Published in final edited form as:

Clin Pharmacol Ther. 2011 July ; 90(1): 90–99. doi:10.1038/clpt.2011.81.

Predicting Adverse Drug Reactions Using Publicly-Available PubChem BioAssay Data

Yannick Pouliot, PhD^{1,2}, Annie P. Chiang, PhD^{1,2}, and Atul J. Butte, MD^{1,2}

¹Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States of America

²Center for Pediatric Bioinformatics, Lucile Packard Children's Hospital, Palo Alto, California, United States of America

Abstract

Adverse drug reactions (ADRs) can have severe consequences, such that the ability to predict ADRs prior to market introduction is desirable. Computational approaches applied to pre-clinical data might be one way to inform drug labeling and marketing with respect to potential ADRs.

Based on the premise that some of the molecular actors of ADRs involve interactions detectable in large, and increasingly public, compound screening campaigns, we generated logistic regression models that correlate post-marketing ADRs with screening data from the PubChem BioAssay database. These models analyze ADRs at the level of organ systems, the System Organ Classes (SOCs). Nine of the 19 SOCs under consideration were found to be significantly correlated with pre-clinical screening data. For 6 of the 8 established drugs for which we could retrodict SOC-specific adversities, prior knowledge was found that support these predictions. We conclude by predicting SOC-specific adversities for three unapproved or recently introduced drugs.

Keywords

Adverse drug reactions; prediction; machine learning; compound screening; pharmacovigilance

INTRODUCTION

Pharmaceutical consumption is continuously increasing due to the aging of the U.S. population, enhanced medication coverage, and the introduction of drugs addressing conditions previously untreatable by medications. Although undeniably beneficial, pharmaceuticals are necessarily associated with rates of morbidity and mortality. Adverse drug reactions (ADRs) are defined as “a response to a drug which is noxious and unintended and which occurs at doses normally used in man for prophylaxis, diagnosis, or therapy of diseases or for modification of physiological function”(1). Serious ADRs result in death, hospitalization, significant disability and other permanent and life-threatening conditions(1). Serious ADRs are a major clinical problem, estimated to account for more than 2 million incidents requiring hospitalization annually(2), and more than 100,000 deaths in the United States(3).

Correspondence to: Atul J. Butte Division of Systems Medicine Stanford University School of Medicine 251 Campus Drive, X-163, MS-5415 Stanford, CA, 94305-5415 abutte@stanford.edu Phone: +1-650-723-3465 Fax: +1-650-723-7070.

CONFLICT OF INTEREST

None.

These statistics reflect the challenge of identifying ADRs. This is partly due to the short-duration/defined population testing paradigm of clinical trials, and the difficulty of recognizing novel ADRs in patients with potentially extensive medical histories. Although progress has been made toward identifying the causes of drug-induced morbidity(4, 5), the process remains difficult and haphazard(6), and aspects of a drug's adversity can remain obscured for years.

Many drugs exhibit unexpected specific organ- or body system-specific ADRs, distinct from “generic” ADRs involving liver or kidney damage. The advent of high-throughput molecular measurement technologies, combined with publicly-available datasets, has the potential to substantially facilitate the identification of novel ADRs in newly introduced drugs whose ADR profile is mostly unknown (7). Since a fraction of organ-specific ADRs is likely due to drugs interacting with unintended targets, predicting such ADRs using data from large-scale compound screening campaigns might be possible, as some of the molecular actors of ADRs must involve interactions at the cellular level, and should thus be potentially detectable in such.

Although attempts at predicting ADRs using preclinical compound characteristics or screening data have been made, most notably by Fliri and colleagues(8, 9), much progress remains to be made. Computational methods have been developed wherein pharmacovigilance data are analyzed in conjunction with a drug's structural properties to predict ADR profiles(10, 11). Other methods for predicting ADRs involve testing in non-human and even yeast species, and suffer from interpretability limitations due to each species' pharmacological idiosyncrasies(12).

Here we used a large, publicly-available compilation of heterogeneous, pre-clinical molecular screening assays to determine whether drug bioactivity across vast screens correlates with post-marketing ADRs manifesting in specific System Organ Classes (SOCs). SOCs are used to group types of ADRs according to where they manifest in the body, as defined by the Medical Dictionary for Regulatory Activities (MedDRA)(13). For example, “eosinophilia” as a side-effect of drug treatment is listed under “Blood and lymphatic system disorders” SOC.

We correlated a drug's propensity toward SOC-specific ADRs, as calculated from the Canadian Adverse Drug Reaction (CVAR) pharmacovigilance database (43), with patterns of screening activity observed in the National Center for Biotechnology Information's PubChem BioAssay database(14). A component of the National Institutes of Health (NIH)'s Molecular Libraries Initiative, PubChem BioAssay currently stores data from more 487,000 screens involving hundreds of thousands of compounds across thousands of molecular targets, thus enabling analyses previously available only to pharmaceutical companies.

Using these molecular screening assay data, we were able to create statistical models for 9 of the 19 SOCs under consideration, and used them to predict unrecognized ADRs for drugs currently or recently approved in the US, as well as drugs not yet marketed in the USA.

RESULTS

Our analytical pipeline searched across 485 drug ingredients in 508 BioAssays in PubChem to identify potential unrecognized adverse drug reactivities manifesting in specific System of Organ Classes (SOCs) (Figure 1). For each drug, the pipeline applied logistic regression to seek individual or pairs of BioAssay bioactivities that optimally correlate with increased drug adversity in specific SOCs, as measured by the Proportional Risk Ratio (PRR) metric(15). Drugs with a SOC-specific PRR ≥ 2 were considered as especially prone to ADRs in that SOC.

Model properties

For each SOC, BioAssays were first ranked based on the p-value of the logistic regression between a drug's binarized SOC-specific PRR and its screening bioactivity (Figure 2). BioAssays with the most significant p-values that most improved Akaike's Information Criterion (AIC) (16) when combined into a single regression equation were selected to compose the SOC's model. A total of 19 univariate or bivariate logistic regression models were generated in this way, one for each SOC grouping of adverse reactions, trained on as many drug ingredients as possible.

We evaluated these models using leave-one-out-cross-validation (LOOCV), which removes one drug ingredient from the dataset and uses the model to predict whether that drug had a significantly high PRR or not. The model's performance is then assessed using Receiver Operating Characteristic(17) (ROC) analysis, and the process repeated for all drug ingredients within the model. The mean Area Under the Curve (AUC), regression coefficient, and p-value are then computed. The mean p-value of recomputed LOOCV regression models ranged from 10^{-2} to 10^{-8} , with mean AUCs ranging from 0.60 to 0.92 (Table 1). Nine models (47%) had AUC values of 0.7 or better (Table 1). The ROC curves for the best two models, "Immune system disorders" (LOOCV mean AUC = 0.92) and "Blood and lymphatic system disorders" (LOOCV mean AUC = 0.79), are depicted in Figure 3 A and B, respectively.

Our models encompass between 70 and 437 drug ingredients per model, with most models relying on BioAssays that interrogate defined molecular targets (Table 1). Of the 37 BioAssays selected by the pipeline, two were assigned to more than one SOC: AID2066 was found to be predictive in SOC's "Gastrointestinal disorders" and "General disorders and administration site conditions", whereas AID2557 was predictive in the "Nervous system disorders" and the "Cardiac disorders" SOC's. Most of the BioAssays in our models were performed by members of the NIH Molecular Library Screening Center Network, or the NIH Molecular Libraries Probe Production Centers Network. These BioAssays were roughly divided across the "screening" (single compound concentration testing) and "confirmatory" (multiple compound concentration testing) categories. The two best performing models involve screens performed *in vivo*: AID119 ("Immune system disorders" SOC) and AID330 ("Blood and lymphatic system disorders" SOC), respectively. AID119 seeks small molecules growth inhibitors of CCRF-CEM leukemia cells, a human acute lymphoblastic leukemia cell line(18). AID330 seeks small molecule inhibitors of tumor growth or survival for mouse P388 leukemia cells *in vivo*, a model of leukemia. Also notable is the selection of 13 BioAssays (46% of selected BioAssays) that measure biochemical activity in a cell-free context (Table 1).

Interestingly, for those screens with defined targets (78% of selected BioAssays), almost none of the molecular targets of the drugs used to train the models are the same as the targets of the BioAssays learned for a given model.

Predictions for marketed drugs

We performed retrospective evaluation of our models using the individual drugs encompassed in these models. Models with a ROC AUC ≥ 0.7 were used to calculate the logistic probability of high PRR (LPHPRR) for individual drugs within a model. For each model, we selected the drug ingredient with the largest LPHPRR for which our prediction of PRR ≥ 2 did not match its current PRR < 2 as calculated from CVAR pharmacovigilance data. In other words, these are drug ingredients for which we predict a high PRR, but for which a low SOC-specific PRR is calculated given current reporting. Potential unrecognized SOC-specific ADRs were thus predicted for 8 drugs, with LPHPRR ranging from 0.56 for the

“Eye disorders” SOC to 0.93 for the “Blood and lymphatic system disorders” SOC (Table 2).

These predictions of SOC-specific ADRs were then assessed by reviewing a database compendium of the literature, as well as each drug's label. For five of the eight compounds (63%), we found mentions of adverse drug reactions in the FDA's drug label that are associated with the SOC under consideration (Table 2). For example, our model predicts a high PRR for cisplatin for the “Blood and lymphatic system disorders” SOC, which did not match the lower calculated PRR given current reporting in CVAR. However, the label for cisplatin itself lists myelosuppression as a “black box” warning, a type of ADR classified under SOC (Table 2). We suspect the label's warning may have inhibited post-marketing adversity reporting of this ADR to regulatory agencies, a known source of under-reporting (19) that can lead to a lower PRR.

Evidence of SOC-specific adversity was found in the DRUGDEX database(20) for the sixth ingredient, clioquinol. This anti-fungal agent, predicted to create adversity in the “Eye disorders” SOC, is already known to be associated with subacute myelo-optic neuropathy (SMON) syndrome in ethnic Japanese (21) (Table 2).

We could not find prior knowledge of carcinogenicity for the skin bleaching agent hydroquinone in humans, as predicted by our model. However, hydroquinone is known to belong to a small group of drugs with genotoxic carcinogenic activity in *in vivo* murine bone marrow micronucleus tests, but not in *in vitro* mutagenesis tests such as the Ames test(22).

Similarity, we could not find prior knowledge for the predicted endocrine SOC-specific adversity for the antimalarial drug pyrimethamine, and suggest this as a potentially novel or unreported class of ADRs for this drug. Overall, 75% of our predictions of adversity in humans could be substantiated by the literature or the drug's label.

Predictions for novel or recently approved drugs

We further applied our models to predict adversity for novel or recently approved drugs not present in the CVAR data set used to train our models. Three compounds were found to meet our requirements for novelty, presence in the models' BioAssays, and being investigated by ongoing clinical trials: tranilast, nitazoxanide and diacerein (Table 3). Of these three, nitazoxanide is the only FDA-approved drug (approved in 2002).

We predict adversity for diacerein within the “Skin and subcutaneous tissue disorders” SOC, and found one supporting literature report pertaining to this prediction (Table 3), wherein diacerein has been anecdotally associated with a single fatal case of toxic epidermal necrolysis(23), a type of ADR included in this SOC. We could not find prior knowledge for our predictions of respiratory system disorders for tranilast, and induction of neoplasms for nitazoxanide. We propose these as potentially novel side effects worthy of special attention in pharmacovigilance efforts.

DISCUSSION

This analysis demonstrates how drugs characterized by an increased frequency of ADRs in specific SOC's can potentially be detected using patterns of biological activity from qualitatively different screens, such as screens evaluating *in vivo* cytotoxicity, bioactivity in cell culture, or molecular interactions in cell-free biochemical assays (Table 1). To our knowledge, this is the first demonstration that post-marketing adverse drug reactions can be correlated with data from diverse, publicly-available preclinical biological assays, building from previous work using proprietary, univariate databases (8, 9). Along with recent

computational approaches based on functional profiling(8, 9, 24) , docking(25), compound structure(10), and integrated data sets(11), our results demonstrate the potential for the identification of hitherto unrecognized ADRs using computational models that integrate pre-clinical screening data with pharmacovigilance data. We deliberately used straight-forward logistic regression to avoid potential model overfitting. .

Because they frequently involve pharmacologically-relevant compounds and targets, the large-scale compound screening campaigns available from PubChem BioAssay present an attractive data set from which to discover potential drug adversities. Many screens involve targets that belong to families with known pharmacologically active targets but are not themselves drug targets, such as KCNJ2, a potassium channel also known as Kir2.1. This protein is the target for AID1672, the BioAssay most correlated with the “Nervous system disorders” SOC (Table 1). Mutated forms of KCNJ2 are associated with congenital long QT Syndrome, and many drugs are known to interact with several other members of the family. However, our approach is fundamentally agnostic of the pharmacological characteristics of the screens it evaluates, such that screens can be selected that do not involve defined molecular targets or were not intended for drug discovery.

Our approach is further predicated on the premise that a fraction of SOC-specific ADRs are at least partly due to drugs interacting with unintended targets (“promiscuity”). We further hypothesized that these interactions should be detectable in large-scale compound screening campaigns, since some of the molecular actors of ADRs must involve interactions at the cellular level, and are thus potentially detectable in such assays. Compound promiscuity in PubChem BioAssay screens has been demonstrated recently, with 25-40% of the compounds in that database exhibiting bioactivity with more than one target(26). This result is congruent with our own finding: the molecular targets of the drugs are typically different from the targets used by the BioAssays in the model.

Given the simplicity of the logistic regression modeling algorithm, it is perhaps surprising that we achieved the selectivity and specificity reported here: half of our models achieved a LOOCV AUC of 0.7 or greater, and all models achieved 0.6 or greater (Table 1). We attribute this performance in part to the rich diversity of screens in the PubChem BioAssay database, which provides good odds of identifying screens that share a biological relationship with the ADRs under consideration. This performance is further reflected in the robustness of the models’ predictions: 75% of the SOC-specific adversity predictions for approved drugs were corroborated by prior knowledge, mostly involving FDA-sanctioned data (Table 2). While no support could be found for our method’s prediction of potential carcinogenicity of hydroquinone in humans, suggestive evidence exists in other mammals: Hydroquinone is a skin bleaching agent with an unusual property: it is carcinogenic in murine *in vivo* bone marrow micronucleus tests but inactive in *in vitro* mutagenesis tests(22). For this reason, studies of hydroquinone’s potential dermal carcinogenicity in mice and rats were launched by the FDA recently under the National Toxicology Program(27).

Similarly interesting predictions were generated for the three drugs new to the US market or otherwise unapproved for which our models could be applied: tranilast, diacerein and nitazoxanide (Table 3). No meaningful prior knowledge was found in support of these predictions. Tranilast was approved in 1982 in Japan and South Korea for the treatment of bronchial asthma, yet our model predicts adverse reactions in the respiratory system. Tranilast is a synthetic tryptophan metabolite that inhibits the release of histamine (28), leukotriene-mediated smooth muscle contraction (29), and collagen synthesis(30).

Nitazoxanide was approved by the FDA in 2002 and is a member of the thiazolides family, a novel class of drugs for the treatment of protozoan infections such as cryptosporidiosis and

giardiasis(31). Its target is believed to be pyruvate:ferredoxin oxidoreductase (PFOR), an enzyme essential to electron transfer reactions used in anaerobic energy metabolism. Our model predicts that nitazoxanide has the potential to induce neoplasia. Nitazoxanide and other thiazolides inhibit the enzymatic activity of glutathione-*S*-transferase μ (GSTP1)(32), a marker of cancer development in many tissues(33). GSTP is a member of a diverse superfamily frequently overexpressed in multidrug-resistant cancer cells(34). Therefore, nitazoxanide's potential neoplastic adversity could be related to its apoptotic activity in human colon cancer cells cultured *in vitro*(32), as it is believed to inhibit the anti-apoptosis activity of glutathione transferase isozymes within the c-Jun N-terminal kinase (JNK) signaling pathway(35), a pathway known to control cell proliferation and apoptosis(36).

Diacerein is an atypical non-steroidal anti-inflammatory drug (NSAID) approved in France for the treatment of osteoarthritis since 1992. A single literature case report associates diacerein with toxic epidermal necrolysis(23), a syndrome classified under the "Skin and subcutaneous tissue disorders" SOC, the SOC predicted by our model. Diacerein directly inhibits the synthesis of interleukin-1 (IL-1) *in vitro*(37), and, indirectly, the synthesis of metalloprotease-13 (collagenase-3; MMP-13) in the subchondral bone of osteoarthritic patients(38). MMP-13 is induced in various skin diseases(39) and mediates cell cycle progression in mouse melanocytes(40), thus providing a rationale for a potential role for diacerein in skin diseases.

Although we make a number of predictions of drug adversity, we fully acknowledge that until these predictions are observed in post-marketing experience in multiple settings, the clinical utility of these drugs should not be altered. Rather, our intent is to provide rational, testable hypotheses able to help inform the identification of unrecognized ADRs in a clinical context, thus shortening the delay during which ADRs go undetected. The approach should also be applicable within the regulatory framework, by better informing surveillance and, eventually, warning statements, and within the drug discovery, development, and approval processes, by providing predictive preclinical assays applicable to novel compounds.

METHODS

Overview of analytical process

We designed an analytical pipeline that relies on a set of integrated databases to correlate a drug's pre-clinical, publically-available screening bioactivity with its pharmacovigilance adversity (Figure 1). The pipeline seeks drug screening bioactivities that correlate with the drug's adversity in individual SOC's as calculated by logistic regression models applied to bioactivity and SOC-specific PRR. For each SOC, the model with the best regression *p*-value was selected, and its selectivity and specificity assessed.

Controlled nomenclature

Identifiers from the Unified Medical Language System(41) (UMLS), version 2007AC, were used to uniquely identify entities in the PubChem BioAssay(14), Substance(42), CVAR (43) and DrugBank(44) databases, as described below.

Data sets

Post-marketing adverse drug reaction data were obtained from CVAR on March 29th, 2010 and loaded into a MySQL relational database (Oracle Corporation, Redwood Shores, CA). At that time, CVAR held spontaneously reported ADRs in Canada from 1965 to 2009. Drug reactions collected in pharmacovigilance databases cannot usually be attributed definitively to a drug, and are thus presumed to be valid by our analytical pipeline. CVAR drug ingredient names were assigned a UMLS unique concept identifier for drugs ("RXCU") to

cross-reference compounds across databases. 2,899 drug ingredients listed in CVAR were assigned an RXCUI, with 485 RXCUIs mapped to compounds in the PubChem BioAssay database (Table S1), associated with 1,498,570 presumed adverse drug reactions. Drug ingredients were not filtered according to type of molecule, such as small molecules and biologics.

CVAR relies upon the Medical Dictionary for Regulatory Activities(13) (MedDRA) to group ADRs based on the tissues and organs where they manifest, the System of Organ Classes (SOC). Analyzing ADRs at the level of a SOC improves the detectability of signals in a manner consistent with how ADRs manifest in clinical practice. After merging the “Immune system disorders” SOC into the “Infections and infestations” SOC and excluding the SOCs “Injury, poisoning and procedural complications”, “Investigations”, “Social circumstances” and “Surgical and medical procedures”, 19 SOCs were found associated with ADRs meeting our requirements (below).

ADRs had to meet three requirements to participate in the calculation of a drug's SOC-specific PRR (described below): (1) association with a SOC; (2) be of type “adverse reaction” and of class “suspect”; (3) have a minimum of 10 reports associated with the drug ingredient. Several ADRs may be associated with a single report, possibly associated with different SOCs. These requirements ensure that SOC-specific PRRs are calculated on a meaningful number of ADRs for which the drug ingredient is the suspected causative agent. Between 1,250 and 178,290s ADRs per SOC were identified in this way (Table S2).

Screening bioactivity data were obtained from PubChem's BioAssay database on April 1st, 2010 and converted into a MySQL database. At that time, the database contained BioAssays involving 466 molecular targets, as well as BioAssays without defined targets (e.g., cytotoxicity assays), involving more than one million Substance Identifiers (SIDs) (Table S3). The process of mapping SIDs to drug ingredients in CVAR is described in Table S4. Informative BioAssays were selected based on the steps described in Table S5.

Normalization of adverse event counts and BioAssay activity

PubChem BioAssay's Activity Scores of compounds within each BioAssay were normalized to a Z-score according to equation (1):

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where x is the Activity Score of the compound, μ and σ are the average and standard deviation of the Activity Score for all compounds associated with the BioAssay, respectively. Raw activity measurements and depositor-submitted activity assessments stored in PubChem BioAssay (“Outcome”) were not used.

PRR was used to assess a compound's propensity toward adverse reaction. This metric is based upon the ratio of the relative frequency of reactions of a given type as compared with all other types of reactions for a drug, and the frequency of reactions of that type for all other drugs in the database. The “SOC-specific PRR” of all drugs was calculated by pooling a drug's ADRs into those SOCs in which they manifest clinically as per equation (2), using the terms defined in Table 4.

$$\text{PRR} = [A / (A + C)] / [B / (B + D)] \quad (2)$$

For logistic regression, SOC-specific PRRs were binarized (“BPRR”) according to equation (3):

$$\text{BinarizedPRR} = \begin{cases} 0, & \text{PRR} < 2 \\ 1, & \text{PRR} \geq 2 \end{cases} \quad (3)$$

The PRR threshold of 2 used here is generally assumed to indicate meaningful potential for adverse drug reactivity(15). An example of the final matrix is depicted in Figure S1. Compounds without ADRs in a particular SOC were assigned a SOC-Specific PRR of 0 if at least 10 ADR reports involving ADRs in other SOCs were present.

Associating adverse events with pre-clinical assay measurements

Since the number of CVAR drug ingredients shared between BioAssays decreases very rapidly as BioAssays are intersected, a forward- or backward-stepwise predictor selection in which all predictors (BioAssays) are evaluated together could not be performed. Instead, the construction of the logistic regression model was performed in two steps. First, the BioAssay with the most significant univariate logistic regression coefficient was identified (“anchor assay”), followed by the second most significant BioAssay that, when added to the model, most improved the Akaike's Information Criterion(16) (AIC) of the resulting model without unduly impacting the significance of the anchor assay. For models with dual BioAssays, no interaction was assumed between them, and drugs must be present in both BioAssays.

To avoid potentially biasing models toward BioAssays with structurally related compounds, the Tanimoto coefficient (45) was calculated for drug ingredients composing a model by evaluating all pairs of drugs for a Tanimoto coefficient ≥ 0.9 . In a few instances a small fraction of a model's drugs satisfied this threshold (<10%). These were evaluated to determine whether they could bias the model by being overly associated with specific features within the model, for example, BPRR = 1, or Z-score ≥ 2 . No such over-representation was observed in our models (results not shown).

Models were evaluated with a combination of leave-one-out cross validation (LOOCV) and Receiver Operating Characteristic (ROC) analysis: Individual drug ingredients were removed from the dataset, the model re-computed and evaluated using the ROCR module(46). This process was repeated for all drug ingredients within the model, and the average ROC AUC, regression coefficient, and p-value were generated for each SOC.

Screening target specificity

We assessed the target specificity of compounds screened in our models' BioAssays by comparing the known molecular interactors of a compound with the target associated with the BioAssay as stated by PubChem. DrugBank's drug-target associations were used for this purpose. Comparisons were made using GenBank GI numbers and target names.

Prediction of unrecognized ADRs in marketed drug ingredients

As test of our method's predictive power, we sought to identify drug ingredients with unrecognized ADRs using models with ROC AUC ≥ 0.7 . Ingredients meeting three requirements were selected: largest logistic probability of high PRR (LPHPRR), LPHPRR ≥ 0.5 , but observed PRR < 2 . In our models, an LPHPRR ≥ 0.5 indicates a compound predicted to exhibit a PRR ≥ 2 .

Three sources were consulted to determine prior association of the selected drug ingredient with the predicted SOC: the U.S. FDA drug label (DailyMed(47)); the Warnings and Adverse Effects sections of each ingredient's record in the DRUGDEX database(20), a compilation of drug data and knowledge derived from the literature and regulatory agencies;

and the FDA's MedWatch database(48). Types of ADRs equivalent to the MedDRA Primary Terms linked to the SOC predicted to be associated with the drug ingredient were taken to indicate that the ingredient was already known to be associated with that SOC.

ADR prediction for novel drugs

We also tested our models' ability to predict adverse drug reactions in novel medications with limited or no known post-marketing adversity. Four conditions were applied for a drug ingredient to be considered "novel": (1) not approved by the FDA at the time of writing, or approved within the past ten years; (2) included in an ongoing clinical trial as listed in ClinicalTrials.gov(49) as of October 2010; (3) not included in the CVAR data set used to train our models due to lack of ADR reports ; (4) present in the set of compounds screened in the BioAssays associated with a model. The bioactivity of novel ingredients was used to calculate the LPHPRR using models with ROC AUC = 0.7. For each SOC, the drug ingredient with the best LPHPRR and LPHPRR = 0.5 was retained. Predictions were assessed against prior knowledge according to the process described above, as well as searches in PubMed and EMBASE(50).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by funding from the Lucile Packard Foundation for Children's Health, the National Institute of General Medical Sciences (R01 GM079719), the National Library of Medicine (T15 LM007033), the Howard Hughes Medical Institute, and the Pharmaceutical Research and Manufacturers of America Foundation. We also thank the Hewlett Packard Foundation for an equipment grant. We thank Alexander A. Morgan, Chirag J. Patel and Dr. Shivkumar Venkatasubrahmanyam for statistical and programming suggestions.

REFERENCES

1. International Conference on Harmonization. ICH harmonised tripartite guideline. Guideline for good clinical practice. 1996; E6(r1)
2. Lazarou J. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. JAMA: The Journal of the American Medical Association. 1998; 279:1200–5. [PubMed: 9555760]
3. IOM. To err is human. Building a safer health system. Institute of Medicine; Washington, D.C.: 2000. 2000.
4. Alfirevic A, Pirmohamed M. Adverse drug reactions and pharmacogenomics: Recent advances. Personalized Medicine. 2008; 5:11–23.
5. Wilke RA, et al. Identifying genetic risk factors for serious adverse drug reactions: Current progress and challenges. Nature Reviews Drug Discovery. 2007; 6:904–16.
6. Woosley RL, Cossman J. Drug development and the FDA's critical path initiative. Clin Pharmacol Ther. 2007; 81:129–33. [PubMed: 17186012]
7. Chiang AP, Butte AJ. Data-driven methods to discover molecular determinants of serious adverse drug events. Clinical Pharmacology & Therapeutics. 2009; 85:259–68. [PubMed: 19177064]
8. Fliri AF, Loging WT, Thadeio PF, Volkmann RA. Analysis of drug-induced effect patterns to link structure and side effects of medicines. Nature Chemical Biology. 2005; 1:389–97.
9. Fliri A, Loging W, Thadeio P, Volkmann R. Biological spectra analysis: Linking biological activity profiles to molecular structure. Proc Natl Acad Sci USA. 2005; 102:261–6. [PubMed: 15625110]
10. Hammann F, Gutmann H, Vogt N, Helma C, Drewe J. Prediction of adverse drug reactions using decision tree modeling. Clinical Pharmacology & Therapeutics. 2010; 88:52–9. [PubMed: 20220749]

11. Bender A, 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. [PubMed: 17477341]
12. Hillenmeyer M, Ericson E, Davis R, Nislow C, Koller D, Giaever G. Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome Biology*. 2010; 11:R30. [PubMed: 20226027]
13. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Safety*. 1998; 20:109–17. [PubMed: 10082069]
14. Wang Y, et al. An overview of the PubChem BioAssay resource. *Nucleic Acids Research*. 2009; 38:D255–D66. [PubMed: 19933261]
15. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*. 2001; 10:483–6. [PubMed: 11828828]
16. Akaike H. New look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; AC-19:716–23.
17. Egan, J. Signal detection theory and roc analysis. Academy Press; New York: 1975.
18. Foley GE, Lazarus H, Farber S, Uzman BG, Boone BA, McCarthy RE. Continuous culture of human lymphoblasts from peripheral blood of a child with acute leukemia. *Cancer*. 1965; 18:522–9. [PubMed: 14278051]
19. Hasford J, Goettler M, Munter KH, Müller-Oerlinghausen B. Physicians' knowledge and attitudes regarding the spontaneous reporting system for adverse drug reactions. *Journal of Clinical Epidemiology*. 2002; 55:945–50. [PubMed: 12393084]
20. DRUGDEX System (Thomson Reuters (Healthcare) Inc.). [October 2010]
21. Bareggi SR, Cornelli U. Clioquinol: Review of its mechanisms of action and clinical uses in neurodegenerative disorders. *CNS Neuroscience & Therapeutics*. 2010 no-no.
22. Tweats DJ, et al. Report of the IWGT working group on strategy/interpretation for regulatory in vivo tests. Ii. Identification of in vivo-only positive compounds in the bone marrow micronucleus test. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*. 2007; 627:92–105. [PubMed: 17113817]
23. Dereure O, Hillaire-Buys D, Augias D, Griffe O, Blanc F. Fatal toxic epidermal necrosis: Responsibility of diacerein? A controversy. *Dermatology*. 1998; 196:431–2. [PubMed: 9669123]
24. Sato T, Matsuo Y, Honma T, Yokoyama S. In silico functional profiling of small molecules and its applications. *Journal of Medicinal Chemistry*. 2008; 51:7705–16. [PubMed: 19032081]
25. Xie L, Wang J, Bourne PE. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Computational Biology*. 2007; 3:2324–32.
26. Han L, Wang Y, Bryant SH. A survey of across-target bioactivity results of small molecules in PubChem. *Bioinformatics*. 2009; 25:2251–5. [PubMed: 19549631]
27. US Food and Drug Administration. Hydroquinone studies under the national toxicology program (NTP).
28. Koda A, Kurashina Y, Nakazawa M. The inhibition mechanism of histamine release by n-(3,4-dimethoxycinnamoyl) anthranilic acid. *International archives of allergy and applied immunology*. 1985; 77:244–5. [PubMed: 2409023]
29. Mesa MG. New approach to the mechanism of antiasthmatic action of tranilast. *Allergologia et immunopathologia*. 1990; 18:53–6. [PubMed: 1696424]
30. Yamada H, Tajima S, Nishikawa T. Tranilast inhibits collagen synthesis in normal, scleroderma and keloid fibroblasts at a late passage culture but not at an early passage culture. *Journal of Dermatological Science*. 1995; 9:45–7. [PubMed: 7537081]
31. White C. Nitazoxanide: A new broad spectrum antiparasitic agent. *Expert review of anti-infective therapy*. 2004; 2:43–9. [PubMed: 15482170]
32. Müller J, Sidler D, Nachbur U, Wastling J, Brunner T, Hemphill A. Thiazolides inhibit growth and induce glutathione-s-transferase pi (GSTP1)-dependent cell death in human colon cancer cells. *International Journal of Cancer*. 2008; 123:1797–806.

33. Howells REJ, et al. Association between glutathione-s-transferase gstp1 genotypes, GSTP1 over-expression, and outcome in epithelial ovarian cancer. *International journal of gynecological cancer*. 2004; 14:242–50. [PubMed: 15086723]
34. Tew KD. Glutathione-associated enzymes in anticancer drug resistance. *Cancer Research*. 1994; 54:4313–20. [PubMed: 8044778]
35. Laborde E. Glutathione transferases as mediators of signaling pathways involved in cell proliferation and cell death. *Cell Death Differ*. 2010; 17:1373–80. [PubMed: 20596078]
36. Wagner EF, Nebreda ÁR. Signal integration by JNK and p38 MAPK pathways in cancer development. *Nature Reviews Cancer*. 2009; 9:537–49.
37. Yaron M, Shirazi I, Yaron I. Anti-interleukin-1 effects of diacerein and rhein in human osteoarthritic synovial tissue and cartilage cultures. *Osteoarthritis and Cartilage*. 1999; 7:272–80. [PubMed: 10329302]
38. Boileau C, Tat S, Pelletier J-P, Cheng S, Martel-Pelletier J. Diacerein inhibits the synthesis of resorptive enzymes and reduces osteoclastic differentiation/survival in osteoarthritic subchondral bone: A possible mechanism for a protective effect against subchondral bone remodelling. *Arthritis Research & Therapy*. 2008; 10:R71. [PubMed: 18578867]
39. Papakonstantinou E. Expression of matrix metalloproteinases and their endogenous tissue inhibitors in skin lesions from patients with tuberous sclerosis. *Journal of the American Academy of Dermatology*. 2004; 51:526–33. [PubMed: 15389186]
40. Meierjohann S, et al. MMP13 mediates cell cycle progression in melanocytes and melanoma cells: In vitro studies of migration and proliferation. *Molecular cancer*. 2010; 9:201. [PubMed: 20667128]
41. Olivier B. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*. 2004; 32:D267–D70. [PubMed: 14681409]
42. Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated platform of small molecules and biological activities. 2008; 4:217–41.
43. [March 2009] Canada Vigilance Adverse Reaction Online Database (Health Canada).
44. Wishart DS, et al. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*. 34:D668–D72. [PubMed: 16381955]
45. Tanimoto, T. IBM internal report, nov. 17th 1957. 1957. 1957.
46. Sing T. ROCR: Visualizing classifier performance in R. *Bioinformatics*. 2005; 21:3940–1. [PubMed: 16096348]
47. DailyMed (National Library of Medicine). [October 2010]
48. MedWatch: The FDA Safety Information and Adverse Event Reporting Program (U.S. Food and Drug Administration).
49. McCray AT, Ide NC. Design and implementation of a national clinical trials registry. *Journal of the American Medical Informatics Association*. 2000; 7:313–23. [PubMed: 10833169]
50. EMBASE (Elsevier B.V.). Accessed October 2010.

Figure 1. Analytical pipeline used to correlate drug adversity to screening bioactivity

Outline of the analytical process applied to drug ingredients found in the CVAR pharmacovigilance database that are also present in the PubChem BioAssay screening database. The pipeline identifies the one or two BioAssays that are most correlated with CVAR potential adverse drug reactions, and creates a logistic regression model using these BioAssays. The best model is evaluated using LOOCV and ROC analysis.

Figure 2. Regression p-values for BioAssays evaluated by the analytical pipeline
The p-values of regression models for all 508 BioAssays and all SOC's are plotted.

Figure 3. Selectivity and specificity for the top two best performing models

The prediction performance of the first two best performing models is illustrated using ROC plots. Panel A: model for “Immune system disorders” SOC. Panel B: “Blood and lymphatic system disorders” SOC. Blue line: performance obtained from individual LOOCV runs. Red line: performance averaged over all LOOCV runs, along with the false positive range (black line).

Table 1

Summary properties of logistic regression models

SOC	Anchor assay					Secondary assay						
	Mean LOOCV AUC (SEM)	Drugs ingredients in model	AID	Assay type	Objective & target	Mean p-value (SEM)	Mean coefficient (SEM)	AID	Assay type	Objective & target	Mean p-value (SEM)	Mean coefficient (SEM)
Immune system disorders	0.92 (5.93E-4)	70	119	confirmatory (<i>in vivo</i>)	Small molecule inhibitors of tumor cell growth in human CCRF-CEM leukemia cells [<i>M. musculus</i>]	2.95E-4 (1.05E-5)	1.15E+0 (5.08E-3)	2451	confirmatory (cell-free)	Small molecule inhibitors of Inhibitors of fructose-1,6-bisphosphate aldolase [<i>G. Lamblia</i>]	9.95E-1 (1.56E-5)	-4.02E+0 (4.20E-3)
Blood and lymphatic system disorders	0.79 (3.18E-4)	185	330	screening (<i>in vivo</i>)	Small molecule inhibitors of mouse P388 leukemia [<i>M. musculus</i>]	1.15E-4 (1.30E-6)	2.34E-1 (6.33E-4)					
Reproductive system and breast disorders	0.77 (1.67E-4)	271	1461	confirmatory (cell-based)	Small molecule inhibitors of neuropeptide S receptor (NPSR) signaling [<i>H. sapiens</i>]	7.49E-8 (1.45E-9)	5.55E-1 (3.87E-4)	1285	screening	Small molecule inhibitors of amyloid precursor protein (APP) translation in SH-SY 5Y neuroblastoma [<i>H. sapiens</i>]	3.01E-2 (1.71E-4)	-1.23E+0 (2.83E-3)
Neoplasms benign, malignant and unspecified	0.76 (8.02E-4)	115	543	confirmatory	Small molecules cytotoxic to H-4-II-E hepatoma cell line [<i>M. musculus</i>]	2.16E-5 (8.19E-7)	9.39E-1 (2.21E-3)	1476	confirmatory (cell-free)	Small molecule inhibitors of cruzain chain A cysteine protease [<i>T. cruzi</i>]	1.96E-1 (1.05E-3)	-6.40E-1 (4.61E-3)
Respiratory, thoracic and mediastinal disorders	0.76 (4.65E-4)	177	774	confirmatory (cell-free)	Small molecule inhibitors of enzymes frequently used to reach an NAD/NADH endpoint	2.14E-3 (3.85E-5)	-9.21E+0 (2.31E-2)	2462	screening	Small molecule inhibitors of chain A-Bcl2-A1-mediated apoptosis [<i>M. musculus</i>]	8.07E-2 (1.28E-3)	-5.35E-1 (7.30E-3)
Nervous system disorders	0.74 (2.72E-4)	376	1672	screening	Small molecule inhibitors of inward-rectifying potassium ion channel Kir2.1 in HEK293 cells [<i>M. musculus</i>]	1.70E-7 (5.27E-9)	3.08E-1 (1.70E-4)	2557	screening	Small molecule allosteric modulators of chain A-I-domain of integrin alpha1beta1 in U937 lymphoma [<i>H. sapiens</i>]	4.67E-2 (2.03E-4)	-3.27E-1 (4.21E-4)
Endocrine disorders	0.72 (7.21E-4)	126	885	confirmatory (cell-free)	Small molecule inhibitors of	1.22E-3 (6.53E-5)	8.75E-1 (2.65E-3)	1616	confirmatory	Small molecule inhibitors of polyglutamine	9.15E-2 (1.75E-3)	5.17E-1 (3.13E-3)

SOC	Anchor assay					Secondary assay						
	Mean LOOCV AUC (SEM)	Drugs ingredients in model	AID	Assay type	Objective & target	Mean p-value (SEM)	Mean coefficient (SEM)	AID	Assay type	Objective & target	Mean p-value (SEM)	Mean coefficient (SEM)
Eye disorders	0.72 (1.74E-4)	286	2553	screening	cytochrome P450 3A4 [<i>H. sapiens</i>]	5.04E-4 (1.09E-5)	1.97E-1 (2.23E-4)	1566	confirmatory	aggregation [<i>S. cerevisiae</i>]	6.98E-2 (5.44E-4)	2.47E-1 (7.16E-4)
					Small molecule inhibitors of transient receptor potential cation channel C6 (TRPC6) in HEK293 cells [<i>M. musculus</i>]					Small molecule inhibitors of NOD2 nucleotide-binding oligomerization domain in HEK-293 cells [<i>M. musculus</i>]		
Skin and subcutaneous tissue disorders	0.70 (1.96E-4)	250	781	screening (cell-free)	Small molecule inhibitors of 14-3-3 phosphoserine/threonine binding proteins - Bad interactions [<i>B. taurus</i>]	6.23E-5 (1.37E-6)	4.95E-1 (4.66E-4)	818	screening	Small molecules cytotoxic to HCT116 colon tumor cells Lacking beta catenin [<i>H. sapiens</i>]	6.99E-3 (3.93E-5)	-6.17E-1 (1.05E-3)
Metabolism and nutrition disorders	0.69 (2.16E-4)	366	1601	confirmatory	Small molecule inhibitors of mutant SOD1-mediated cell death in neuroblastoma neuro 2A cells [<i>M. musculus</i>]	1.21E-3 (1.22E-5)	6.76E-1 (6.13E-4)	1581	confirmatory	Small molecule inhibitors of PAR4 - aPKC interactions [<i>E. coli</i>]	5.56E-2 (3.30E-4)	3.90E-1 (5.29E-4)
Psychiatric disorders	0.69 (1.26E-4)	380	1530	screening (cell-free)	Small molecule inhibitors of MEK2-MEK5 PB1 domain interactions [<i>H. sapiens</i>]	1.46E-5 (2.05E-7)	5.97E-1 (3.48E-4)	1621	confirmatory	Small molecule inhibitors of West Nile virus cytopathicity in Vero E6 cells [<i>C. aethiops</i>]	3.0E-1 (2.15E-3)	-5.57E-1 (3.56E-2)
Renal and urinary disorders	0.68 (4.15E-4)	189	1242	screening	Small molecule potentiators of clotrimazole action in multidrug-tolerant <i>C. albicans</i> cells	4.85E-4 (7.99E-6)	8.41E-1 (1.35E-3)	567	screening	Small molecule agonists of 5-hydroxytryptamine receptor subtype 1a (5HT 1a) [<i>H. sapiens</i>]	1.02E-1 (2.90E-3)	3.60E-1 (1.99E-3)
Vascular disorders	0.68 (2.98E-4)	263	760	screening (cell-free)	Small molecule inhibitors of wildtype Rab2 [<i>C. lupus familiaris</i>]	2.68E-3 (2.30E-5)	-1.35E+0 (1.96E-3)	2242	confirmatory (cell-free)	Small molecule activators of acid alpha-glucosidase preprotein [<i>H. sapiens</i>]	6.33E-3 (1.25E-4)	3.17E-1 (4.60E-4)
General disorders and administration site conditions	0.66 (2.52E-2)	460	2445	screening	Small molecule potentiators of	2.35E-3 (9.28E-4)	5.24E-1 (6.49E-4)	2066	screening	Small molecule modulators of general amino acid	5.81E-1 (9.21E-4)	-1.94E+0 (1.17E-1)

SOC	Anchor assay					Secondary assay						
	Mean LOOCV AUC (SEM)	Drugs ingredients in model	AID	Assay type	Objective & target	Mean p-value (SEM)	Mean coefficient (SEM)	AID	Assay type	Objective & target	Mean p-value (SEM)	Mean coefficient (SEM)
					oxytocin receptor [<i>H. sapiens</i>]					permease AGP1 [<i>S. cerevisiae</i>]		
Gastrointestinal disorders	0.65 (9.9E-5)	440	2066	screening	small molecule modulators of protein targets in the rapamycin target-containing pathway [<i>S. cerevisiae</i>]	1.21E-5 (3.82E-7)	1.05E+0 (6.04E-4)	1947	screening (cell-free)	Small molecule inhibitors of Fam108B serine hydrolase [<i>M. musculus</i>]	1.02E-1 (2.43E-4)	-5.70E-1 (6.38E-4)
Cardiac disorders	0.63 (2.29E-4)	269	1359	screening	Small molecule modulators of neuropeptide Y receptor Y2 [<i>H. sapiens</i>]	1.76E-4 (2.64E-6)	2.43E-1 (3.48E-4)	2557	screening	Small molecule allosteric modulators of chain A - I-domain of integrin alpha1beta1 in U937 lymphoma [<i>H. sapiens</i>]	5.35E-2 (3.54E-4)	-2.01E-1 (3.76E-4)
Ear and labyrinth disorders	0.63 (1.47E-4)	366	1588	confirmatory (cell-free)	Small molecule inhibitors of polyglutamine aggregation	1.14E-3 (1.13E-5)	-4.39E-1 (3.90E-4)	1584	confirmatory (cell-free)	Small molecule inhibitors of mutant huntingtin aggregation	6.30E-2 (3.75E-4)	-2.57E-1 (3.78E-4)
Musculoskeletal and connective tissue disorders	0.61 (1.39E-4)	431	2101	confirmatory (cell-free)	Small molecule modulators of N370S glucocerebrosidase [<i>H. sapiens</i>]	2.24E-1 (1.77E-4)	1.05E-3 (1.79E-5)	1478	confirmatory (cell-free)	Small molecule inhibitors of cruzain cysteine protease [<i>T. cruzi</i>]	-2.25E+0 (3.63E-4)	9.88E-1 (7.02E-6)
Hepatobiliary disorders	0.60 (1.10E-4)	437	2391	confirmatory	Small molecule inhibitors of respiratory syncytial virus (RSV) cytopathicity in Hep2 laryngeal carcinoma cells [<i>H. sapiens</i>]	7.75E-4 (8.71E-6)	5.14E-1 (4.18E-4)	2435	screening	Small molecule agonists of human oxytocin receptor in CHO cells	2.05E-1 (4.03E-4)	-3.70E-1 (9.38E-4)

Table 2

FDA-approved drugs predicted to manifest unrecognized adversity

SOC	Drug ingredient	Observed PRR	logistic probability of PRR ²	Class	Related known ADRs	
					FDA drug label	DRUGDEX
Immune system disorders	Fluocinolone acetonide	1.26	0.74	Glucocorticoid topical anti-inflammatory	"... may prolong the course and may exacerbate the severity of many viral infections of the eye (including herpes simplex)"	Possible exacerbation of viral infections
Blood and lymphatic system disorders	Cisplatin	1.55	0.93	Antineoplastic	Myelosuppression	
Reproductive system and breast disorders	Loratadine	0.70	0.80	Antihistaminic	Dysmenorrhea (low incidence)	
Neoplasms benign, malignant and unspecified	Hydroquinone	0.00	0.68	Hyperpigmentation agent		
Nervous system disorders	Papaverine hydrochloride	0.28	0.70	Peripheral Vasodilator	Vertigo, headache, excessive sedation	Major: Raised intracranial pressure. Minor: Headache, somnolence, vertigo
Endocrine disorders	Pyrimethamine	0.00	0.62	Antimalarial		
Eye disorders	Clioquinol	1.92	0.56	Antibacterial, antifungal		Associated with subacute myeloptoptic neuropathy (SMON) syndrome in ethnic Japanese
Skin and subcutaneous tissue disorders	Pergolide	0.19	0.66	Dopamine Agonist	Rashes and sweating "frequently" observed	

Table 3

Predicted SOC-specific adversity for novel or recently approved drugs

Drug ingredient	SOC	FDA approval status	logistic probability of PRR ²	Class	Related known ADRs
Tranilast	Respiratory, thoracic and mediastinal disorders	Not approved	0.78	Antiasthma	
Nitazoxanide	Neoplasms benign, malignant and unspecified	2002	0.68	Antiprotozoal	
Diacerein	Skin and subcutaneous tissue disorders	Not approved	0.52	Non-steroidal anti-inflammatory	"Fatal TOXIC EPIDERMAL NECROLYSIS (LYELL'S SYNDROME), possibly induced by diacerein, was reported in a 71-year-old woman." (DRUGDEX)

Table 4

2×2 contingency table used to calculate PRR

	Drug of interest	All other drugs
ADRs associated with SOC	A	B
ADRs in other SOC's	C	D