Published in final edited form as:

Drug Saf. 2018 November ; 41(11): 1059–1072. doi:10.1007/s40264-018-0688-5.

Predicting Adverse Drug Effects from Literature- and Database-Mined Assertions

Mary K La¹, Alexander Sedykh^{2,3}, Denis Fourches⁴, Eugene Muratov², and Alexander Tropsha^{2,*}

¹Division of Practice Advancement and Clinical Education, UNC Eshelman School of Pharmacy, 301 Pharmacy Lane, Chapel Hill, NC 27599, USA

²Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, 301 Pharmacy Lane, Chapel Hill, NC 27599, USA

³Sciome LLC, 2 Davis Drive, Research Triangle Park, NC 27709, USA

⁴Department of Chemistry, North Carolina State University, Raleigh, NC, 27695, USA

Abstract

Introduction—Given that adverse drug effects have led to post-market patient harm and subsequent drug withdrawal, failure of candidate agents in the drug development process, and other negative outcomes, it is essential to attempt to forecast adverse drug effects and other relevant drug-target-effect relationships as early as possible. Current pharmacologic data sources, providing multiple complementary perspectives on the drug-target-effect paradigm, can be integrated to facilitate the inference of relationships between these entities.

Objective—This study aims to identify both existing and unknown relationships between chemicals (C), protein targets (T), and adverse drug effects (ADEs: E) based on literature evidence.

Materials and Methods—Cheminformatics and data mining approaches were employed to integrate and analyze publicly-available clinical pharmacology data and literature assertions interrelating drugs, targets, and ADEs. Based on these assertions, a C-T-E relationship knowledge base was developed. Known pairwise relationships between Cs, Ts, and Es were collected from several pharmacological and biomedical data sources. These relationships were curated and integrated according to Swanson's paradigm to form C-T-E triangles. Missing C-E edges were then inferred as chemical-ADE (C-E) relationships.

Results—Unreported associations between drugs, targets, and ADEs were inferred, and inferences were prioritized as testable hypotheses. Several chemical-ADE inferences, including *testosterone* → *myocardial infarction*, were identified using inferences based on the literature

*Corresponding Author: Alexander Tropsha, PhD, 100K Beard Hall, UNC Eshelman School of Pharmacy, Chapel Hill, NC 27514, Telephone: (919) 966-2955 | Fax: (919) 966-0204 | alex_tropsha@unc.edu.

Competing Interests:

Mary La, Alexander Sedykh, Denis Fourches, Eugene Muratov, and Alexander Tropsha declare no conflict of interest that are directly relevant to the content of this study.

sources published prior to confirmatory case reports. Timestamping approaches confirmed the predictive ability of this inference strategy on a larger scale.

Conclusions—The presented workflow, based on free-access databases and an association-based inference scheme, provided novel C-E relationships that have been validated *post-hoc* in case reports. With refinement of prioritization schemes for the generated chemical-ADE inferences, this workflow may provide an effective computational method for the early detection of potential drug candidate ADEs that can be followed by targeted experimental investigations.

Keywords

Swanson's ABC model; text mining; inference; adverse effect

1. Introduction

Adverse drug effects (ADEs) are non-preventable, negative pathological or physiological consequences resulting from the use of a drug product that may or may not manifest as adverse drug reactions (ADRs) [1]. Between 1999–2006, annual rates of ADR-related mortalities were as high as 0.12/100,000 cases in the United States [2]. Because ADRs are notoriously underreported, the annual morbidity and mortality rates may actually be much higher [2]. ADRs significantly increase hospital length of stay and overall cost of hospitalization [3], and contribute to the annual \$5.6 billion in adverse drug event costs according to the US Agency for Healthcare Research and Quality [4].

Anticipating and managing ADEs have been critical areas of research for decades. Indeed, the structure of the Food and Drug Administration's (FDA) new drug approval process allows for multiple opportunities to identify ADEs before the drug candidate reaches the market. Furthermore, post-marketing surveillance may discover additional ADEs, especially ones that occur infrequently. Nevertheless, rare but severe ADEs have caused several drugs to be withdrawn from the market, *e.g.*, the non-steroidal anti-inflammatory drug pirofen in 1990 for fatal hepatotoxicity, the antihistamine terfenadine in 1998 for prolonged QT_c interval, and the HMG-coA reductase inhibitor cerivastatin in 2001 for increased rhabdomyolysis risk [5].

Informatics techniques have been pursued to study ADEs. For instance, Don Swanson developed the following text-mining inference paradigm [6] using published literature: if an article associated chemical entity (A) with a physiological process (B), and if another article associated the same physiological process (B) with an effect (C), then an association between chemical (A) and effect (C) could be inferred. Swanson used this ABC paradigm to correctly predict that magnesium supplementation would reduce the occurrence of migraines, an inference that has been subsequently validated experimentally [7].

The publicly-accessible FDA Adverse Event Reporting System (FAERS) database [8] is a critical tool for identifying rare ADEs in post-marketing surveillance. There have been efforts to mine FAERS for drug interaction adverse events [9]. Electronic health records and Medline have also served as knowledge bases for other data mining investigations [10]. A

detailed review of available data sources for systems pharmacology is available from Bai et al. [11].

In silico strategies to predict drugs' off- and on-target interactions as well as associated adverse and therapeutic effects have been actively pursued (Table 1). A substantial number of those computational studies were dedicated to drug repurposing (Table 1, "C-T" or "C-D" in the "Goal" column). For example, Campillos *et al.* [17] used side-effect profile similarity to impute new pharmacological targets for known drugs, while Lounkine *et al.* [18] employed structural similarity of drugs. Yang *et al.* [12] used side effects as features of drugs to build classification models of drug indications, while Simon *et al.* [22] used protein interaction profiles of drugs. Moreover, there are many studies focusing on large-scale prediction of chemical-protein interactions *per se* [26,27], which can provide useful input. Another group of studies derives statistical models to predict drug side-effects (Table 1, "C-E" in the "Goal" column) based on chemical structure, drug-target interaction profiles, and even drug indications as features [13–15,24].

Finally, an increasing number of studies attempt to consider the entire network of drug/target/effect relationships (Table 1, "Network" or "Inference" in the "Methods" column), often reporting the integrated view in the form of a database [19,20]. Predictions are generated in the form of probabilities of network connections [15] or as network-based inferences [21].

Despite these advances, there are still several challenges facing informatics-based ADE prediction. Collections such as FAERS are voluntary; many ADE occurrences do not get reported due to perceived lack of time or utility on the part of the clinician or patient. Case reports may or may not be able to discern a mechanistic explanation for observed drug/ADE associations. Mining of electronic health record (EHR) data has proven fruitful for establishing ADEs [28], clinical trajectories [29], and other endpoints [30], though underlying mechanisms for these outcomes can be challenging to elucidate through this approach. Thus, additional approaches for detecting, validating, or identifying mechanisms behind ADEs should be sought.

Following the aforementioned Swanson's ABC paradigm, this study aims to identify unknown relationships between chemicals (C), protein targets (T), and ADEs (E) based on scientific reports in the literature. The "chemicals" of interest in this study are compounds recognized as approved or investigational pharmacotherapeutic agents. In our approach described in detail below, known relationships between Cs, Ts, and Es are collected initially from several pharmacological and biomedical data sources. These relationships are curated and integrated according to Swanson's paradigm to form C-T-E triangles. They are then priority-scored to facilitate inference of previously unreported chemical-ADE (C-E) relationships.

2. Materials and Methods

2.1 Chemicals (C)

The initial list of chemicals was extracted from the National Institutes of Health (NIH) Chemical Genomics Center's Pharmaceutical Collection (NPC), a publicly-accessible database containing ca. 15,000 approved and investigational drugs (<https://tripod.nih.gov/npc/>) [31]. A chemical curation process (Figure 1) pared this list down to 7,523 compounds. Since the intended focus of this study was on small-molecule organic compounds, those that were tagged in NPC as "biologics", "inorganics", or "structure undefined" (these were often mixtures) were removed. Compounds with molecular weight > 3 kDa were removed, as well. Additional inorganics, metalorganics, and mixtures were removed via Pipeline Pilot Professional Client ver. 9.1 (Dassault Systèmes, Vélizy-Villacoublay, France) and ChemAxon Standardizer ver. 6.0 (ChemAxon, Budapest, Hungary). Duplicate chemicals detected via canonical SMILES and InChIKey, forms of line notation for chemical substances, were set aside using Knime ver. 2.6.2 (KNIME.COM AG, Zurich, Switzerland). Our standard automated curation workflow [32–34] was then applied to the resulting unique active pharmaceutical ingredient (API) list, and additional errors, such as 2D structural duplicates, were analyzed by HiT QSAR [35]. Cases in which a mixture was misrepresented by only one compound were removed. Then, incorrect name-to-structure annotations, stereochemistry errors, and structures in certain cases were fixed. In addition, incorrect PubChem CIDs were fixed or removed. Secondary to our group's established workflow, multiple entries for the same API were mapped to a single chemical entity. These many-to-one relationships are represented in the online repository (https://dataverse.unc.edu/dataverse/CTE_Inferences). Although our curation workflow is automated, manual inspection of the resulting list is inherent to this process. We manually checked ~10% of the final API list and found one entry where an excipient was labeled as the primary chemical was removed, leaving 7,448 chemicals.

2.2 Targets (T)

The list of protein targets was obtained from the Therapeutic Target Database [36], another publicly-accessible database containing pharmacological information on over 2,000 known or investigational therapeutic targets. To facilitate the interoperability between multiple data sources, a restricted set of 1,280 targets with associated UniProt IDs [37] were retained for this project (Online Resource, Supplemental Figure SF1).

2.3 ADEs (E)

A list of ADEs was obtained from the Side Effect Resource (SIDER) [38], a public knowledge source listing of established (*known*) ADEs of therapeutic agents extracted from the drug products' package inserts. Only the 4,492 ADEs with unique UMLS concept IDs (Unified Medical Language System; ver. 2013AB; [39]) were retained for this project.

2.4 Establishing pairwise associations

Known C-T, T-E and C-E associations, or edges, were collected. C-T edges were obtained from two data sources: the Search Tool for Interactions of Chemicals (STITCH) [40], and

the Comparative Toxicogenomics Database (CTD) [41]. STITCH contains information on chemical-protein interactions for 300,000 small molecules, 2.6 million proteins, and over 1,000 organisms. CTD describes associations between chemicals, genes, and diseases in the framework of gene ontology and biochemical pathways, covering 887,000 chemical-gene associations, 13.5 million gene-disease associations, and 1.5 million chemical-disease associations. T-E edges were obtained from CTD, as well. Known C-E edges were extracted from SIDER, CTD, and ChemoText (CT) [42]. CT is a knowledge base of relationships between chemicals, targets, and ADEs extracted from Medline annotations, covering 1.2 million chemical-target associations, 1.9 million target-disease associations, and 9.4 million chemical-disease associations. A summary of the data sources used to obtain these C-T, T-E, and C-E edges is shown in Table 2.

2.5 Data curation and representation

Each data source may refer to its chemicals, targets, and ADEs using different identifier schemes. For example, STITCH refers to chemicals using STITCH IDs (which are derived from PubChem compound identifiers), whereas CTD uses Chemical Abstracts Service (CAS) registry numbers. Thus, a common identifier scheme for chemicals, targets, and ADEs had to be established. For chemicals, an internal identifier system (Ch0000-Ch7522) was used for the remainder of the study. Although after finishing curation, the number of chemicals was decreased to 7,448, the original Ch0000-Ch7522 chemical identification scheme was kept. UniProt IDs [37] were selected to refer to targets. UMLS concept unique identifiers and MedDRA Lowest Level Term IDs (Medical Dictionary for Regulatory Activities; accessed 2013; [43]) were used to refer to ADEs*. Interconversion between the different identifier schemes was facilitated by UniProt's ID mapping service [37] and the Chemical Translation Service [44]. Likewise, the pool of C-T, T-E, and C-E edges extracted from the various data sources were merged to consolidate information for the same entity that was represented across multiple data sources. The number of C-T, T-E and known C-E edges before and after merging are listed in Table 2. Auxiliary information stored with each edge included the number and PubMed IDs (PMIDs) of their article lists in which the edge's vertices, *e.g.*, C and T for a given C-T edge, were co-annotated. These PMIDs were obtained from the data sources informing each edge listed in Table 2.

2.6 Inference procedure

Swanson's ABC paradigm [6] was used to make C-E inferences in the following way: if an edge existed between chemical C and target T ($C \rightarrow T$), and another edge existed between T and E ($T \rightarrow E$), then an edge between C and E ($C \rightarrow E$) was inferred. It was possible for a C-E inference to have multiple linking targets. That is, C-E could be inferred not only via $C \rightarrow T_1$ and $T_1 \rightarrow E$, but also via $C \rightarrow T_2$ and $T_2 \rightarrow E$. The unique C-E inference was stored with its list of linking targets, as well as the number and PMIDs of articles which were co-annotated with the C-T and T-E edges that informed the C-E inference. After the C-E inferences were made, the entire list of C-E edges could be categorized into three sets: 1) known (*i.e.*,

*MedDRA® the Medical Dictionary for Regulatory Activities terminology is the international medical terminology developed under the auspices of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

previously reported, and obtained from SIDER, CTD, and/or CT) and inferred C-E, 2) known and not inferred C-E, and 3) inferred and not known C-E.

2.7 Inference scoring

To prioritize C-E inferences for further investigation, three scoring systems were considered for ranking the C-E inferences. Article counts (AC) described the number of unique articles in which C-T and/or T-E were co-annotated. For example, if chemical C_I was mentioned with target T_I in n articles, then C_I-T_I would have an AC of n . Several data sources from which edge relations are extracted are already annotated by the source developers. For example, CTD utilizes a combination of manual curation and text mining-assisted ranking/prioritization of articles [41], and ChemoText utilizes the Medline annotations designated by National Library of Medicine indexers [42]. Linking target counts (LTC) described the number of linking targets bridging the C-E inference. For instance, if chemical C_I was linked with effect E_I via T_1 , T_2 , T_3 , T_4 and T_5 , then C_I-E_I would have an LTC of 5. Mutual information (MI) relies on article count tables, and reflects how well-connected two vertices are to each other (see Online Resource) [45]. Receiver operating characteristic (ROC), ROC enrichment, and precision-recall (PR) curves were used to compare the performance of each scoring system and select one scoring system for ranking C-E inferences. Known C-E edges were used as the “true” cases, while the inferred C-E edges were decoy cases. The prioritization procedure ranks all inferences (based on a particular metric score), then takes various top portions (L) of the ranked list as the “predicted positive” part (with the remainder $\wedge L$ being “predicted negative”). This affords calculation of “true positives” – known associations in L, “false positives” – other inferences in L, “false negatives” – known association in $\wedge L$, and “true negatives” – other inferences in $\wedge L$. This process is repeated for progressively larger portions of the ranked list and essentially reveals how well the scoring method retrieves known associations.

2.8 Substudy 1 - Restriction of target and ADE lists

To facilitate closer analysis of the C-E inferences, the lists of Ts and Es were reduced. The known C-T and T-E edges were analyzed to find the top 100 occurring Ts in each set of known C-T and T-E edges. The union of these sets of Ts would be retained from the original target list. Similarly, the set of Es was filtered to retain the “pharmacovigilance-relevant” ADEs, which have been associated with drugs’ market withdrawal, black box warnings, and/or hospitalizations based on analysis of scientific literature, FAERS, and other biomedical references [46,47]. It was anticipated *a priori* that there would be many inferred C-E edges, and narrowing down potential sets of C-E inferences to be inspected further helped make the additional investigations more tractable. Also, not all targets and effects are clinically informative (e.g. *nausea* is fairly non-specific). Selecting a reduced set of targets and effects to evaluate further serves as a pilot study for the ability to view these relations at multiple levels of granularity.

2.9 Substudy 2 – Restriction of ADE list only

Out of concern that limiting the set of Ts to just the most promiscuous targets would result in “noisy” inferences (i.e., many C-E inferences would be ranked equally by LTC), the

restrictions on the T list were lifted, and just limited the set of Es to those that were pharmacovigilance-relevant as described in Substudy 1.

2.10 Substudy 3 – Inference and inference strategy validation

This inference strategy was validated in several ways. Case reports linking C and E for a given C-E inference, and published after the C-E inference was made, provided primary literature confirmation for some C-E inferences. The proportion of known C-E associations captured by inferred C-E associations was computed.

Additional inference strategy validation was conducted via a time-split validation technique, with the aims of 1) validating the forward-prediction abilities of the ABC inference strategy, and 2) evaluating the proportion of C-E inferences that could be confirmed over time. All edges from the described integrated knowledge base were further characterized by timestamping, which overall sought to identify the earliest year by which an edge could be established. For known C-T, T-E, and C-E edges with article lists (*i.e.*, papers that co-annotate or establish an association between known C-T, T-E, or C-E), their timestamps were determined by the earliest publication year of the papers in their article lists. If a known edge was listed in its data source(s) without an article list, no timestamp was assigned. For inferred C-E associations, their timestamps were determined by the later of the constituent C-T and T-E edges' timestamps. That is, the timestamp represents the latest year at which there was enough information on C-T and/or T-E so that the C-E inference could be made. If either constituent C-T or T-E edge did not have a timestamp, the inferred C→E edges timestamp defaulted to that of whichever edge did have an assigned timestamp.

The timestamping validation was conducted as shown in Figure 2.

3. Results

From the selected data sources (Table 2) and after additional rounds of chemical curation, 7,448 chemicals, 1,280 targets, and 4,492 ADEs, as well as 76,400 C-T edges, 1,128,542 T-E edges, and 362,408 known C-E edges, were extracted. Supplementary Figure SF2 depicts the extent of entities covered by the identified edges based on raw data. It is acknowledged that the biomedical literature grows each day with new C→E associations and other discoveries. Significant changes to our results are not expected, but minor changes in size of data and some of the derivative values may occur.

Among chemicals with at least one connection, 44% were associated with 5 or fewer targets; at the same time, 39% of such chemicals had associations with over 100 ADEs (Online Resource, Supplementary Figure SF3). Once information regarding vertices (chemicals, targets, and ADEs) and edges was compiled and organized into CTE triangles, 5,841,274 unique C-E associations were inferred. These correspond to ~17% of all possible associations in a matrix of 7,448 chemicals and 4,492 ADEs, *i.e.*, 83% sparsity. This CTE network is visualized in Figure 3. There were 341,453 C-E edges that were both previously reported (known) and inferred from the ABC strategy (Figure 4). Inference of C-E edges via the ABC paradigm was able to capture 94% of the known C-E edges. The majority of inferred C-E edges (57%) had few (≤ 5) linking targets.

3.1 Substudy 1 – Restriction of Target and ADE Lists

The restriction of the set of Ts provided a list of 153 targets (see online repository at https://dataverse.unc.edu/dataverse/CTE_Inferences), retained from the original (n = 1,280) target list. From the C-T edge list, the most common Ts were *CYP3A4*, *TNF-α*, and *MDR1*. The set of Es was filtered to retain 164 ADEs that were “pharmacovigilance-relevant” (see online repository at https://dataverse.unc.edu/dataverse/CTE_Inferences). There were 47,078 C→E inferences remaining after the number of Ts and Es were reduced (new sparsity: 96%). Pulmonary valve incompetence, thrombocytopenic purpura, and aortic valve incompetence were the three most common ADEs.

Of the three scoring systems considered, article count appeared to have the best overall performance (by area under the ROC) in this reduced C-E inference set, though linking target count had the best early enrichment (*i.e.*, more known C-E associations ranked higher in priority listing). The top 45 linking target count-prioritized C-E inferences (“generated hypotheses”) were selected for further investigation (Online Resource, Supplemental Table ST1).

3.2 Substudy 2 – Restriction of ADE List Only

When just limiting the set of Es to those that were pharmacovigilance-relevant (164 ADEs), there were 286,815 C-E inferences remaining from the original ~6 million C-E inferences (new sparsity: 77%). There were 76 unique ADEs represented. Drug-induced liver disease, renal failure acute, acute renal insufficiency, and liver disorder were the most common ADEs on this list.

For this E-only limited C-E inference list, linking target count appeared to have the best overall performance for separating known inferences (area under the ROC = 0.84, Figure 5(a)), though mutual information had the best early enrichment (~100-fold at 0.01% of the ranked list, Figure 5(b)). The C-E inferences were thus sorted by mutual information, and the top ten prioritized C-E inferences by each scoring system are shown in Table 3.

3.3 Substudy 3 – Inference and inference strategy validation

Case studies were identified from primary literature published after the inference was made (the latter date being, at the latest, January 11, 2013). When the restricted Ts/restricted Es list of C-E inferences were prioritized by linking target count, the inference between paroxetine and thrombocytopenic purpura (TP) ranked first. Paroxetine, a selective serotonin reuptake inhibitor, was linked to TP via 22 targets (see Figure 3). These targets include caspase-3 (CASP3) and signal transducer and activator of transcription 3 (STAT3). Paroxetine increases CASP3 activation. CASP3 activation is associated with idiopathic thrombocytopenic purpura [48], and its inhibition is associated with blocking apoptosis mediated by plasma from thrombotic thrombocytopenic purpura and hemolytic-uremia syndrome patients [49]. Paroxetine suppresses STAT3 expression [50], and decreased STAT3 dimerization and general lack of JAK2 pathway activation is associated with thrombocytopenic purpura [51,52].

When the restricted Es list of C-E inferences was prioritized by mutual information, the inference between testosterone and myocardial infarction via 183 linking targets ranked ninth. These targets include tumor necrosis factor alpha (TNF- α) and c-Fos. Testosterone increases TNF- α production [53]. While the relationship between increased TNF- α production and myocardial infarction is not definitively a causative one, increased levels of the inflammatory cytokine have been observed during initial and recurrent myocardial infarctions [54], and have been associated with post-event myocardial injury [55]. Similarly, testosterone induces c-Fos expression [56], which may have a role in myocardial infarction pathogenesis [57]. The FDA has required labeling changes regarding the potential cardiovascular risk of testosterone [58], in part due to two independent retrospective studies [59,60] that found an association between testosterone and negative cardiovascular outcomes, including myocardial infarction and stroke. However, these studies' findings have come under controversy because of methodological concerns, as well as prior evidence that 1) low testosterone levels have been associated with increased cardiovascular events, and 2) testosterone replacement therapy for patients with low testosterone has been associated with reduced cardiovascular events [61].

The top 50 mutual information-ranked C-E inferences (again, restricted by pharmacovigilance-relevant Es) were selected for further investigation (Online Resource, Supplemental Table ST2). When reviewing the biomedical literature for inference validation, 9 (18%) of these inferences appear to have a causal relationship, 9 (18%) appear to have a protective/effect-mitigating relationship, 6 (12%) appear to have mixed effects, 4 (8%) appeared to have no effect (*i.e.*, incorrect inference), and 22 (44%) could not be confirmed. This does not mean that these inferences were erroneous, but no articles that co-annotated the drug and effect could be found to the authors' knowledge as of this writing.

Other inferences that were validated in the literature reports that appeared after the publications used for making the inference include ciprofloxacin $\rightarrow$ posterior reversible encephalopathy syndrome (PRES) [62], and rifaximin $\rightarrow$ toxic epidermal necrolysis syndrome (TENS) [63].

Following the workflow illustrated in Figure 2, the first aim of time-split validation sought to evaluate how well the ABC inference strategy predicts later-known C-E associations. Consider the example of a cutoff year of 1992 (Step 2 of Figure 2). There were 241,817 known C $\rightarrow$ E associations established (*i.e.*, with timestamp) after 1992 (Step 3). Based on the C $\rightarrow$ T and T $\rightarrow$ E associations established prior to 1992, there were 497,537 C $\rightarrow$ E inferences that could be made prior to 1992, of which 69,844 C $\rightarrow$ E inferences were also confirmed after 1992 (Steps 4–5). This yields a post-1992 coverage rate of $69,844 / 241,817 = 29\%$, *i.e.*, 29% of C-E associations established after 1992 could have been inferred using the described ABC strategy prior to 1992 (Step 6). In addition, the time-split validation assessed the rate of C $\rightarrow$ E inferences that would later be validated post-cutoff year. There is a validation rate of the pre-1992 C $\rightarrow$ E inferences of $69,844 / 497,537 = 16\%$, *i.e.*, 16% of C-E inferences that could have been made prior to 1992 were later confirmed (Step 6). These steps were repeated in 5-year increments for the years 1987–2007 (Figure 6).

4. Discussion

In this work, publicly-available knowledge bases describing known associations between pharmacotherapeutic agents, targets, and ADEs were used to infer unknown associations between drugs and ADEs. While our strategy was able to recover 94% of previously reported C-E associations, there was still a small fraction (~21K) of known C-E associations that remained unpredicted. This may occur if no targets are discovered during the initial data collection and curation steps that link C to E, or it could happen for known C-E associations if the pathophysiology behind the association has not been fully elucidated. For example, the bisphosphonate alendronic acid is known to cause osteonecrosis of jaw (ONJ), and this pair is on our list of previously-reported C-E associations. However, it could not be inferred through the ABC strategy since the alendronic acid-T associations and T-ONJ shared no common linking targets. The mechanism linking bisphosphonates to ONJ is still not well understood [64].

The results of our study compare favorably with those of previous computational approaches to ADE prediction. Yamanishi and colleagues [13] used several approaches to analyze 658 drugs, 969 ADEs, and 23,061 drug-ADE associations. The best-performing one yielded an area under the precision-recall curve (AUPR) of 0.21 ± 0.0024 . When scoring by mutual information, our inference strategy yielded an AUPR of 0.48 (Figure 5(c)). Mathur *et al.* [25] used network analysis techniques on known relations between drugs, targets, and diseases to predict associations between biological processes and drugs. Several databases, including Swiss-Prot, OMIM, and GeneRIF, were used to inform known relationships. They report 2,222 drug-disease associations involving 76 drugs and 169 diseases, of which 144 (6.4%) were confirmed in DrugBank [25]. Of the ~6 million C-E inferences produced with the ABC strategy described above, 341,453 (5.8%) C-E inferences were confirmed in SIDER [38], ChemoText [42], and/or CTD [41] (Figure 4). Despite our lower confirmed C-E rate, a larger number of chemicals and effects were included for consideration in this study, with the C-E inferences involving 4,053 chemicals and 1,743 effects (raw data). “Chemicals” include not only drugs but also drug-like compounds, whose physiologic effects may not be as well characterized as drugs in most of the available pharmacology or ADE databases.

The time-split validation identified “novel” C-E inferences at 5-year time points from 1987 to 2007 (Figure 6). The percentage of known C-E associations that could have been predicted earlier increased over time, while the percentage of inferred C-E associations that were later confirmed appeared to decrease over time. Both these trends are likely due to the natural growth of information (discoveries about C-T and T-E) available over time, combined with the decreasing number of C-E associations established after the cutoff year, as the cutoff year increases.

There are several aspects of this approach that are of benefit in clinical pharmacovigilance. A case report can take some time to develop from observation to publication, and therefore into discoverable public knowledge. The approach in the present work provides a method for using data already gathered to potentially corroborate an ADE noticed in practice, in less time than it would typically require for said case report to get published and discovered. The described inference approach was designed to draw from publicly-available (*i.e.*, free to

access) knowledge bases; access to proprietary databases may be cost-prohibitive and difficult to maintain long-term. Our study seeks to build on the multiple *in silico* studies [12–25,28–30,65–72] that have evaluated the CTE relationship by integrating heterogeneous data resources bridging molecular pharmacology through clinical surveillance in scope, and ranging from traditional biochemical studies through text mining (*e.g.*, ChemoText [42]) in origin of assertions. While limits can be imposed on the types of drugs, targets, and ADEs returned, our resource of integrated databases was not built with these limits in place and therefore affords some measure of flexibility by interest area.

This study had certain limitations. The inference strategy is limited by the “universe” of knowledge it is based on, which affects both quantity and quality of inferences. Only an *association* between a C and an E can be inferred; it is not possible to distinguish, without further characterization of edges, between causative, therapeutic, or no-effect C-E associations. It may be computationally expensive to update the current collection of C-T, T-E, known C-E, and inferred C-E edges, since new associations would need to be made, integrated into the current framework, and have scores revised. This inference strategy does not replace clinical judgment; in fact, manual intervention is required to process even the prioritized list of C-E inferences. For example, “Convulsion”, “Convulsions generalized”, “Clonic convulsion”, and “Convulsive seizure” are all considered separate concepts in this CTE system, but have been manually grouped together as seizure, especially if they were all inferred to be associated with the same chemical with the same ranking score. This explains why there are some older dates listed in the “previously unreported” C-E association list (see Online Resource, Supplemental Tables ST1 and ST2), in that the C-E association could have been identified and linked with an ADE synonym not accounted for in the current framework. Moreover, this inference strategy does not attempt to predict the dose at which an ADE would occur for a particular agent, nor does it predict the frequency of this ADE occurring with exposure to the agent. A scoring system that integrates the optimal characteristics of the metrics featured in the present work may be able to provide a more refined prioritized C-E inference list.

There are several future directions that could be explored. Identifying other subsets of C-E inferences for further study, as well as improving the scoring system used to prioritize C-E inferences, may be instructive. One way would be to obtain prioritization ranks based on each metric and then work with these ranked lists, using various rank-merging algorithms. Though pharmacovigilance-relevant ADEs were the focus of the current study, it would be interesting to investigate inferences implicating the top 200 marketed drugs, or to focus on inferences involving G-protein coupled receptors, many of which are drug targets and/or mechanisms of drug interactions. The same inference strategy could be used for drugs in the development pipeline, to identify potential ADEs that the drug could cause prior to *in vivo* testing. New ADE predictions could be made or strengthened based on structural similarity with chemicals participating in known C-E associations. Rules of similarity for chemicals sharing the same T and E could be rationalized. Inferences could be made for not only C-E associations, but also C-T and T-E edges. These studies may help with elucidating a drug’s mechanism of action or off-target effects, or link pharmacological action at a target with a specific ADE. The results of these hypothesis-generating investigations could be further validated (or refuted) using basic science, clinical, and computational approaches.

5. Conclusion

The increasing costs of bringing new drug entities to market strongly motivate the search for *in silico* methods of identifying ADEs of a drug candidate early in the development process. We created and characterized an integrated knowledge base of chemical-target, target-ADE, and chemical-ADE associations, grounded in publicly-available pharmacology databases. We used this knowledge base to investigate the feasibility and performance of Swanson's ABC inference paradigm to predict ADEs for selected drugs and drug-like compounds. This inference strategy generated a large volume of chemical-ADE hypotheses, which were prioritized for further investigation using scoring methods. Many inferences were validated *post-hoc*, as demonstrated by identified case reports and our timestamping validation. Our strategy could be used to identify previously unreported chemical-ADE associations for further investigation. Given the hypothesis-generating nature of the study, this strategy is well-suited for user groups in the research, development, and academia spaces as a springboard for further investigations into the mechanisms behind a C-T, T-E, or C-E association, or groups of these relations thereof. Regulatory groups such as the FDA might find utility in this knowledge base as one resource that might lend further credence to a detected pharmacovigilance signal, as this resource represents the integration of pharmacological and biomedical literature data. This has implications in saving time and cost in drug development, as well as patient outcomes in terms of being better equipped to anticipate ADEs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Mr. Alexander Gartland for fruitful discussions that helped to improve the quality of the manuscript.

The MedDRA® trademark is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) on behalf of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

Funding:

This study was supported in part by the National Institutes of Health (grant 1U01CA207160).

References

1. Aronson JK. Distinguishing hazards and harms, adverse drug effects and adverse drug reactions: Implications for drug development, clinical trials, pharmacovigilance, biomarkers, and monitoring. *Drug Saf.* 2013; 36:147–53. [PubMed: 23417506]
2. Shepherd G, Mohorn P, Yacoub K, May DW. Adverse drug reaction deaths reported in United States vital statistics, 1999–2006. *Ann Pharmacother.* 2012; 46:169–75. [PubMed: 22253191]
3. Suh DC, Woodall BS, Shin SK, Hermes-De Santis ER. Clinical and economic impact of adverse drug reactions in hospitalized patients. *Ann Pharmacother.* 2000; 34:1373–9. [PubMed: 11144691]
4. Agency for Healthcare Research and Quality. Reducing and Preventing Adverse Drug Events To Decrease Hospital Costs [Internet]. AHRQ Arch; Available from: <https://archive.ahrq.gov/research/findings/factsheets/errors-safety/aderia/ade.html> [Accessed 02 Apr 2014]

5. Ninan B, Wertheimer A. Withdrawing Drugs in the U.S. Versus Other Countries. *Innov Pharm* [Internet]. 2012; 3(3) [Accessed 19 Apr 2018] Article 87. Available from: <https://pubs.lib.umn.edu/index.php/innovations/article/view/269/263>.
6. Swanson DR. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med*. 1988; 31:526–57. [PubMed: 3075738]
7. Ramadan NM, Halvorson H, Vande-Linde A, Levine SR, Helpert JA, Welch KM. Low brain magnesium in migraine. *Headache*. 1989; 29:416–9. [PubMed: 2759849]
8. US Food and Drug Administration. [Accessed 19 Apr 2018] openFDA Datasets: FAERS [Internet]. Available from: <https://open.fda.gov/data/faers/>
9. Harpaz R, Haerian K, Chase HS, Friedman C. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annu Symp Proc AMIA Symp*. 2010; 2010:281–5. [PubMed: 21346985]
10. Schuemie MJ, Coloma PM, Straatman H, Herings RMC, Trifirò G, Matthews JN, et al. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Med Care*. 2012; 50:890–7. [PubMed: 22929992]
11. Bai JPF, Abernethy DR. Systems pharmacology to predict drug toxicity: integration across levels of biological organization. *Annu Rev Pharmacol Toxicol*. 2013; 53:451–73. [PubMed: 23140241]
12. Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. *PLoS One*. 2011; 6:e28025. [PubMed: 22205936]
13. Yamanishi Y, Pauwels E, Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces. *J Chem Inf Model*. 2012; 52:3284–92. [PubMed: 23157436]
14. Cheng F, Li W, Wang X, Zhou Y, Wu Z, Shen J, et al. Adverse drug events: database construction and in silico prediction. *J Chem Inf Model*. 2013; 53:744–52. [PubMed: 23521697]
15. Cami A, Arnold A, Manzi S, Reis B. Predicting Adverse Drug Events Using Pharmacological Network Models. *Sci Transl Med*. 2011; 3:114ra127.
16. Chen B, Ding Y, Wild DJ. Assessing Drug Target Association Using Semantic Linked Data. *PLoS Comput Biol*. 2012; 8:e1002574. [PubMed: 22859915]
17. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science*. 2008; 321:263–6. [PubMed: 18621671]
18. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*. 2012; 486:361–7. [PubMed: 22722194]
19. Oprea TI, Nielsen SK, Ursu O, Yang JJ, Taboureau O, Mathias SL, et al. Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. *Mol Inform*. 2011; 30:100–11. [PubMed: 22287994]
20. Davis AP, Wiegiers TC, Roberts PM, King BL, Lay JM, Lennon-Hopkins K, et al. A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database*. 2013; 2013:bat080. [PubMed: 24288140]
21. King BL, Davis AP, Rosenstein MC, Wiegiers TC, Mattingly CJ. Ranking Transitive Chemical-Disease Inferences Using Local Network Topology in the Comparative Toxicogenomics Database. *PLoS One*. 2012; 7:e46524. [PubMed: 23144783]
22. Simon Z, Peragovics A, Vigh-Smeller M, Csukly G, Tombor L, Yang Z, et al. Drug effect prediction by polypharmacology-based interaction profiling. *J Chem Inf Model*. 2012; 52:134–45. [PubMed: 22098080]
23. Wang Y, Chen S, Deng N, Wang Y. Drug Repositioning by Kernel-Based Integration of Molecular Structure, Molecular Activity, and Phenotype Data. *PLoS One*. 2013; 8:e78518. [PubMed: 24244318]
24. Wallach I, Jaitly N, Lilien R. A Structure-Based Approach for Mapping Adverse Drug Reactions to the Perturbation of Underlying Biological Pathways. *PLoS One*. 2010; 5:e12063. [PubMed: 20808786]
25. Mathur S, Dinakarpandian D. Drug Repositioning using Disease Associated Biological Processes and Network Analysis of Drug Targets. *AMIA Annu Symp Proc*. 2011; 2011:305–11. [PubMed: 22195082]

26. Kim Kjærulff S, Wich L, Kringelum J, Jacobsen UP, Kouskoumvekaki I, Audouze K, et al. ChemProt-2.0: visual navigation in a disease chemical biology database. *Nucleic Acids Res.* 2013; 41:D464–469. [PubMed: 23185041]
27. Jacob L, Vert J-P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinforma Oxf Engl.* 2008; 24:2149–56.
28. Lo HZ, Ding W, Nazeri Z. Mining Adverse Drug Reactions from Electronic Health Records. 2013 IEEE 13th Int Conf Data Min Workshop; 2013. 1137–40.
29. Jensen K, Soguero-Ruiz C, Mikalsen KO, Lindsetmo R-O, Kouskoumvekaki I, Girolami M, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep.* 2017; 7:46226. [PubMed: 28387314]
30. Yadav P, Steinbach M, Kumar V, Simon G. Mining Electronic Health Records (EHRs): A Survey. *ACM Comput Surv.* 2018; 50:85:1–85:40.
31. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, et al. The NCGC Pharmaceutical Collection: A comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci Transl Med.* 2011; 3:80ps16. Available from: <https://tripod.nih.gov/npc/>.
32. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model.* 2010; 50:1189–204. [PubMed: 20572635]
33. Fourches D, Muratov E, Tropsha A. Curation of chemogenomics data. *Nat Chem Biol.* 2015; 11:535. [PubMed: 26196763]
34. Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model.* 2016; 56:1243–52. [PubMed: 27280890]
35. Kuz'min VE, Artemenko AG, Muratov EN. Hierarchical QSAR technology based on the Simplex representation of molecular structure. *J Comput Aided Mol Des.* 2008; 22:403–21. [PubMed: 18253701]
36. Yang H, Qin C, Li YH, Tao L, Zhou J, Yu CY, et al. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.* 2016; 44:D1069–74. [PubMed: 26578601]
37. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* 2015; 43:D204–212. [PubMed: 25348405]
38. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol.* 2010; 6:343. [PubMed: 20087340]
39. National Library of Medicine. [Accessed 2015 Mar 02] Unified Medical Language System (UMLS) [Internet]. 2013. Available from: <https://www.nlm.nih.gov/research/umls/>
40. Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P. STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Res.* 2012; 40:D876–80. [PubMed: 22075997]
41. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, et al. The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.* 2013; 41:D1104–1114. [PubMed: 23093600]
42. Baker NC, Hemminger BM. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J Biomed Inform.* 2010; 43:510–9. [PubMed: 20348023]
43. [Accessed 19 Apr 2018] MedDRA [Internet]. 2013. Available from: <https://www.meddra.org/>
44. Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O. The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics.* 2010; 26:2647–8. [PubMed: 20829444]
45. Cover TM, Thomas JA. *Elem Inf Theory.* 2. Wiley Interscience; 2006. Relative entropy and mutual information; 19–20.
46. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: the importance of reporting suspected reactions. *Arch Intern Med.* 2005; 165:1363–9. [PubMed: 15983284]
47. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salamé G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Saf.* 2009; 18:1176–84. [PubMed: 19757412]

48. Houwerzijl EJ, Blom NR, van der Want JJJ, Esselink MT, Koornstra JJ, Smit JW, et al. Ultrastructural study shows morphologic features of apoptosis and para-apoptosis in megakaryocytes from patients with idiopathic thrombocytopenic purpura. *Blood*. 2004; 103:500–6. [PubMed: 12969975]
49. Mitra D, Kim J, MacLow C, Karsan A, Laurence J. Role of caspases 1 and 3 and Bcl-2-related molecules in endothelial cell apoptosis associated with thrombotic microangiopathies. *Am J Hematol*. 1998; 59:279–87. [PubMed: 9840908]
50. Taler M, Gil-Ad I, Lomnitski L, Korov I, Baharav E, Bar M, et al. Immunomodulatory effect of selective serotonin reuptake inhibitors (SSRIs) on human T lymphocyte function and gene expression. *Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol*. 2007; 17:774–80.
51. Metjian A, Abrams CS. New advances in the treatment of adult chronic immune thrombocytopenic purpura: role of thrombopoietin receptor-stimulating agents. *Biol Targets Ther*. 2009; 3:499–513.
52. Schipperus M, Fijnheer R. New therapeutic options for immune thrombocytopenia. *Neth J Med*. 2011; 69:480–5. [PubMed: 22173361]
53. Metcalfe PD, Leslie JA, Campbell MT, Meldrum DR, Hile KL, Meldrum KK. Testosterone exacerbates obstructive renal injury by stimulating TNF- α production and increasing proapoptotic and profibrotic signaling. *Am J Physiol Endocrinol Metab*. 2008; 294:E435–443. [PubMed: 18073317]
54. Ridker PM, Rifai N, Pfeffer M, Sacks F, Lepage S, Braunwald E. Elevation of tumor necrosis factor- α and increased risk of recurrent coronary events after myocardial infarction. *Circulation*. 2000; 101:2149–53. [PubMed: 10801754]
55. Li D, Zhao L, Liu M, Du X, Ding W, Zhang J, et al. Kinetics of tumor necrosis factor α in plasma and the cardioprotective effect of a monoclonal antibody to tumor necrosis factor α in acute myocardial infarction. *Am Heart J*. 1999; 137:1145–52. [PubMed: 10347344]
56. Giannakopoulou M, Bozas E, Philippidis H, Stylianopoulou F. Protooncogene c-fos Involvement in the Molecular Mechanism of Rat Brain Sexual Differentiation. *Neuroendocrinology*. 2001; 73:387–96. [PubMed: 11408780]
57. Zhang S, Zhang M, Goldstein S, Li Y, Ge J, He B, et al. The effect of c-fos on acute myocardial infarction and the significance of metoprolol intervention in a rat model. *Cell Biochem Biophys*. 2013; 65:249–55. [PubMed: 23054911]
58. US Food and Drug Administration, Center for Drug Evaluation and Research. [Accessed 19 Apr 2018] Drug Safety and Availability - FDA Drug Safety Communication: FDA cautions about using testosterone products for low testosterone due to aging; requires labeling change to inform of possible increased risk of heart attack and stroke with use [Internet]. 2015. Available from: <https://www.fda.gov/Drugs/DrugSafety/ucm436259.htm>
59. Vigen R, O'Donnell CI, Barón AE, Grunwald GK, Maddox TM, Bradley SM, et al. Association of testosterone therapy with mortality, myocardial infarction, and stroke in men with low testosterone levels. *JAMA*. 2013; 310:1829–36. [PubMed: 24193080]
60. Finkle WD, Greenland S, Ridgeway GK, Adams JL, Frasco MA, Cook MB, et al. Increased risk of non-fatal myocardial infarction following testosterone therapy prescription in men. *PloS One*. 2014; 9:e85805. [PubMed: 24489673]
61. Morgentaler A, Zitzmann M, Traish AM, Fox AW, Jones TH, Maggi M, et al. Fundamental Concepts Regarding Testosterone Deficiency and Treatment: International Expert Consensus Resolutions. *Mayo Clin Proc*. 2016; 91:881–96. [PubMed: 27313122]
62. Ali WHAB. Ciprofloxacin-associated posterior reversible encephalopathy. *BMJ Case Rep*. 2013; 2013
63. Patel AS, Supan EM, Ali SN. Toxic epidermal necrolysis associated with rifaximin. *Am J Health Syst Pharm*. 2013; 70:874–6. [PubMed: 23640348]
64. Sharma D, Ivanovski S, Slevin M, Hamlet S, Pop TS, Brinzaniuc K, et al. Bisphosphonate-related osteonecrosis of jaw (BRONJ): diagnostic criteria and possible pathogenic mechanisms of an unexpected anti-angiogenic side effect. *Vasc Cell*. 2013; 5:1. [PubMed: 23316704]
65. Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed*. 1998; 57:149–53. [PubMed: 9822851]

66. Shang N, Xu H, Rindflesch TC, Cohen T. Identifying plausible adverse drug reactions using knowledge extracted from the literature. *J Biomed Inform.* 2014; 52:293–310. [PubMed: 25046831]
67. Hristovski D, Rindflesch T, Peterlin B. Using literature-based discovery to identify novel therapeutic approaches. *Cardiovasc Hematol Agents Med Chem.* 2013; 11:14–24. [PubMed: 22845900]
68. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. *J Biomed Inform.* 2009; 42:633–43. [PubMed: 19124086]
69. Preiss J, Stevenson M, Gaizauskas R. Exploring relation types for literature-based discovery. *J Am Med Inform Assoc JAMIA.* 2015; 22:987–92. [PubMed: 25971437]
70. Tari L, Vo N, Liang S, Patel J, Baral C, Cai J. Identifying novel drug indications through automated reasoning. *PloS One.* 2012; 7:e40946. [PubMed: 22911721]
71. Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. Literature mining, ontologies and information visualization for drug repurposing. *Brief Bioinform.* 2011; 12:357–68. [PubMed: 21712342]
72. Doulaverakis C, Nikolaidis G, Kleontas A, Kompatsiaris I. Panacea, a semantic-enabled drug recommendations discovery framework. *J Biomed Semant.* 2014; 5:13.

Fig. 1.
Workflow for selecting study chemicals from the NPC database. NPC: NCGC
Pharmaceutical Collection; MW: molecular weight

Fig. 2.
Time-split validation workflow. CTE: Chemical-Target-Effect; Y: year, C-E: Chemical-Effect; C-T: Chemical-Target; T-E: Target-Effect

Fig. 3. Network visualization. (a) Network visualization of interconnected chemicals (blue), targets (green), and ADEs (red) interconnections (b) Connectivity clusters in each domain (c) A subnetwork for the paroxetine-thrombocytopenic purpura inference

Fig. 4.
Comparison of known vs. inferred C-E edges (coverage of known C→E edges = 94%)

Fig. 5. Substudy 2 scoring system performance, derived from raw data. (a) ROC curve, (b) ROC enrichment curve, and (c) Precision-Recall curve

Fig. 6. Time-split validation results. (a) Post-cutoff year coverage rates from time-split validation. C→E inferences that could be made with C→T and T→E edges published prior to certain years (1987–2007 in 5-year intervals), and their coverage of the known C→E edges published after that given year. The total number of known C→E edges established after a given cutoff year is listed above each bar. (b) Pre-cutoff year validation rates from time-split validation. The percentage of C→E inferences that could be made prior to a given cutoff year that were later validated. The total number of inferred C→E edges that could be established prior a given cutoff year is listed above each bar

Table 1

Computational studies linking drugs, targets, and side-effects/diseases.

Study	Chemicals (C)	Targets (T)	Effects (E) / Indications (D)	Goal ^a	Method ^b
Yang <i>et al</i> [12]	303	-	584/145	C-D by C-E	ML
Yamanishi <i>et al</i> [13]	658	1,368	969	C-E by C, C-T	KM, CCA
Cheng <i>et al</i> [14]	953	(1,200)	13,255	C-E	DB, Inference
Cami <i>et al</i> [15]	809	-	852	C-E	Network
Chen <i>et al</i> [16]	157	1,683	-	C-T	Network
Campillos <i>et al</i> [17]	746	(<1,000)	1,182	C-T by C-E	Similarity
Lounkine <i>et al</i> [18]	656	73 (339)	-/317	C-T(D) by C	Similarity
Oprea <i>et al</i> [19]	988	(1,750)	174/?	C-D	DB, PCA
Davis <i>et al</i> [20]	5,562	9,167	2,697	C-T-E	DB
King <i>et al</i> [21]	6,406	20,898	3,999	C-T-E	Inference
Simon <i>et al</i> [22]	1,177	149	-/177	C-D by C-T	CCA, LDA
Wang <i>et al</i> [23]	593	(<4,323)	1,450/313	C-D	KM, Similarity
Wallach <i>et al</i> [24]	730	830	506	C-E by pathway	Inference
Mathur <i>et al</i> [25]	948	>1,480	-/581	C-D by pathway	Inference

^a C-chemical, D-disease, E-(side) effect, T-target.

^b ML-machine learning methods (e.g., support vector machines), KM-Kernel methods, CCA-canonical correlation analysis, DB-database creation, PCA-principal component analysis, LDA-linear discriminant analysis.

Table 2

Data sources from which information regarding the associations between chemicals, targets, and ADEs in this study were obtained.

Entity ^a	Data source	Citation	Original # Entities ^b	# Entities Included for AE Inference ^c
C→T	Search Tool for Interactions of Chemicals (STITCH)	Kuhn <i>et al</i> [40]	787,889 associations	76,400 associations
	Comparative Toxicogenomics Database (CTD)	Davis <i>et al</i> [41]	1,435,948 associations	
T→E	CTD	(see above)	12,364,098 associations	1,128,542 associations
Known C→E	ChemoText (CT)	Baker <i>et al</i> [42]	10,916 associations	362,408 associations
	CTD	(see above)	4,269,771 associations	
	SIDER	Kuhn <i>et al</i> [38]	520,145 associations	

^a For the details on C, T, and E entries, see online repository at https://dataverse.unc.edu/dataverse/CTE_Inferences;

^b After initial data curation;

^c After final data curation.

Table 3
Top 10 C→E inferences as ranked by mutual information and linking target count

The original list of C-E inferences has been restricted in Es to retain pharmacovigilance-relevant Es (164 Es).

	Ranked by Mutual Information (MI score ^a)	Ranked by Linking Target Count (LTC score ^b)
1	doxorubicin → chest pain (8.63×10^{-5})	estradiol → confusional state (498)
2	pilocarpine → torsades de pointes (8.19×10^{-5})	clofezone → vomiting (470)
3	tretinoin → extrapyramidal disorder (6.40×10^{-5})	clofezone → hemolytic anemia (462)
4	testosterone → drug-induced liver disease (6.30×10^{-5})	clofezone → diarrhea (462)
5	testosterone → seizure [Convulsions generalized, Clonic convulsion, Convulsive seizure] (4.79×10^{-5})	clofezone → drug eruption (462)
6	indomethacin → torsades de pointes (4.16×10^{-5})	clofezone → pneumonia (460)
7	pilocarpine → sudden death (4.08×10^{-5})	clofezone → thrombosis [Venous thrombosis, Deep vein thrombosis] (458)
8	phenylephrine → thrombocytopenia (3.84×10^{-5})	clofezone → rhabdomyolysis (458)
9	testosterone → myocardial infarction (3.82×10^{-5})	clofezone → pancreatitis (457)
10	testosterone → hypotension (3.80×10^{-5})	estradiol → death (455)

^a Mutual information had mean 4.56×10^{-7} , median 1.79×10^{-8} , and range 9.67×10^{-17} to 1.96×10^{-3} ;

^b Linking target count had mean 16.7, median 6, and range 1 to 501.