logistic regression (LR), naïve Bayes (NB), K-nearest neighbor (KNN), random forest (RF), and SVM—were investigated for the prediction task. To build the LR model, we used the L2-regularized logistic regression solver in LIBLINEAR.[38] An object-oriented Matlab(R) ML package called CLOP[39] was used to implement the NB classifier. The popular ML software, WEKA,[40] was used for the KNN and RF modeling. Lastly, LIBSVM[41] was applied as the SVM learner for prediction.

Evaluation

Model evaluation

For each ADR, a classifier was built and evaluated using a five-fold cross-validation on 832 drugs. As a consequence, n classifiers will be constructed for n side effects where n is 1385. Performance of the proposed method was assessed by a receiver operating characteristic (ROC) curve, which is a graphical plot of sensitivity or true positive rate against false positive rate (1 − specificity). Sensitivity is defined as the proportion of actual positives that are correctly identified as such (ie, SN = TP/(TP +FN)), and specificity measures the proportion of actual negatives that are correctly predicted as such (ie, SP = TN/(TN +FP)), where FN is false negative, FP is false positive, SN is sensitivity, SP is specificity, TP is true positive, and TN is true negative. The ROC curve can be plotted by varying threshold values for prediction scores above which the output is predicted as positive and negative otherwise.

Area under the ROC curve (AUC), accuracy, precision, and recall were calculated as well. AUC provides a single measurement of the performance of a ROC curve. Accuracy (ACC) is the proportion of true results obtained (ie, ACC = (TP + TN)/(TP + FP + FN + TN)). Precision (P) is defined as the proportion of true positives against all predicted positive results (ie, P = TP/(TP+FP)). Recall is also known as the true positive rate or sensitivity, which is defined above.

To summarize the global performance across 1385 ADRs, there are two possible approaches. One can compute an evaluation measure for each ADR and then average the measures over all ADRs to obtain an overall score, which is called macro-averaging. Another approach is to merge the prediction scores for all drugs over all ADRs, and then compute the overall measure, which is referred to as micro-averaging. The study by Pauwels *et al*[27] reported a global AUC across all ADRs by merging the prediction scores for all ADRs into one big matrix and drawing a global ROC curve from the matrix, which is a similar approach to micro-averaging. Here, we followed their approach to generate the global AUC and accuracy. In addition, we reported micro-averaging precision and recall. The reported accuracy, precision, and recall were obtained from the best cut-off points or operating points of the global ROC curve, so that it gives the best tradeoff between false positives and false negatives.

Statistical significance test

In order to assess whether the improvement in performance by adding feature spaces to the baseline chemical space is significant, the two-sample Kolmogorov–Smirnov test (KS test)[42 43] was computed. The two-sample KS test is a general non-parametric method for comparing two samples to test whether the two underlying probability distributions differ. We calculated the KS test over the AUC scores generated by different feature sets for each ADR. For example, in the case of comparing the baseline chemical space 'chem' with the combined set 'chem +bio', a set of AUC scores is generated for predicting each of the 1385 ADRs using each feature set, and then the KS test assesses if the AUC scores generated by 'chem+bio' are stochastically larger than the scores generated by 'chem'. Finally, since we were making multiple comparisons for different feature pairs, the p values from the KS test were corrected by Bonferroni correction.[44]

Clinical validation

To demonstrate the clinical significance of the proposed model, we evaluated the model's ability to predict post-market ADRs that caused the withdrawals of cerivastatin (Baycol) and rofecoxib (Vioxx). Cerivastatin is a statin used to lower cholesterol and prevent cardiovascular disease and was voluntarily withdrawn from the market in 2001 because of reports of fatal rhabdomyolysis. Rofecoxib is a non-steroidal anti-inflammatory drug used to treat osteoarthritis, acute pain conditions, and dysmenorrhea, and was withdrawn in 2004 over safety concerns

about increased risk of heart attack. A physician manually reviewed both drugs' ADRs in SIDER and identified seven ADRs related to rhabdomyolysis for Baycol and four ADRs related to heart attack for Vioxx (see table 4). For each of the 11 ADRs, we built a prediction model based on the remaining drugs and applied it to either Baycol or Vioxx. To compare the effect of different feature sets, we reported the prediction results for 'chem', 'chem+bio', and 'chem+bio+pheno'. As these seven ADRs related to rhabdomyolysis correlated highly and the use of the other six ADRs as features to predict the remaining ADR may make the task easier, we created a higher-level ADR for rhabdomyolysis by grouping all seven ADRs into one (the same applies for heart attack). We then built the prediction models and reported the performance of the grouped ADRs for rhabdomyolysis as well as for heart attack.

RESULTS

Feature assessment

First, we assessed the abilities of different feature combinations to predict known side effects using SVM through a fivefold cross-validation with chemical structures as the baseline feature. To conduct a fair and accurate comparison across different feature sets, the same experimental conditions were maintained by using the same training drugs and test drugs for each fold. SVM parameters were empirically optimized using the AUC as an objective function. The best results for SVM were obtained by a Radial Basis Function (RBF) kernel with kernel parameter $g = 0.008$ and penalty parameter $C = 2$. When chemical structure alone was adopted, the best resulting AUC was 0.9054, which is similar to the finding (AUC = 0.8930) of Pauwels *et al*.[27] Figure 2 shows the ROC curves for different feature sets based on cross-validation experiments, and table 2 summarizes the evaluation results.

When the feature spaces were compared independently (table 2), the phenotypic features appeared to be the most informative (highest AUC of 0.9542), and 'chem' and 'bio' achieved similar AUC. Adding biological features on top of chemical structures improved AUC slightly (from 0.9054 to 0.9098), whereas the increase obtained by adding phenotypic features was dramatic (from 0.9054 to 0.9526). When all three levels of features were combined ('chem+bio+pheno'), the performance was almost the same as the 'chem+pheno' or 'pheno' alone. For example, the ROC curves of 'chem+pheno' and 'chem+bio+pheno' in figure 2 almost overlap. On the other hand, if we focus on precision and recall, the improvement by adding biological features was more obvious (~3% in precision and ~1% in recall). Adding the phenotypic features yielded much larger increases, with ~21% in precision and ~15% in recall. Statistical analysis using the KS test[42 43] showed that the improvement in AUC was significant for the addition of biological features to the chemical features (p = 1.45E-07), as well as for the addition of biological and phenotypic features to the chemical features (p = 1.10E-15). Compared with 'pheno' alone, the addition of 'chem' and 'bio' produced a reduction in the global AUC; however, the reduction was not statistically significant according to the KS test (p=0.177).

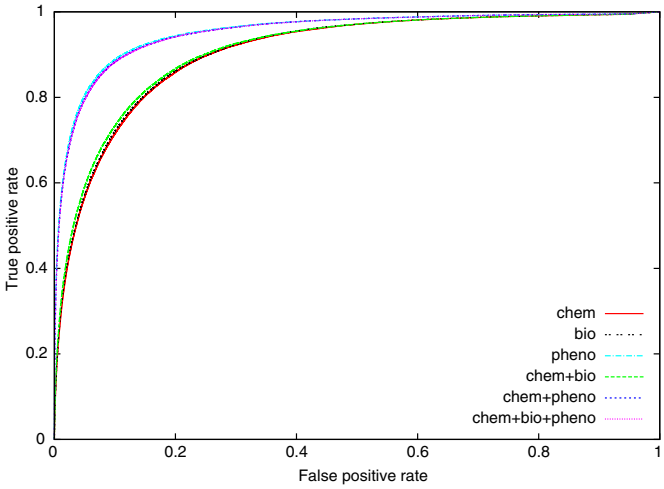

Figure 2 Receiver operating characteristic curves in fivefold cross-validation for various feature sets using support vector machine: (1) chemical structures, 'chem'; (2) biological properties, 'bio'; (3) phenotypic properties, 'pheno'; (4) chemical and biological properties, 'chem+bio'; (5) chemical and phenotypic properties, 'chem+pheno'; (6) chemical, biological, and phenotypic properties, 'chem+bio+pheno'.

The resulting ROC curves of the common ADRs (ie, ADR_50+) are shown in figure 3, and corresponding results are summarized in table 2. When compared with the results of all ADRs, a decrease in AUC and accuracy was observed as expected because rare ADRs that may distort the measures were excluded from the calculation. Thus in figure 3, there are larger separations between the ROC curves. For instance, when all ADRs were used in the calculation, the biological properties only increased the AUC by 0.004, but when we only considered the common ADRs, the increment was 0.02.

Method comparison

We compared the abilities of five ML algorithms—LR, NB, KNN, SVM, and RF—to predict known side effects of drugs by a five-fold cross-validation using all chemical, biological, and phenotypic properties as the feature set. Parameters for all classifiers presented here were empirically optimized using the AUC score. The best result for LR was obtained with parameters $C = 10$ and *epsilon* = 1, and for KNN the optimized number of neighbors is

Table 2 Feature comparison—performance of SVM over all versus common ADRs

	ADR_All				ADR_50+			
Feature set	**AUC**	**ACC**	**Precision**	**Recall**	**AUC**	**ACC**	**Precision**	**Recall**
Chem	0.9054	0.9538	0.4337	0.4925	0.7659	0.8268	0.4539	0.5569
Bio	0.9069	0.9543	0.4324	0.5043	0.7729	0.8287	0.4666	0.5521
Pheno	0.9542	0.9678	0.6607	0.6460	0.9175	0.8891	0.6933	0.7142
Chem+bio	0.9098	0.9551	0.4623	0.5008	0.7849	0.8327	0.4776	0.5728
Chem+pheno	0.9526	0.9669	0.6488	0.6443	0.9141	0.8857	0.6757	0.7215
Chem+bio+pheno	0.9524	0.9669	0.6617	0.6306	0.9138	0.8856	0.6750	0.7227

ADR_All considers all ADRs and ADR_50+ are the common ADRs caused by at least 50 drugs. All AUC, ACC, Precision, and Recall are micro-averages across ADRs in the corresponding dataset.
ACC, accuracy; ADR, adverse drug reaction; AUC, area under the receiver operating characteristic curve; Bio, biological property; Chem, chemical structure; Pheno, phenotypic property; SVM, support vector machine.

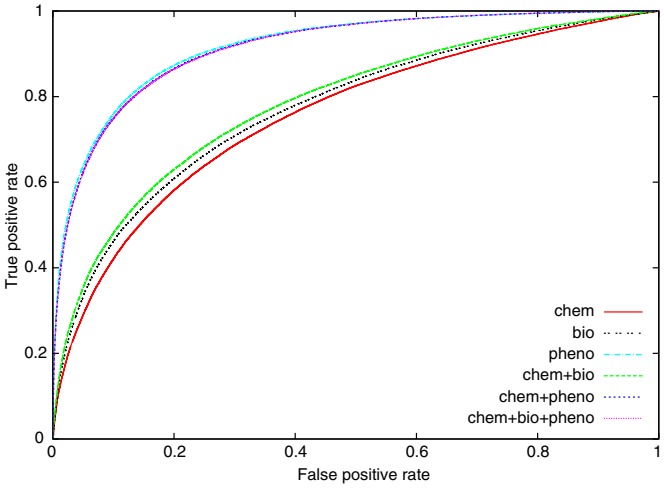

Figure 3 Receiver operating characteristic curves in fivefold cross-validation on various feature sets for common adverse drug reactions using support vector machine: (1) chemical structures, 'chem'; (2) biological properties, 'bio'; (3) phenotypic properties, 'pheno'; (4) chemical and biological properties, 'chem+bio'; (5) chemical and phenotypic properties, 'chem+pheno'; (6) chemical, biological, and phenotypic properties, 'chem+bio+pheno'.

$k = 55$. For RF, we grew 100 decision trees in each ensemble. ROC curves of the five methods are shown in figure 4. AUC and accuracy over all ADRs versus the common ADRs are summarized in table 3.

As shown in figure 4, SVM performed the best followed by RF, KNN, NB, and LR. For LR, NB, and KNN, the AUC score is almost the same when calculated across all ADRs, but diverges greatly when calculated across the common ADRs. Nevertheless, all measures of RF and SVM outperform others by a large margin. Although over all ADRs, AUC scores of SVM and RF are almost the same, SVM produced a higher precision of 66.17% and recall of 63.06% compared with RF (63.10% for precision and 62.50% for recall).

Clinical validation examples

Table 4 shows the prediction results on ADRs related to rhabdomyolysis for Baycol and heart attack for Vioxx. Prediction performance was in the order 'chem' < 'chem+bio' < 'chem+bio+pheno'. The classifiers based on 'chem' detected only one ADR related to rhabdomyolysis and none for heart attack. The classifiers based on 'chem+bio' detected five of seven rhabdomyolysis-related ADRs, but none for heart attack. For the classifiers using all features, five of seven rhabdomyolysis-related ADRs and two of four heart attack-related ADRs were predicted successfully. For the two grouped ADRs for rhabdomyolysis and heart attack, all classifiers predicted them successfully, which was probably due to increased sample sizes after grouping.

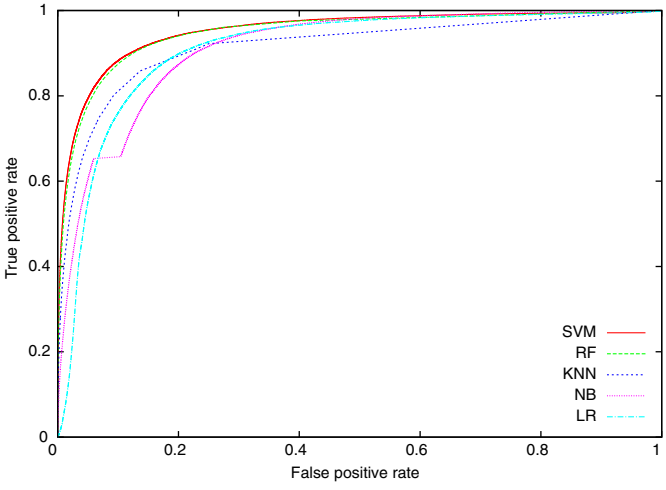

Figure 4 Receiver operating characteristic curves for method comparison. KNN, K-nearest neighbor; LR, logistic regression; NB, naïve Bayes; RF, random forest; SVM, support vector machine.

DISCUSSION

In this study, we conducted a large-scale ADR prediction of FDA-approved drugs and investigated three types of feature: (1) chemical structures; (2) biological properties—protein targets, transporters, enzymes, and pathways; (3) phenotypic characteristics—indication and other known ADRs. Our evaluation showed that drug phenotypic information (when available) is informative for ADR prediction, indicating its potential use for early detection of post-market ADR signals. In addition, our study demonstrated that the combination of chemical and biological features improved the AUC as well as precision (~3% increase) and recall (~1%), suggesting that such a data fusion approach is promising for preclinical screening of potential ADRs. The combination of all three types of information ('chem+bio+pheno') had lower global AUC than the 'pheno'-only classifier (but this was not statistically significant), indicating that the simple feature combination method may not work well in this case. We then compared the true positive predictions by classifiers that used individual feature sets ('chem', 'bio', or 'pheno') and measured the overlap between each pair of classifiers. As shown in figure 5, 5072 ADRs were detected by 'chem' or 'bio' but not by 'pheno', and 10 581 ADRs were detected by 'pheno' but not by 'chem' or 'bio', indicating that ADRs predicted by each feature type are complementary, and higher performance could be achieved through development of more advanced methods for feature integration. We further analyzed the significance of associations between each of the 4276 features and each of the 1385 ADRs using χ^2 statistics in which a feature is regarded as informative if the $p<0.05$. Distribution of the informative features is shown in online supplementary table S1.

During revision of this paper, Cami *et al*[45] published a similar study, where they proposed an integrative approach for predicting new ADRs by utilizing structure attributes of the network formed by known drug–ADR relationships from drug safety data, as well as specific drug information including Anatomical Therapeutic Chemical taxonomy, molecular descriptors, and *Medical Dictionary for Regulatory Activities* (MedDRA) taxonomy of adverse events. Thus we believe that the models built on large-scale approved drugs have the potential to detect clinically important ADRs at both preclinical and post-market phases for new drugs.

In a further analysis, we found that the contribution of phenotypic features was mostly due to other known ADRs rather than indications. A major reason that existing ADRs contributed significantly to performance could be the existence of high correlations between ADRs. For instance, nausea and headache co-occurred with 596 of the total 832 drugs, and 49 pairs of ADRs co-occurred with more than 400 drugs. As SIDER represents ADRs as unified medical language system (UMLS)[46] concept unique identifiers (CUIs), one side effect may be represented by a group of CUIs (see table 4 for seven concepts related to rhabdomyolysis). To predict one ADR CUI by using other ADR CUIs in the same group may introduce biases and

Table 3 Algorithm comparison using the full feature set over all versus common ADRs

	ADR_All				ADR_50+			
Method	**AUC**	**ACC**	**Precision**	**Recall**	**AUC**	**ACC**	**Precision**	**Recall**
LR	0.9102	0.9486	0.4152	0.5671	0.7648	0.8023	0.5321	0.6908
NB	0.9116	0.9527	0.3537	0.6302	0.8627	0.8431	0.3929	0.7214
KNN	0.9161	0.9595	0.5300	0.5787	0.8508	0.8530	0.5633	0.6401
RF	0.9491	0.9653	0.6310	0.6250	0.9052	0.8784	0.6522	0.7057
SVM	0.9524	0.9669	0.6617	0.6306	0.9141	0.8857	0.6750	0.7227

The full feature set here refers to chemical + biological + phenotypic properties. ADR_All considers all ADRs, and ADR_50+ are the common ADRs caused by at least 50 drugs. All AUC, ACC, Precision, and Recall are micro-averages across ADRs in the corresponding dataset.
ACC, accuracy; ADR, adverse drug reaction; AUC, area under the receiver operating characteristic curve; KNN, K-nearest neighbor; LR, logistic regression; NB, naïve Bayes; RF, random forest; SVM, support vector machine.

overestimate the performance of the model. Therefore, an appropriate grouping schema for ADRs will be investigated in the future. The drug indication information only improved the AUC slightly from 0.9054 (ie, chemical structures only) to 0.9110 (ie, chemical structures + indications). One possible way to improve this is to build a better representation of the indication data. Currently, similar diseases with different CUIs were observed for drug indications in SIDER, for example, C0019693 for 'HIV infection' and C0019699 for 'HIV positive'. Thus, for future work, it may be useful to group the indications.

The improvement produced by biological features was not as much as we initially expected, which may be the result of a few issues. First, the body's response to a drug is a complex process. When a drug enters the body and interacts with its intended targets, favorable effects are expected. However, at the same time, a drug often binds to other protein pockets with varying affinities (off-target interactions), leading to observed side effects. Furthermore, the biological features (ie, protein targets, transporters, enzymes, and pathway) used in this study are relatively simple and probably do not provide the details of molecular processes associated with the drugs.

One problem with the proposed ADR prediction model is imbalanced samples. Of the 1385 ADRs in our dataset, 554 were observed to be associated with fewer than five drugs. Therefore, for these ADR predictions, the dataset has an approximate 1:166 positive to negative ratio, which causes a serious problem for classification algorithms. In the case of an imbalanced classification problem such as this, the large preponderance class will dominate the decision process, which produces classification bias toward the majority class (negative class in this case). As a result, the precision for these ADR predictions would be close to 0%, but accuracy would be near 100%. To compare with results reported in Pauwels *et al*,[27] we followed their approach to report global AUC values. However, owing to the imbalance problem, the global AUC could be very high (over 0.9 in this task), but the actual ability to detect and predict positive samples (the ADRs) could be low. Therefore we reported precision and recall in addition to the AUC. As expected, although 'chem' features achieved over 0.9 AUC, precision and recall were <0.5 (table 2). Furthermore, when the global AUC and accuracy is used, any improvements in the prediction accuracy of the common ADRs might be diluted by the 554 rare ADRs; thus the contribution of the feature addition could be severely underestimated. For example, after the inclusion of biological properties, the AUC remained relatively similar, but the precision actually improved from 43.37% to 46.23%, with relatively similar recall of 50%. We also analyzed different feature sets by only focusing on ADRs associated with at least 50 drugs so that we have sufficient positive samples. As expected, the results showed more significant contributions by each feature addition in terms of AUC, accuracy, precision and recall because rare ADRs that may distort the measures were excluded. For example, in the case of biological properties, its improvement in AUC was 0.02 for common ADRs as opposed to 0.004 for all ADRs.

Different methods have been proposed to address the imbalanced classification problem.[47–49] As a further analysis, we tested a simple method for addressing the sample imbalance

Table 4 Clinical validation examples of cerivastatin and rofecoxib

UMLS CUI	**Known ADRs in SIDER**	**Chem**	**Chem+bio**	**Chem+bio+pheno**
Cerivastatin (Baycol)				
C0035410	Rhabdomyolysis	No	Yes	Yes
C0026848	Myopathy	No	Yes	Yes
C0027121	Myositis	No	Yes	Yes
C0231528	Myalgia	Yes	Yes	Yes
C0026821	Muscle cramps	No	Yes	Yes
C0011633	Dermatomyositis	No	No	No
C0027080	Myoglobinuria	No	No	No
	Group above ADRs	Yes	Yes	Yes
Rofecoxib (Vioxx)				
C0027051	Myocardial infarction	No	No	Yes
C0008031	Chest pain	No	No	Yes
C0004238	Atrial fibrillation	No	No	No
C0018802	Congestive heart failure	No	No	No
	Group above ADRs	Yes	Yes	Yes

ADR, adverse drug reaction; Bio, biological property; Chem, chemical structure; CUI, concept unique identifier; Pheno, phenotypic property; UMLS, unified medical language system.

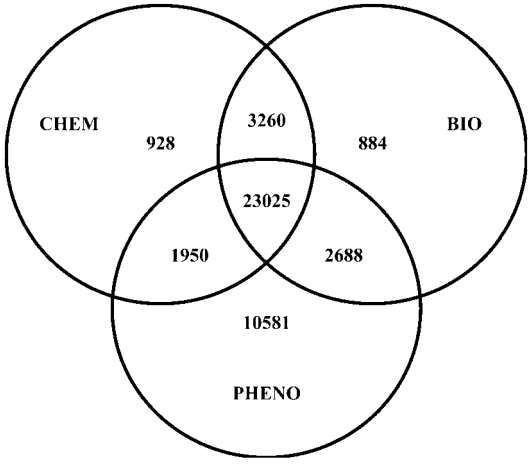

Figure 5 Overlap of the true positive predictions using CHEM (chemical structure), BIO (biological properties), or PHENO (phenotypic properties) features.

problem by adjusting the class weights of the RF and SVM classifiers (ie, weight = 1 − (class samples/total samples)) and observed improvement in AUC only for RF (increased from 0.9491 to 0.9524). SVM did not improve with class weight adjustment because it is very sensitive to parameters; thus parameters must be reoptimized when weights are adjusted. In the future, we plan to explore other techniques such as feature selection and resampling algorithms as suggested previously.[47–49]

Furthermore, the clinical validation examples of Baycol and Vioxx support the utility by detecting post-market adverse drug events using information from other medications in the database. For Baycol, the model based on 'chem' detected only one ADR related to rhabdomyolysis, while the use of 'chem+bio' was able to detect five of seven related ADRs, and the addition of 'pheno' did not result in more predictions. For Vioxx, 'chem+bio+pheno' was required to detect two of four ADRs related to heart attack. This highlights the utility of chemical and biological data for detecting and predicting likely adverse events, as well as the need for incorporating human adverse event data (phenotypic) as in SIDER to allow detection of other signals. These results suggest that our model has the potential to make clinically important ADR predictions early rather than waiting for sufficient post-market population response data to accumulate.

The study has several limitations, and there is scope for much future work to be carried out. For one, we would like to investigate algorithms that have better interpretability, which can return important features associated with ADRs. Moreover, in this study, representation for phenotypic features was relatively simple. More sophisticated methods (eg, categorizing drug indications via ontologies) could be further examined. Furthermore, a drug acts by inducing perturbations to biological systems, which involve various molecular interactions such as protein–protein interactions, signaling pathways, and pathways of drug action and metabolism.[50] Therefore, in future work, we also plan to incorporate more detailed features such as interaction networks and drug bioactivities into the integrative framework for identification of ADRs.

CONCLUSION

This study proposed a new drug surveillance framework for ADR prediction by integrating chemical (ie, compound signatures), biological (ie, protein targets, transporters, enzymes, and pathways), and phenotypic (ie, indications and other known side effects) properties. Using a set of 1385 side effects for 832 drugs from the SIDER database, we developed ML models to integrate the different sources of information for prediction. Five ML algorithms—LR, NB, KNN, RF, and SVM—were systematically compared through fivefold cross-validations, and SVM was found to outperform the others. The AUC score for SVM was increased from 0.9054 when only chemical structures were used to 0.9524 when all three types of information were integrated. The precision increased from 43.37% to 66.17%, and recall increased from 49.25% to 63.06%. Most importantly, with rofecoxib and cerivastatin used as case studies, the proposed model was able to predict clinically important ADRs. These results suggest that such data fusion approaches are promising for large-scale ADR prediction.

Contributors ML, YW, and HX were responsible for the overall design, development, and evaluation of this study. YW, YC, ML, and XC worked on the machine learning experiments. JS and ZZ extracted the biological features for this study. MM designed and reviewed the clinical validation experiments of Baycol and Vioxx. ML and HX did the bulk of the writing, and ZZ, XC, and MM also contributed to writing and editing of this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

Funding This study was supported in part by grants from the NHLBI 5U19HL065962 and the NCI R01CA141307. ML is supported by the NLM training grant 3T15LM007450-08S1. JS is partially supported by the 2010 NARSAD Young Investigator Award. ZZ is partially supported by the 2009 NARSAD Maltz Investigator Award. MM is supported by a Veterans Administration HSR&D Career Development Award (CDA-08-020).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The mapping file between drugs in the SIDER database and DrugBank will be available upon request after publication. In addition, the entire training dataset used in our study will be available upon request as well.

REFERENCES

1. **Pirmohamed M,** Breckenridge AM, Kitteringham NR, *et al*. Adverse drug reactions. *BMJ* 1998;**316**:1295–8.
2. **Lazarou J,** Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* 1998;**279**:1200–5.
3. **Moore TJ,** Cohen MR, Furberg CD. Serious adverse drug events reported to the Food and Drug Administration, 1998–2005. *Arch Intern Med* 2007;**167**:1752–9.
4. **Giacomini KM,** Krauss RM, Roden DM, *et al*. When good drugs go bad. *Nature* 2007;**446**:975–7.
5. **Whitebread S,** Hamon J, Bojanic D, *et al*. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov Today* 2005;**10**:1421–33.
6. **Bate A,** Lindquist M, Edwards IR, *et al*. A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol* 1998;**54**:315–21.
7. **van Puijenbroek EP,** Bate A, Leufkens HG, *et al*. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol Drug Saf* 2002;**11**:3–10.
8. **Ahmed I,** Thiessard F, Miremont-Salame G, *et al*. Pharmacovigilance data mining with methods based on false discovery rates: a comparative simulation study. *Clin Pharmacol Ther* 2010;**88**:492–8.
9. **Harpaz R,** Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* 2010;**11** (Suppl 9):S7.
10. **Harpaz R,** Perez H, Chase HS, *et al*. Biclustering of adverse drug events in the FDA's spontaneous reporting system. *Clin Pharmacol Ther* 2011;**89**:243–50.
11. **Evans SJ,** Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001;**10**:483–6.
12. **Szarfman A,** Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf* 2002;**25**:381–92.
13. **DuMouchel W.** Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat* 1999;**53**:177–202.
14. **Ahmed I,** Dalmasso C, Haramburu F, *et al*. False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics* 2010;**66**:301–9.
15. **Tatonetti NP,** Denny JC, Murphy SN, *et al*. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clin Pharmacol Ther* 2011;**90**:133–42.
16. **Chiang AP,** Butte AJ. Data-driven methods to discover molecular determinants of serious adverse drug events. *Clin Pharmacol Ther* 2009;**85**:259–68.
17. **Pouliot Y,** Chiang AP, Butte AJ. Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clin Pharmacol Ther* 2011;**90**:90–9.
18. **Fliri AF,** Loging WT, Thadeio PF, *et al*. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat Chem Biol* 2005;**1**:389–97.
19. **Campillos M,** Kuhn M, Gavin AC, *et al*. Drug target identification using side-effect similarity. *Science* 2008;**321**:263–6.
20. **Scheiber J,** Chen B, Milik M, *et al*. Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J Chem Inf Model* 2009;**49**:308–17.
21. **Fukuzaki M,** Seki M, Kashima H, *et al*. Side effect prediction using cooperative pathways. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Washington, DC, 2009:142–7.
22. **Xie L,** Li J, 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.
23. **Bender A,** Scheiber J, Glick M, *et al*. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* 2007;**2**:861–73.
24. **Scheiber J,** Jenkins JL, Sukuru SC, *et al*. Mapping adverse drug reactions in chemical space. *J Med Chem* 2009;**52**:3103–7.
25. **Yamanishi Y,** Kotera M, Kanehisa M, *et al*. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010;**26**:i246–54.
26. **Hammann F,** Gutmann H, Vogt N, *et al*. Prediction of adverse drug reactions using decision tree modeling. *Clin Pharmacol Ther* 2010;**88**:52–9.

27. **Pauwels E,** Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics* 2011;**12**:169.
28. **Wang Y,** Bolton E, Dracheva S, *et al*. An overview of the PubChem BioAssay resource. *Nucleic Acids Res* 2010;**38**(Database issue):D255—66.
29. **Kuhn M,** Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**:343.
30. **Chen B,** Wild D, Guha R. PubChem as a source of polypharmacology. *J Chem Inf Model* 2009;**49**:2044—55.
31. **Bolton E,** Wang Y, Thiessen PA, *et al*. PubChem: integrated platform of small molecules and biological activities. *Chapter 12 in Annual Reports in Computational Chemistry*. Washington, DC: American Chemical Society, 2008.
32. **Knox C,** Law V, Jewison T, *et al*. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 2010;**39**(Database issue):D1035—41.
33. **Wishart DS,** Knox C, Guo AC, *et al*. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;**36**(Database issue):D901—6.
34. **Wishart DS,** Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**(Database issue):D668—72.
35. **Kanehisa M,** Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27—30.
36. **Kanehisa M,** Goto S, Furumichi M, *et al*. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010;**38**(Database issue):D355—60.
37. **Kanehisa M,** Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**(Database issue):D354—7.
38. **Fan RE,** Chang KW, Hsieh CJ, *et al*. LIBLINEAR: a library for large Linear classification. *J Mach Learn Res* 2008;**9**:1871—4.
39. *CLOP—Challenge Learning Object Package*. http://clopinet.com/CLOP/
40. **Hall M,** Frank E, Holmes G, *et al*. The WEKA data mining software: an update. *SIGKDD Explorations* 2009;**11**:10—18.
41. **Chang CC,** Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**:1—27.
42. **Birnbaum ZW,** Tingey FH. One-sided confidence contours for probability distribution functions. *Ann Math Stat* 1951;**22**:592—6.
43. **Conover WJ.** *Practical Nonparametric Statistics*. New York: John Wiley & Sons, 1971.
44. **Miller RGJ.** *Simultaneous Statistical Inference*. New York: Springer-Verlag, 1991.
45. **Cami A,** Arnold A, Manzi S, *et al*. Predicting adverse drug events using pharmacological network models. *Sci Transl Med* 2011;**3**:114—27.
46. **Bodenreider O.** The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**(Database issue):D267—70.
47. **Chawla NV,** Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321—57.
48. **Chen C,** Liaw A, Breiman L. *Using Random Forest to Learn Imbalanced Data*. Berkeley: University of California, 2004. Report No. 666.
49. **Lessmann S.** Solving imbalanced classification problems with support vector machines. *Int Conf Artif Intelligence* 2004:214—20.
50. **Tatonetti NP,** Liu T, Altman RB. Predicting drug side-effects by chemical systems biology. *Genome Biol* 2009;**10**:238.