Advanced Review

Drug repurposing and adverse event prediction using high-throughput literature analysis

Spyros N. Deftereos,* Christos Andronis, Ellen J. Friedla, Aris Persidis and Andreas Persidis

Drug repurposing is the process of using existing drugs in indications other than the ones they were originally designed for. It is an area of significant recent activity due to the mounting costs of traditional drug development and scarcity of new chemical entities brought to the market by bio-pharmaceutical companies. By selecting drugs that already satisfy basic toxicity, ADME and related criteria, drug repurposing promises to deliver significant value at reduced cost and in dramatically shorter time frames than is normally the case for the drug development process. The same process that results in drug repurposing can also be used for the prediction of adverse events of known or novel drugs. The analytics method is based on the description of the mechanism of action of a drug, which is then compared to the molecular mechanisms underlying all known adverse events. This review will focus on those approaches to drug repurposing and adverse event prediction that are based on the biomedical literature. Such approaches typically begin with an analysis of the literature and aim to reveal indirect relationships among seemingly unconnected biomedical entities such as genes, signaling pathways, physiological processes, and diseases. Networks of associations of these entities allow the uncovering of the molecular mechanisms underlying a disease, better understanding of the biological effects of a drug and the evaluation of its benefit/risk profile. *In silico* results can be tested in relevant cellular and animal models and, eventually, in clinical trials. © 2011 John Wiley & Sons, Inc. *WIREs Syst Biol Med* 2011 3 323–334 DOI: 10.1002/wsbm.147

INTRODUCTION

Despite enormous investments in basic science and technology, drug development still requires considerable time, in the order of 10–15 years, and costs between $500 million and $2 billion.[1–3] At the same time the number of new drugs that are approved by the U.S. Food and Drug Administration (FDA) continues to decline,[4] whereas that of reported serious or life-threatening adverse drug reactions (ADRs) increases.[5,6]

In silico analysis has been proposed as a method to predict the ADRs of new drugs[7] and to identify new uses for existing ones,[3] an approach known as 'drug repurposing' or 'drug repositioning'.[8,9] Some believe that extensive use of ADR prediction technologies could reduce the cost of drug development by 50%.[10] New uses of existing drugs, on the other hand, cost much less to develop compared with *de novo* drug discovery, mainly due to the accumulated data on their preclinical properties and their established safety profiles.[8] The use of topiramate, an anti-epileptic drug initially marketed by Johnshon&Johnshon as Topamax, for the treatment of obesity and that of fluoxetine (Prozac), the first-in-class selective

*Correspondence to: s.deftereos@biovista.com

Biovista Inc., Charlottesville, VA, USA

Based on a presentation to the 8th International Conference on Pathways, Networks and Systems Medicine. *www.aegeanconferences.org*

DOI: 10.1002/wsbm.147

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

serotonin reuptake inhibitor (SSRI), for the treatment of premenstrual dysphoria (Sarafem, Eli Lilly) are examples of repurposed drugs.[11]

Among the contemporary approaches to *in silico* analysis, literature-based discovery (LBD) has received significant attention. LBD relies on the processing of publicly accessible biomedical literature data to uncover indirect or innate relationships among seemingly unconnected biological entities.[12–14] In analogy to system biology, which examines biology in terms of the dynamic structure and inter-relationships of all the components of a cell or organism, and not the individual constituents in isolation,[15] LBD treats large sets of scientific literature as a vast system of interconnections between research parameters; it navigates through these interconnections to describe the basic molecular mechanisms underlying a disease, to better understand the biological effects of a drug, to evaluate its benefit/risk profile, and to arrive at novel discoveries. These can be either repurposing opportunities or unreported ADRs. In this paper, we review algorithms rooted in LBD, including systems literature analysis (SLA), a workflow for drug discovery that has already yielded compounds in the preclinical stage and is being increasingly used for the prediction of ADRs.[12–14,16] We emphasize successful case studies of discoveries that have been experimentally or clinically validated.

A HISTORICAL PERSPECTIVE

Swanson's ABC Model

The possibility of linking different scientific disciplines through intermediate (or shared) concepts was first described by Swanson in 1986, in what has come to be known as the ABC model.[17] In that model A, B, and C denote separate scientific concepts, where A is reported to be related to B in one set of publications and B is reported to be related to C in another, while A is not reported to be directly related to C. The two known relations of A-to-B and B-to-C allow one to infer that A may be indirectly related to C, through B (Figure 1). The unknown A–B–C relation, from which the name of the model is derived, might thus constitute a novel scientific discovery.

Discovery under the ABC model can be pursued as either a closed or an open process. A closed discovery process begins with known starting and target concepts, A and C. The task in this case is to identify and evaluate all relevant intermediate (B) concepts that support the relation of A and C. An open discovery process, on the other hand, begins with a known starting concept A, whereas the relevant target concept C is not known beforehand and should be identified by the algorithm (Figure 1). The types of intermediate concepts used and the algorithms by which target concepts are ranked vary between implementations. One example is shown in Figure 2.

Use of this methodology led to a novel hypothesis in the late 1980s that fish oil might be beneficial in the treatment of Raynaud's syndrome.[18] A second hypothesis, that magnesium deficiency might be implicated in the pathophysiology of migraine,[19] was not entirely novel, as the topic had been discussed in earlier publications.[20–22] However, papers predating Swanson's publication were admittedly scarce. It was later shown in a double-blind placebo-controlled clinical trial that fish oil improves tolerance to cold exposure and delays the onset of vasospasm

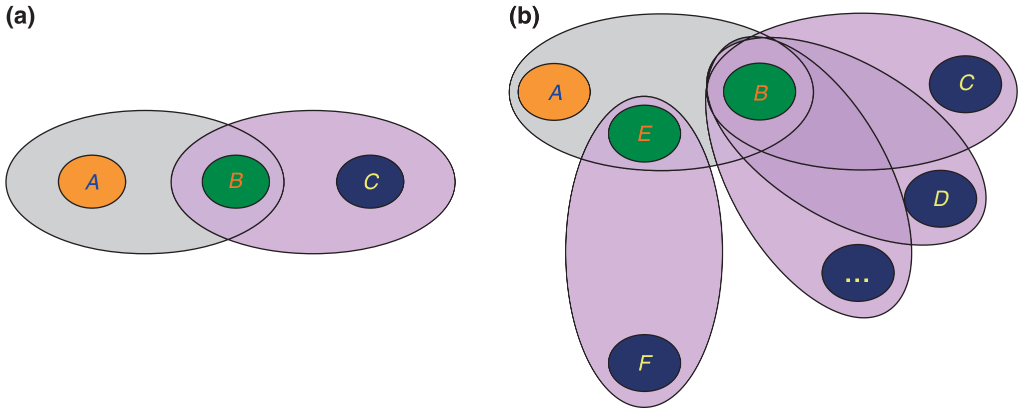

FIGURE 1 | (a) Concept A is related to concept B, as reported in one set of papers (indicated by the gray ellipse on the left), while B is related to C according to another set of papers (indicated by the purple ellipse on the right). Although A is not known to be directly related to C one can infer an indirect relation through B. As A and C are known beforehand, this is a *closed* discovery process. (b) Concept A is related to concept B, and through that to C, D, etc. Furthermore, A is related to F through E. In this *open* discovery process, the target concepts are not known beforehand, and their selection is part of the process output.

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

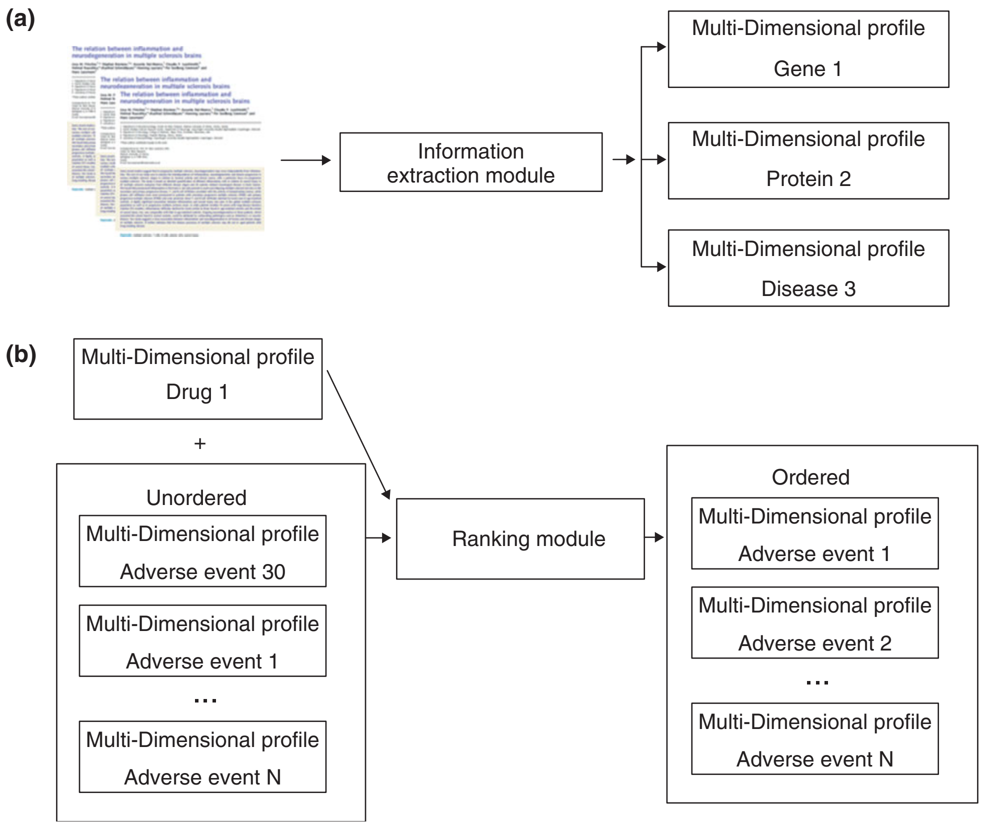
(a)
Information extraction module
Multi-Dimensional profile Gene 1
Multi-Dimensional profile Protein 2
Multi-Dimensional profile Disease 3
(b)
Multi-Dimensional profile Drug 1
+
Unordered
Multi-Dimensional profile Adverse event 30
Multi-Dimensional profile Adverse event 1
...
Multi-Dimensional profile Adverse event N
Ranking module
Ordered
Multi-Dimensional profile Adverse event 1
Multi-Dimensional profile Adverse event 2
...
Multi-Dimensional profile Adverse event N

FIGURE 2 | The Information Extraction and Ranking algorithms implemented in Biovista's Clinical Outcome Search Space (COSS) platform.

in patients with primary Raynaud's syndrome,[23] while the role of magnesium in the prophylaxis of migraine, particularly pediatric and premenstrual, is now established.[24] Swanson's group used the same methodology in 1996 to propose novel mechanistic hypotheses for the effects of estrogens[25] and indomethacin,[26] which were known at that time, in the treatment of Alzheimer's disease (AD).

The ABC model has been implemented in the *Arrowsmith* software,[27] which uses Medline as the literature corpus and seeks for occurrences of the concepts of interest in the title and MEdical Subject Headings (MESH) of publications. Here concepts are expressed as single *words* or *phrases*, while the strength of a relation between the starting and target concepts A and C is measured in terms of the number of intermediate concepts B.[28]

Variations of the ABC Algorithm

Swanson's approach to discovery served as a basis for several subsequent efforts, which aimed to improve the original algorithm by modifying one or more of its parameters. In LBD by lexical statistics, concepts are expressed as either *words* or *complete phrases*, and the strength of a relation between the starting and target concepts is expressed in terms of four statistical measures, as opposed to the plain number of intermediate concepts used in *Arrowsmith*;[28,29] these are: (1) the number of publications within the subset of Medline that is used as the literature corpus, which refer to the concepts of interest (*document frequency*); (2) the total number of occurrences of these concepts within all publications comprising the corpus (*token frequency*), which can be higher than document frequency because a concept may be referred to more than once within each publication; (3) the *relative frequency* of the concepts in the literature corpus versus the entire Medline; and (4) the product (*token frequency*)×(*inverse document frequency*), where *inverse document frequency* is defined by the equation:

$$\text{Inverse document frequency} = \log\left(\frac{\begin{array}{c}\text{total number of}\\ \text{publications in Medline}\end{array}}{\begin{array}{c}\text{number of Medline publications}\\ \text{referring to the particular concept}\end{array}}\right) \quad (1)$$

Further to the additional measures, the lexical statistics approach, contrary to *Arrowsmith*,

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

capitalizes on complete Medline records, which include both titles and abstracts. This feature is shared by many subsequent variations of the original algorithm.

LitLinker, on the other hand, is an open discovery approach. It searches for concepts of interest in MESH and uses a statistical method based on word probability distributions to identify MESH terms that are closely related to the starting and target concepts, while it incorporates a knowledge-based approach to allow pruning of synonymous or noninteresting concepts.[30] *LitLinker* was reported to have uncovered three novel relations: that of AD to endocannabinoids, of migraine to AMPA receptors, and of schizophrenia to secretin.[30] These relations, however, have been criticized for not being novel, on the basis of prior scientific papers disclosing them explicitly.[31]

In the DAD system, the ABC model is enhanced by Natural Language Processing techniques for better identification of concepts in biomedical literature, and by a knowledge-based approach that capitalizes on the semantic information incorporated in the UMLS thesaurus,[32] to help prune spurious or noninteresting relations.[33] In 2003 it was proposed, on the basis of results obtained by the use of this software, that thalidomide might be a potential treatment for myasthenia gravis and chronic hepatitis C.[34] Although the former hypothesis has not been validated at the time of writing of this paper, thalidomide was reported to be an effective add-on treatment for chronic hepatitis C, in a report of six patients.[35] Wren et al. used a similar conceptual approach (although implemented differently) and proposed that chlorpromazine, the first-in-class typical antipsychotic, could be used in the treatment of cardiac hypertrophy.[36] This hypothesis has been validated in a rodent model of cardiac hypertrophy, but has not been corroborated in clinical trials at the time of writing, which is perhaps not surprising in view of the cardiac adverse reactions of the drug. Indeed, chlorpromazine can cause arrhythmias and should be used with caution in patients with cardiovascular comorbidities.[37]

Further variations of the original algorithm include the use of association rule mining to identify co-occurring MESH terms,[38] and use of Support Vector Machines in conjunction with dependency-parse trees to capture predicate–argument relationships among the words of a sentence.[39] The principal component analysis statistical technique has also been utilized in an unsupervised approach to mining Medline for associations between genes and phenotypic characteristics.[40] The authors demonstrated that the feature sets extracted from Medline abstracts can be used to identify novel genotype–phenotype relations.

RECENT APPROACHES TO LITERATURE-BASED DRUG REPURPOSING

Focused applications of LBD to drug repurposing have appeared in recent years. These rely on the biomedical literature alone or combined with additional data sources, and are pursued by both academic institutions and companies. In the latter case, they are both offered as services and employed to enhance internal product pipelines of the stakeholders or to extend the life cycle of a drug.

Reliance on additional data sources that complement the literature is exemplified by the work of Li et al., who created drug–protein connectivity maps from molecular interaction networks and Medline abstracts.[41] Here the authors processed the subset of Medline that refers to AD and identified all drugs cited in these papers; they subsequently used a term frequency statistical criterion to isolate the drugs that are significantly 'enriched' in the AD literature compared to the entire Medline, and they constructed a drug–protein connectivity matrix on the basis of this information. A key difference from previous approaches is that the literature on AD was not retrieved by using 'Alzheimer's Disease' and its synonyms as search terms; rather, they used a seed list of 560 proteins that are related to AD according to Online Mendelian Inheritance in Man (OMIM)[42] and Online Human Interaction Database (OPHID)[43] databases. This allows the identification of both explicit and implicit drug–protein associations to a disease context and carries the potential of uncovering a more extended set of drug candidates for the disease of interest (here AD).

In the MedGene approach, Medline records are scanned for references to genes and diseases, with the purpose of producing a comprehensive set of gene–disease relations.[44] The strength of the derived relations is calculated by statistical methods, including χ^2 analysis, Fisher's exact probabilities, relative risk of gene occurrence, and relative risk of disease occurrence. These gene–disease relations have been evaluated against a microarray expression dataset comparing breast cancer and normal breast tissue. A significant correlation was observed between MedGene results and genes showing an expression difference of 10-fold or more in the microarray experiment ($r = 41$, $p < 0.05$).[44] The literature-derived genes and their associated proteins can then serve as putative targets for drug discovery in the

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

respective disease areas, as discussed in a recent review.[45] This technique can also be used in reverse, with the aim of identifying all diseases associated with a particular protein. Previously unknown disease associations can be further evaluated as repurposing opportunities for compounds binding the particular protein.[45,46]

Chemical abstract service registry (CAS)[47] numbers and enzyme commission (EC)[48] classification information are additional descriptors of scientific papers and have been combined with MESH to arrive at novel literature-based hypotheses, derived on the basis of coword clustering.[49]

A number of companies previously engaged in providing software and services based on, among other data types, the analysis of the scientific literature have recently entered the field of drug repurposing. Companies like Ingenuity,[50] GeneGo,[51] and Ariadne Genomics[52] have all been providing pathway-based solutions to the pharmaceutical industry and have recently started aiming their tools toward drug repurposing. Ingenuity capitalizes on its proprietary database of biomedical relations, which are manually extracted from the literature. GeneGo and Ariadne Genomics, on the other hand, are integrating a variety of bioinformatics and cheminformatics resources with literature analysis to identify novel drug targets for a given drug. Clients of these companies may opt to gain direct access to the analytical tools, for the purpose of performing their own analyses, or to seek the advice of the companies' internal teams on their research questions.

GeneGo combines pathway and chemical structural analyses to produce a list of putative indications for a compound of interest. GeneGo identifies drug targets by (1) finding the known targets for the given compound, (2) identifying similar compounds through structural similarity algorithms, and (3) performing Quantitative Structure–Activity Relationship (QSAR) predictions on target affinity to identify new potential targets.

Ariadne Genomics integrates the analyses of the datasets produced by high-throughput experiments, such as microarrays, with literature analytics, in order to derive relationships among genes, proteins, cell processes, and diseases. They used this approach to repurpose fulvestrant, an estrogen receptor antagonist currently used in the treatment of hormone receptor-positive metastatic breast cancer, to glioblastoma, a brain malignancy.[53]

Biovista[54] capitalizes extensively on the use of literature analysis to infer novel relationships between drugs, new indications and potential ADRs. Central to Biovista's efforts is its Clinical Outcome Search Space (COSS) computational platform, a system incorporating text mining and data analysis and visualizations tools. COSS relies on an extensive proprietary database of context-crossing relations among biomedical entities, which utilizes custom technological solutions to achieve high-performance access to the underlying data. These technologies allow Biovista's medical subject experts to rapidly evaluate mechanistic hypotheses as these are formed during the discovery process. Biovista has repurposed dimebon, a histamine receptor antagonist currently used in allergic conditions, to primary and secondary progressive multiple sclerosis, two neurodegenerative forms of multiple sclerosis. It has also repurposed pirlindol, a monoaminoxidase-A inhibitor currently marketed as an antidepressant to the same indications. In both cases Biovista has obtained positive efficacy results in Myelin Oligodendrocyte Glycoprotein (MOG)-induced Experimental Allergic Encephalomyelitis (EAE), a mouse model of multiple sclerosis.[55,56]

RECENT APPROACHES TO LITERATURE-BASED ADVERSE EVENT PREDICTION

Despite the medical and fiscal importance of adverse event prediction, the number of publications on the issue is surprisingly small, especially when the methodology of prediction is required to include the biomedical literature.

In a recent report, the Center for Food Safety and Applied Nutrition of the U.S. FDA combined multiple sources of data in an attempt to predict cardiac ADRs in humans.[57] The authors created a database of cardiac ADRs and used it to (1) construct QSAR models that could predict cardiac ADRs of untested chemicals, (2) identify different properties of pharmaceutical molecules that correlate with rare and unexpected cardiac ADRs observed in patients, and (3) identify plausible Mechanisms of Action (MoAs) by which the drugs might have caused the ADRs, on the basis of these *in silico* data. In this approach, drugs were classified according to (1) the clinical indications for which they were prescribed, (2) their primary target, (3) their MoA, and (4) their structural similarity to other drugs, known to bind to specific receptors.[57] Drug-related ADRs were derived from FDA's Spontaneous Reporting System (SRS) and Adverse Event Reporting System (AERS) post-market surveillance databases and a supplement of adverse event data from published medical literature,[58,59] while drug MoAs and target affinities were compiled for 2124 FDA-approved

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

drugs through QSAR modeling.[60] It was found that cardiac ADRs correlate with a small number of MoAs, namely those affecting cardiovascular functions (such as α-adrenoceptor, β-adrenoceptor, and calcium channel blocker) and cardioneurological functions (5-hydroxytryptomine receptor, dopamine receptor, and acetylcholinesterase).[57] The authors suggested that screening of new chemical entities for the presence of these MoAs could predict a major portion of cardiac ADRs that might occur in patients, and that this technology might be used proactively for the early detection of ADRs in clinical trials and for the investigation of rare, unexpected, and idiosyncratic ADRs that are identified by pharmacovigilance and post-market surveillance. A similar approach had been used by the same authors to predict hepatobiliary and urinary tract ADRs.[61,62] More recently, association rule mining, a well-established data mining method, has been used to detect in AERS associations between multiple drugs and potential ADRs.[63]

Alomar et al. used biomedical literature as the primary source of data from which the authors isolated reports on ADRs. They identified key risk factors in these reports, including age, gender, race, maternity, obesity, comorbidities, concomitant medications, and allergies, and they expressed these factors as categorical variables. These were used in turn to create a mathematical model for predicting the expected frequency of a given ADR in a patient taking specific drugs and having a given number of risk factors.[64] Although this is an interesting exploratory approach, no rigorous validation has been provided or documented.

The biomedical literature together with databases of protein-chemical [PDSP Ki Database,[65] Protein Data Bank (PDB),[66] KEGG,[67] Reactome,[68] NCI-Nature Pathway Interaction Database,[69] DrugBank,[70] and MATADOR[71]] and drug–ligand interactions (GLIDA,[72] PharmGKB,[73,74] Comparative Toxicogenomics Database (CTD),[75] and BindingDB[76]) are incorporated in STICH, a comprehensive resource of drug–drug interactions.[77] Although the use of this and other similar resources in drug repurposing scenarios has been discussed extensively (see, for example, Refs 78 and 79), their potential utility in ADR prediction has been overlooked. This is a surprising observation, given the inherent similarity of these two research areas. SIDER, on the other hand, is a manually curated database of ADRs for all FDA-approved drugs, which have been extracted from the corresponding drug labels. Although not based on the scientific literature *per se*, SIDER has been used to capture the phenotypic effects of drugs[80] and, from this perspective, is a useful resource for ADR prediction applications.

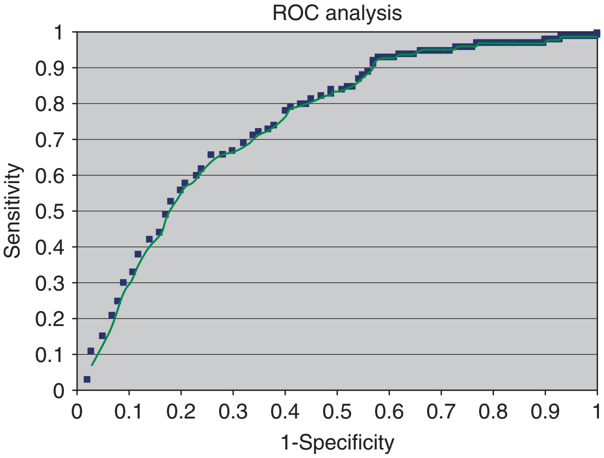

FIGURE 3 | Receiver operating characteristic curve (ROC) analysis of the overall performance of Clinical Outcome Search Space (COSS) in classifying drugs, according to their propensity to produce a predefined set of adverse drug reactions (ADRs). The area under the curve (AUC) value is 0.75 (95% CI 0.70–0.79).

Recently text mining of Medline abstracts was used to annotate genes and to identify their copublications with pathology terms and biological processes and pathways. This information was used to produce keyword-fingerprints of selected compounds and to predict their toxicologic characteristics. In the case of two genotoxic carcinogens, two nongenotoxic carcinogens, and two peroxisome proliferators activators, the fingerprints that were derived from the literature correlated well with the histopathological events that were induced by the individual compounds in rat livers and with the results of microarray analysis of these tissues.[81]

Biovista utilizes its COSS platform for the prediction of ADRs for compounds of interest. This aspect of the company's services is also used to strengthen its own repurposing candidates with in-depth analyses of their benefit/risk profiles, before advancing them to preclinical development. In a classification task, where drugs were classified according to their propensity for producing a predefined set of ADRs, COSS achieved an overall receiver operating characteristic curve (ROC) area under the curve (AUC) value of 0.75 [95% confidence interval (CI) 0.70–0.79, $p < 0.0001$]. At the optimal configuration, sensitivity was 0.78 (95% CI 0.72–0.83) and specificity 0.60 (95% CI 0.58–0.62; Figure 3, data not shown).

CONCLUSIONS

Building on Swanson's original work, early efforts in literature-based discovery have explored the potential

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of the methodology and the effect of several proposed improvements on the final output. Although they were not specifically focused on drug repurposing, some have identified new indications for existing drugs that proved correct in subsequent clinical trials. Unquestionably, these efforts have laid out the LBD framework and have paved the way to contemporary drug repurposing and ADR prediction.

In recent years, when the allocation of pharmaceutical research costs per approval drug for a unique therapeutic indication has grown significantly, LBD efforts have focused more intensely on drug repurposing. From the technology point of view, emphasis has been made on the integration of the biomedical literature with additional resources, such as drug–ligand and chemical databases. Although certain groups develop technologies specifically aimed at drug repurposing, others arrive at novel drug–target and disease–target associations. Strictly speaking, these are two different applications of LBD; however, drug–target–disease associations are immediately amenable to the identification of repurposing opportunities.

Literature-based ADR prediction is maturing slower than drug repurposing. Currently, pharmacovigilance is the main approach used in the industry, where detection, assessment, understanding, and prevention of short- and long-term ADRs are based on the detection of disproportionate reporting signals.[82] It requires a rigorous post-marketing surveillance (PMS) process to assess the safety of a company's product in the worldwide marketplace using primarily spontaneous reports based on good pharmacovigilance practices (GPVP).[83] In the best case, a careful analysis of spontaneous reports can only alert, signal, and confirm an association. In certain circumstances, one high quality and carefully documented report with a successful dechallenge and a positive drug rechallenge may have sufficient internal validity to declare a causal association; other than this scenario, spontaneous case reports cannot demonstrate a definitive causal relationship. Future strategic investigations beyond standard practices using pharmacoepidemiologic methods are needed for hypothesis testing and risk quantification in observational data to determine risk management strategies. Pharmacovigilance is the risk evaluation component spanning the full life cycle of a drug, from early clinical investigational trials through the duration of its global marketing, but is particularly focused on PMS and risk mitigation, if warranted. ADRs are, of course, thoroughly monitored during clinical trials, while during preclinical development the potential ADRs of a compound are predicted on the basis of *in vitro* tests and animal trials in various species (mice, rats, dogs, monkeys, etc). On the one hand, recording of ADRs that occurred in the clinic constitutes a reactive approach, i.e., it is a response to an event that has already happened. Nothing can be done to change the molecular properties of the drug at that time. On the other hand, animal results do not always translate to humans accurately, due to several factors, including species differences and genetic variability.[84] Contemporary *in silico* approaches to ADR prediction concentrate on docking studies and Structure–Activity Relationship (SAR)/QSAR models for the identification of structural features with ADR propensity.[85,86] The use of LBD for ADR prediction has not received much attention at present, despite its considerable methodological similarity to drug repurposing.

The issue of computational performance should not be overlooked in any application of LBD. In practice any discovery, including the repurposing of drugs and the prediction of ADRs, entails the scrutiny of computer output by human experts. Although researchers have tried to minimize human interaction, it still comprises an important step of all discoveries. Furthermore, the answer to a given research question might involve several LBD runs, executed sequentially; the output of each run of the algorithm might provide part of the answer, which can then be used as input to a next run, until a final conclusion is reached. Any software and hardware infrastructure used in this context should, thus, perform adequately, to allow the quick exploration of research questions, as they are created during real-life projects. This is not a trivial problem, as open discovery processes usually rely on the repeated processing of the entire biomedical literature. Modern high-performance computing architectures, such as GRID and Cloud Computing,[87] are part of the solution; efficient LBD software is the remaining part.

The Holy Grail of all predictive technologies that are used in biomedicine, either *in silico* or not, is how accurately their results can be translated to clinical outcomes in humans. This is especially true in the case of LBD, the sensitivity and specificity of which in drug repurposing and ADR prediction remains to be determined. This is understandably a tedious process, as the clinical development of a drug—even that of a repurposed drug with proven safety—takes many years. Similarly, predicted ADRs may not be observed for a long time; still that does not mean that they will not be observed in the future. Until the time comes that we will have gathered adequate prospective data to evaluate the statistical performance of these technologies, our only alternative is to rely on surrogate markers of success.

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Clinically or Experimentally Validated Efforts Toward the Use of Literature-Based Discovery (LBD) in Drug Repurposing and Adverse Event Prediction

Author/Company	System Name	Architecture/Corpus	Initial use	Clinical/ Experimental Validation	Comments	References
Swanson	Arrowsmith	Relies on Medline titles and MESH	Drug repurposing	Use of fish oil in Raynaud's syndrome was validated in a clinical trial		18, 23
Wren		Medline title and abstract	Drug repurposing	Use of chlorpromazine in cardiac hypertrophy was validated in a rodent model	Fuzzy logic is used to rank concept relations	37
Weeber	DAD	Medline title and abstract, UMLS	Drug repurposing	Use of thalidomide for the treatment of chronic hepatitis C was validated in a case series	Uses natural language processing for the better identification of concepts in the literature	34–38
Hu	MedGene	Medline records	Extraction of gene-disease relations	Evaluated against the results of a breast cancer microarray experiment	Relies on statistical comparisons (χ^2 and other tests) between literature sets	45
Biovista	Clinical Outcome Search Space (COSS)	● Biomedical literature ● Domain expert interpretation of the results ● Additional data: ○ Adverse event reporting system ○ Patent literature ○ Structural analysis	● Drug repurposing ● Adverse event prediction ● Identification of at-risk populations ● Target profiling	● Validation of two repurposed drugs in animal models of multiple sclerosis ● ROC analysis of adverse event prediction accuracy		55–56
Frijters		Medline abstracts	Assessment of compound toxicity	Good correlation with the histopathological events that were induced by the test compounds in rat livers and with the results of microarray analysis of these tissues		81

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In vitro and animal experiments proving the efficacy of repurposing candidates in the new target indications is one example of such a marker. Use of past literature data to predict ADRs and corroboration of the results in the more recent literature is another example. Current efforts that have employed such surrogate markers for the validation of their computational results are outlined in Table 1. As the field of literature-based drug repurposing and ADR prediction evolves, it is critical to shift our focus from 'discovery' to 'validated discovery'; and to produce the statistics required to back our hypotheses.

REFERENCES

1. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff (Millwood)* 2006, 25:420–428.
2. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ* 2003, 22:151–185.
3. Boguski MS, Mandl KD, Sukhatme VP. Drug discovery. Repurposing with a difference. *Science* 2009, 324:1394–1395.
4. Grabowski H. Are the economics of pharmaceutical research and development changing? Productivity, patents and political pressures. *Pharmacoeconomics* 2004, 22(2 suppl 2):15–24.
5. Wadman M. Experts call for active surveillance of drug safety. *Nature* 2007, 446:358–359.
6. Moore TJ, Cohen MR, Furberg CD. Serious adverse drug events reported to the Food and Drug Administration, 1998–2005. *Arch Intern Med* 2007, 167:1752–1759.
7. Gilbert J, Henske P, Singh A. Rebuilding Big Pharma's Business Model. *In Vivo*, the Business & Medicine Report, Windhover Information, vol. 21; 2003.
8. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004, 3:673–683.
9. Carley DW. Drug repurposing: identify, develop and commercialize new uses for existing or abandoned drugs. Part II. *Drugs* 2005, 8:310–313.
10. PricewaterhouseCoopers. Pharma 2005 silicon rally: the race to e-R&D. *Paraxel's Pharmaceutical R&D Statistical Sourcebook* 2002/2003.
11. Netterwald J. Recycling existing drugs. *Drug Disc Dev* 2008, 16–22.
12. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol* 2010, 6(9).
13. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 2006, 7: 119–129.
14. Persidis A, Deftereos S, Persidis A. Systems literature analysis. *Pharmacogenomics* 2004, 5943–947.
15. Deftereos S, Andronis C, Friedla EJ, Persidis A, Persidis A. Drug repurposing and adverse event prediction using high-throughput literature analysis: case studies. *8th International Conference on Pathways, Networks, and Systems Medicine*, Rhodes July 9–14, 2010.
16. Kitano H. Systems biology: a brief overview. *Science* 2002, 295:1662.
17. Swanson DR. Complementary structures in disjoint science literatures. In: Bookstein A, Chiaramella Y, Salton G, Raghavan VV, eds. *Proceedings of the 14th Annual International ACM/SIGIR Conference*. New York: ACM Press; 1991, 280–289.
18. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986, 30:7–18.
19. Swanson DR. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med* 1988, 31:526–557.
20. Vosgerau H. Migraine therapy with magnesium glutamate. *Ther Ggw* 1973, 112:640 passim.
21. Altura BM. Calcium antagonist properties of magnesium: implications for antimigraine actions. *Magnesium* 1985, 4:169–175.
22. Altura BM, Altura BT. New perspectives on the role of magnesium in the pathophysiology of the cardiovascular system. I. Clinical aspects. *Magnesium* 1985, 4:226–244.
23. DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am J Med* 1989, 86:158–164.
24. Schiapparelli P, Allais G, Castagnoli Gabellari I, Rolando S, Terzi MG, Benedetto C. Non-pharmacological approach to migraine prophylaxis: part II. *Neurol Sci* 2010, 31(suppl 1):S137–S139.
25. Smalheizer NR, Swanson DR. Linking estrogen to Alzheimer's disease: an informatics approach. *Neurology* 1996, 47:809–810.
26. Smalheizer NR, Swanson DR. Indomethacin and Alzheimer's disease. *Neurology* 1996, 46:583.
27. Smalheiser NR, Torvik VI. The place of literature-based discovery in contemporary scientific practice. In: Weeber M, Bruza P, eds. *Literature-based Discovery*. Springer; Berlin 2008, 13–22.

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

28. Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. *J Am Soc Inf Sci* 1999, 50:574–587.

29. Gordon MD, Lindsay RK. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *J Am Soc Inf Sci* 1996, 47:116–128.

30. Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. *J Biomed Inform* 2006, 39:600–611.

31. Kostoff R. Validating discovery in literature-based discovery. *J Biomed Inform* 2007, 40:448–450.

32. Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. *J Am Med Rec Assoc* 1990, 61:40–42.

33. Weeber M, Klein H, de Jong-van den Berg LTW, Vos R. Using concepts in literature based discovery: simulating Swanson's Raynaud—fish oil and migraine— magnesium examples. *J Am Soc Inf Sci Technol* 2001, 52:548–57.

34. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LTW, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J Am Med Inform Assoc* 2003, 10:252–259.

35. Caseiro MM. Treatment of chronic hepatitis C in non-responsive patients with pegylated interferon associated with ribavirin and thalidomide: report of six cases of total remission. *Rev Inst Med Trop Sao Paulo* 2006, 48:109–112.

36. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* 2004, 20:389–398.

37. Thorazine (Chlorpromazine) drug information: uses, side effects, drug interactions and warnings at RxList. Available at: http://www.rxlist.com/thorazine-drug.htm. (Accessed October 7, 2010).

38. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Improving literature based discovery support by genetic knowledge integration. *Stud Health Technol Inf* 2003, 95:68–73.

39. Ozgür A, Xiang Z, Radev DR, He Y. Literature-based discovery of IFN-γ and vaccine-mediated gene interaction networks. *J Biomed Biotechnol* 2010, 2010:426–479.

40. Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* 2005, 3:e134.

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

42. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005, 33:D514–D517.

43. Brown KR, Jurisica I. Online predicted human interaction database. *Bioinformatics* 2005, 21:2076–2082.

44. Hu Y, Hines LM, Weng H, Zuo D, Rivera M, Richardson A, LaBaer J. Analysis of genomic and proteomic data using advanced literature mining. *J Proteome Res* 2003, 2:405–412.

45. Loging W, Harland L, Williams-Jones B. High-throughput electronic biology: mining information for drug discovery. *Nat Rev Drug Discov* 2007, 6:220–230.

46. Potts SJ, Edwards DJ, Hoffman R. Challenges of target/compound data integration from disease to chemistry: a case study of dihydrofolate reductase inhibitors. *Curr Drug Discov Technol* 2005, 2:75–87.

47. Chemical Abstracts Service. Available at: www.cas.org/. (Accessed October 7, 2010).

48. Webb EC. *Enzyme nomenclature 1992 recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. San Diego, CA: Published for the International Union of Biochemistry and Molecular Biology by Academic Press; 1992. ISBN 0-12-227164-5. Available at: http://www.chem.qmul.ac.uk/iubmb/enzyme/. (Accessed October 7, 2010).

49. Stegmann J, Grohmann G. Hypothesis generation guided by co-word clustering. *Scientometrics* 2003, 56:111–135.

50. Ingenuity. Available at: http://www.ingenuity.com/. (Accessed October 7, 2010).

51. GeneGo. Available at: http://www.genego.com/. (Accessed October 7, 2010).

52. Ariadne Genomics. Available at: http://www.ariadnegenomics.com/. (Accessed October 7, 2010).

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

54. Biovista. Available at: http://www.biovista.com/. (Accessed October 7, 2010).

55. Deftereos SN, Andronis C, Virvillis V, Konstanti O, Persidis A. Dimebon ameliorates disease severity in the MOG-induced experimental allergic encephalomyelitis animal model of progressive multiple sclerosis. *American Neurological Association 135th Annual Meeting*, San Fransisco, September 12–15, 2010.

56. Deftereos SN, Andronis C, Sharma A, Aris Persidis A. Systematic drug repurposing for CNS indications: account of two successful case studies. *American Neurological Association 135th Annual Meeting*, San Fransisco, September 12–15, 2010.

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

57. Matthews EJ, Frid AA. Prediction of drug-related cardiac adverse effects in humans—A: creation of a database of effects and identification of factors affecting their occurrence. *Regul Toxicol Pharmacol* 2010, 56:247–275.

58. Adverse Event Reporting System (AERS). Available at: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm. (Accessed October 7, 2010).

59. Thomson-Reuters-MicroMedex. Available at: http://www.micromedex.com/. (Accessed October 7, 2010).

60. Matthews EJ, Ursem CJ, Kruhlak NL, Benz RD, Sabaté DA, Yang C, Klopman G, Contrera JF. Identification of structure–activity relationships for adverse effects of pharmaceuticals in humans: B. Use of QSAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. *Regul Toxicol Pharmacol* 2009, 54:23–42.

61. Matthews EJ, Kruhlak NL, Benz RD, Sabaté DA, Marchant CA, Contrera JF. Identification of structure activity relationships for adverse effects of pharmaceuticals in humans: C. Use of QSAR and an expert system for the estimation of the mechanism of action of drug-induced hepatobiliary and urinary tract toxicities. *Regul Toxicol Pharmacol* 2009, 54:43–65.

62. Ursem CJ, Matthews EJ, Kruhlak NL, Contrera JF, Benz RD. Identification of structure–activity relationships for adverse effects of pharmaceuticals in humans A: use of FDA post market reports to create a database of hepatobiliary and urinary tract toxicities. *Regul Toxicol Pharmacol* 2009, 54:1–22.

63. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* 2010, 11(suppl 9):S7.

64. Alomar MJ, Hourani AA, Sulaiman SA. Proposal for the development of adverse drug reaction prediction model. *Am J Pharmacol Toxicol* 2008, 3:193–200.

65. Roth B, Lopez E, Patel S, Kroeze W. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* 2000, 6:262.

66. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000, 28:235–242.

67. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006, 34:D354–D357.

68. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005, 33: D428–D432.

69. Pathway Interaction Database. Available at: http://pid.nci.nih.gov. (Accessed October 7, 2010).

70. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res* 2006, 34:D668–D672.

71. Gunther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* 2008, 36:D919–D922.

72. Okuno Y, Yang J, Taneishi K, Yabuuchi H, Tsujimoto G. GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res* 2006, 34:D673–D677.

73. Gong L, Owen RP, Gor W, Altman RB, Klein TE. PharmGKB: an integrated resource of pharmacogenomic data and knowledge. *Curr Protoc Bioinformatics* 2008, Chapter 14(Unit 14):17.

74. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res* 2002, 30:163–165.

75. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, et al. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res* 2009, 37:D786–D792.

76. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding afinities. *Nucleic Acids Res* 2007, 35:D198–D201.

77. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, von Mering C, Jensen LJ, Beyer A, Bork P. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res* 2010, 38:D552–D556.

78. Lee S, Park K, Dongsup K. Building a drug–target network and its applications. *Exp Opin Drug Discov* 2009, 4:1177–1189.

79. Chautard E, Thierry-Mieg N, Ricard-Blum S. Interaction networks: from protein functions to drug discovery. A review. *Pathol Biol (Paris)* 2009, 57:324–333.

80. 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.

81. Frijters R, Verhoeven S, Alkema W, van Schaik R, Polman J. Literature-based compound profiling: application to toxicogenomics. *Pharmacogenomics* 2007, 8:1521–1534.

82. Norén GN, Edwards IR. Modern methods of pharmacovigilance: detecting adverse effects of drugs. *Clin Med* 2009, 9:486–489.

83. Härmark L, van Grootheest AC. Pharmacovigilance: methods, recent developments and future perspectives. *Eur J Clin Pharmacol* 2008, 64:743–752.

© 2011 John Wiley & Sons, Inc.

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

84. Watterson C, Lanevschi A, Horner J, Louden C. A comparative analysis of acute-phase proteins as inflammatory biomarkers in preclinical toxicology studies: implications for preclinical to clinical translation. *Toxicol Pathol* 2009, 37:28–33.

85. Ridings JE, Barratt MD, Cary R, Earnshaw CG, Eggington CE, Ellis MK, Judson PN, Langowski JJ, Marchant CA, Payne MP. Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. *Toxicology* 1996, 106:267–279.

86. Cash GG. Prediction of the genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices. *Mutat Res* 2001, 491:31–37.

87. Reese G. *Cloud Application Architectures: Building Applications and Infrastructure in the Cloud: Transactional Systems for EC2 and Beyond*. Cambridge, MA: O'Reilly Media; 2009.

88. Srinivasan P. Generating hypotheses from MEDLINE. *J Am Soc Inf Sci Technol* 2004, 55:396–413.

1939005x, 2011, 3, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.147 by Tu Muenchen Bibliothek, Wiley Online Library on [01/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License