Research and Applications

Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications

Justin Mower,[1] Devika Subramanian,[2] and Trevor Cohen[3]

[1]Baylor College of Medicine, Quantitative and Computational Biosciences, Houston, Texas, USA, [2]Department of Computer Science, Rice University, Houston, Texas, USA and [3]School of Biomedical Informatics, University of Texas Health Science Center Houston, Texas, USA

Corresponding Author: Justin M. Mower, BS, Quantitative and Computational Biosciences, Baylor College of Medicine, 1 Baylor Plaza MS BCM272, Houston, Texas, 77030; justin.mower@bcm.edu

ABSTRACT

Objective: The aim of this work is to leverage relational information extracted from biomedical literature using a novel synthesis of unsupervised pretraining, representational composition, and supervised machine learning for drug safety monitoring.

Methods: Using ≈80 million concept-relationship-concept triples extracted from the literature using the SemRep Natural Language Processing system, distributed vector representations (embeddings) were generated for concepts as functions of their relationships utilizing two unsupervised representational approaches. Embeddings for drugs and side effects of interest from two widely used reference standards were then composed to generate embeddings of drug/side-effect pairs, which were used as input for supervised machine learning. This methodology was developed and evaluated using cross-validation strategies and compared to contemporary approaches. To qualitatively assess generalization, models trained on the Observational Medical Outcomes Partnership (OMOP) drug/side-effect reference set were evaluated against a list of ≈1100 drugs from an online database.

Results: The employed method improved performance over previous approaches. Cross-validation results advance the state of the art (AUC 0.96; F1 0.90 and AUC 0.95; F1 0.84 across the two sets), outperforming methods utilizing literature and/or spontaneous reporting system data. Examination of predictions for unseen drug/side-effect pairs indicates the ability of these methods to generalize, with over tenfold label support enrichment in the top 100 predictions versus the bottom 100 predictions.

Discussion and Conclusion: Our methods can assist the pharmacovigilance process using information from the biomedical literature. Unsupervised pretraining generates a rich relationship-based representational foundation for machine learning techniques to classify drugs in the context of a putative side effect, given known examples.

Key words: machine learning, representation learning, pharmacovigilance, unsupervised pretraining, literature based discovery

OBJECTIVE

Contemporary approaches for identifying potential on-market drug side effects depend on aggregation of many data sources and manual signal review.[1,2] One source of information to assist this process is the biomedical literature.[3] Due to scale and complexity, this data source necessitates robust and scalable methods.[3–5] The aim of this work is to leverage relational information extracted from the biomedical literature for drug safety monitoring, using a novel synthesis of unsupervised pretraining, representational composition, and supervised machine learning.

BACKGROUND AND SIGNIFICANCE

Drug safety monitoring

Pharmaceuticals are a primary method of therapeutic intervention, with nearly half of the US population utilizing a prescription drug in a given month, and office, outpatient, and emergency department visits including drug therapy in ≈75% or more of cases.[6–9] Unfortunately, pharmaceutical intervention may precipitate pharmaceutical side effects, and adverse drug events (ADEs) are both common and costly. The annual financial cost of drug-related morbidity and mortality in the United States was estimated at 528.4 billion in 2016, equivalent to 16% of total US healthcare expenditures that year.[10] ADEs are unfortunately frequent in both hospitals[11] and outpatient settings.[12] Often, adverse effects of drugs are identified after their approval and release to market. Numerous products have been removed from the market citing safety concerns,[13] underscored by high profile cases such as Vioxx (rofecoxib) and Bextra (valdecoxib).[13–15] Furthermore, a recent study found that nearly one in three drug products approved between 2001 and 2010 had post-market safety events, such as a label change or withdrawal, in the years following release.[16] The prevalence of these post-market safety events is due, in part, to limitations in duration and patient cross-section, inherent in the clinical trial process.[17,18]

To identify previously undetected side effects, drugs are monitored for safety after market release, a process known as pharmacovigilance (PV).[18] PV has been primarily mediated by spontaneous reporting systems (SRS), such as the Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) in the United States.[19] FAERS aggregates large numbers of reports of adverse events from clinicians, researchers, and patients, with over a million reports received in 2014 alone.[20] These data have widely acknowledged limitations, however, such as reporting bias and incompleteness of data.[19,21–24] Consequently, detected signals mined from FAERS require additional review for assessment of plausible causality.[2,25] To assist in this, researchers have sought to improve signal detection through algorithm development and integration of multiple data sources.[13,26–29] One possible data source with information relevant to causality assessment is the biomedical literature. Indeed, this literature is already consulted by reviewers in the PV process.[30,31] However, rapid increases in the biomedical literature make manual review increasingly intractable.[32,33] Scalable methods to analyze this large text repository to assess potential causal links are needed.[2,32,34]

Literature-based discovery

The most common approaches to leveraging the literature for PV are based on concept co-occurrence.[3,35,36] The general idea is that if concepts co-occur with disproportionate frequency, a meaningful statistical association exists between them. In PV, this can be leveraged to mine literature for enriched associations between drugs and potential ADEs.[3] At times, constraints are placed upon these co-occurrence relationships, such as recognition of a causality assertion using natural language processing (NLP),[37] or identification of a medical subject heading (MeSH) term indicating an adverse event.[38] Terminological mapping and expansion can be used to enhance signal detection within these constraints.[37,38] Regardless of constraints or enhancements, *explicit* co-occurrence between a drug and a side effect within a unit of text is a prerequisite for signal detection.

Concepts in the literature may also be related to one another *implicitly*—in some cases exclusively so. Direct co-occurrence models miss these hidden connections. Swanson's seminal work on literature-based discovery (LBD) demonstrated that these indirect connections between concepts can reveal relationships that are both biomedically plausible and therapeutically useful.[4,39,40] On the basis of their co-occurrence with shared *bridging concepts*, Swanson identified fish oil as a potential therapeutic for Raynaud's Disease, a finding later supported by a clinical trial.[39,41] This form of transitive inference, originally envisioned to discover treatments, can also be applied to identify side effects.[42] Traditionally, LBD is accomplished by identifying chains of directly co-occurring terms,[35] a computationally expensive task on account of the combinatorial explosion of possible bridging terms.[4] Perhaps more importantly, such methods do not examine structured relational information—which is to say that the *nature* of the relationships between concepts is not considered. How concepts relate to one another is of particular interest when assessing the biological plausibility of putative associations.

Auspiciously, large amounts of explicitly structured relational information have been extracted from the biomedical literature using NLP. For example, the semantic knowledge representation (SemRep) system extracts concept-relationship-concept assertions (eg drugA: *treats*: diseaseB), known as *semantic predications*.[43] Operating over MEDLINE citations, SemRep extracts on the order of tens of millions of semantic predications. Drawing inference from this relational information is still challenging, as step-wise exploration of the entire logical connection space is also computationally intractable.[36] Consequently, methods that limit the search space using relational constraints,[42,44] and/or some form of matrix factorization,[34,45–48] have been developed to utilize this information at scale.

Discovery patterns are one way of limiting the search space examined when considering explicit relational information.[44,47,49] This approach operates on the premise that some relational pathways will be more enriched for a particular implicit relation than others. For example, when looking for an implicit therapeutic relationship, enrichment might be expected along pathways in which a *stimulated* process is *inhibited* in a disease pathway (or vice versa).[44] Although these can be determined *a priori*, we have also developed methods to infer such discovery patterns from positive examples of a relationship of interest.[44,47] Discovery patterns have successfully been used to examine the role of insulin in Huntington's Disease[44] and to identify or explain other therapeutic relationships,[47,50–52] as well as ADEs.[53] In these methods, restrictions are placed on bridging concepts in terms of semantic type, semantic relationship, or a combination thereof. Consequently, these methods do not consider all possible relational connections between concepts—they are restricted in their considerations by design.

Another approach to PV has been the use of supervised machine learning models trained on manually engineered features alongside curated reference standards of positive and negative examples of drug/ADE pairs.[37,54,55] Feature vectors may incorporate information from the literature in the form of co-occurrence or disproportionality measures, with signal enhancement via mapping and expansion of terms.[37,38] Alternatively or additionally, information from a variety of ontological and structured data sources may be utilized.[54,56,57] For example, a study examining the use of a support vector machine for ADE classification utilized 4276 total phenotypic, biological, and chemical features extracted from four online databases.[54] However, manual feature engineering is a laborious process that constrains the extension of the methods to other data sets.[58,59] In other domains, methods have been proposed and deployed that obviate the need for manual feature engineering by learning representations of data in an unsupervised manner.[59–61] This unsupervised representational pretraining has resulted in better

performance and generalizability in numerous tasks, such as image and speech recognition.[59–62]

Representation learning

In our previous work, we have used representation learning for PV.[53,63,64] Representations of drug/ADE pairs were derived from SemRep output using a method termed predication-based semantic indexing (PSI),[47,65,66] which uses reversible vector transformations to encode the nature of the relationships between concept pairs. Initial results were promising, illustrating several advantages of this encoding scheme, including: (1) compressed representation of large amounts of relational information; (2) mediation of analogical inference;[67] and (3) facilitation of downstream machine learning.[63,64]

This paper moves beyond our previous work by supplanting PSI with a recently developed neural-probabilistic representational approach for semantic predications—called embedding of semantic predications (ESP)[64]—with inclusion of additional reference standards; comparison with recently published results; visualization to interrogate the underlying representations; and an evaluation of generalization to previously unseen drugs. Our hypotheses were that ESP would offer advantages over PSI as a representational basis for machine learning; that considering implicit relationships would improve the performance of literature-based models; and that trained models could be used to identify side effects of previously unseen drugs.

METHODS

Knowledge source

Predications were downloaded from SemMedDB, version 25.1.,[68] containing 82 239 652 predications extracted by SemRep from 25 027 441 MEDLINE citations available before 2016.

Unsupervised pretraining

Concept embeddings were generated utilizing ESP, implemented in the open source semantic vectors package.[64,69] In brief, ESP is a representation learning technique that generates semantic concept embeddings from semantic predications, with advantages over PSI in some predictive modeling experiments.[64] In both PSI and ESP, high-dimensional (on the order of thousands of dimensions) binary vectors are generated consistent with the binary spatter code (BSC), one of a family of representational approaches developed to mediate symbolic operations (eg variable-value binding) on connectionist representations.[70–73] As deployed in ESP and PSI, the pairwise exclusive-OR (XOR) operator, represented by $\otimes$, is applied to bind randomly initialized context embeddings (denoted C, and representing both predicates and their arguments) together, providing a basis for the generation of semantic concept embeddings (denoted S) using predications in SemMedDB. An example is shown in Figure 1 to give intuition for this training process. In PSI, bound products, each representing a predicate-argument pair, are superposed to generate concept embeddings. In ESP, this superposition occurs during the course of training a neural network to predict the object of a predication, given the subject and predicate. The mathematical differences between how this process is accomplished in ESP and PSI are briefly covered in the Supplementary Appendix, but for a more detailed account of these approaches, we refer the interested reader to Cohen and Widdows[64] and Widdows and Cohen[66] respectively. In this research, we generated two sets of concept embeddings: ESP vectors using the parameters detailed in Cohen and Widdows[64] and PSI vectors using the same parameters as in Mower et al.,[63] both at

32 000 dimensions and utilizing SemMedDB version 25.1, consistent with previous work.

Generation of composite feature vectors

After concept embeddings were trained, representations for drug/ADE pairs were composed by binding ($\otimes$) concept embeddings for the drug and ADE concerned. The resulting drug/ADE pair vectors will be similar when composed from similar vector representations. For example, the vector (myocardial infarction)$\otimes$(celecoxib) would be similar to (myocardial infarction)$\otimes$(rofecoxib), if both drugs occur in the predication (*coxib):: INHIBITS:: cox-2. This combination of trained semantic vectors also reveals ways in which two component concepts are related.[53,63] For example, if (ibuprofen)+=(TREATS)$\otimes$(PAIN) and (arthritis)+=(CAUSES)$\otimes$(PAIN), the composition (ibuprofen)$\otimes$(arthritis) will be similar to (TREATS)$\otimes$(CAUSES),[1] indicating that ibuprofen treats something caused by arthritis. Figure 1 shows this composition in a simple case. In practice and at PubMed scale, these compositions contain many such relational "pathways," resulting in an abstract relational embedding.

For this analysis, vector representations were composed for each drug/ADE pair in the Observational Medical Outcomes Partnership (OMOP) and Exploring and Understanding Adverse Drug Reactions (EU-ADR) manually curated reference sets.[1,74] The OMOP set contains 165 ground-truth positive and 234 ground-truth negative examples across four ADEs: myocardial infarction (MI), gastrointestinal bleeding (GIB), liver injury (LI), and kidney injury (KI). Examples containing two drugs (darunavir and sitagliptin) without embeddings in the vector spaces used in this analysis were removed (n = 5), leaving 394 examples (164 positive and 230 negative cases). The EU-ADR reference set contains 94 total examples across 10 ADEs (the four OMOP ADEs and six others). The only unresolved example removed was the positive example pair nimesulide-LI. Except for cardiac valve fibrosis, each ADE is comprised of both positive and negative examples. All ADE terms were either identical to the OMOP set, or extracted from the Supplementary Appendix of Coloma et al.[74] A single term was used per ADE—no terminological expansion was performed.
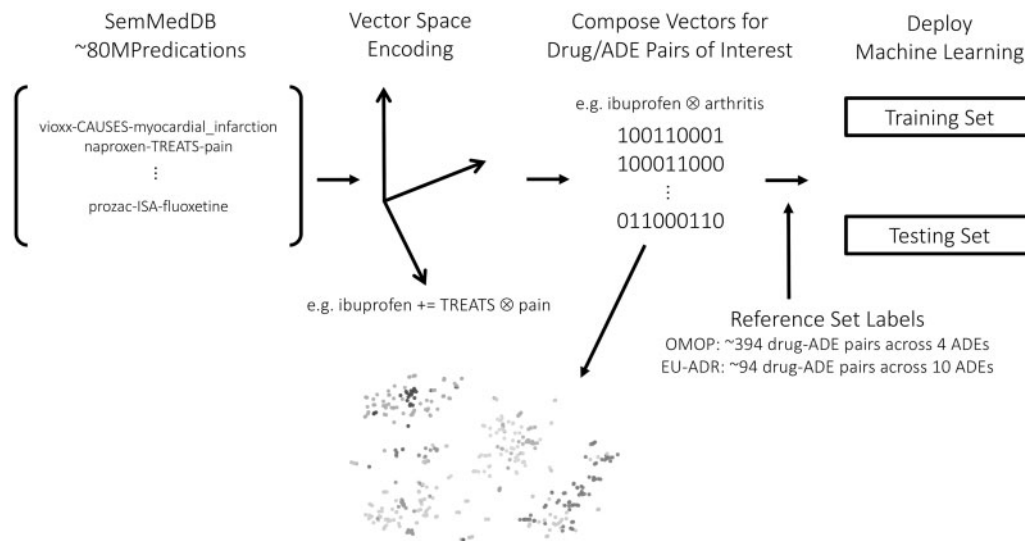
Training and cross-validation

For supervised machine learning, the composite feature vectors were labeled according to their ground-truth assertion in the OMOP or EU-ADR reference set. Experiments were performed using sci-kit learn version 0.19.0[75] and the Anaconda distribution of Python version 3.6.1.[76] We trained k-nearest neighbors (kNN) and logistic regression (LR) models in leave one out (LOO) and stratified 5-fold (S5F) cross-validation (CV) configurations. kNN was chosen, as representations should be amenable to nearest neighbor approaches (since the classification mechanism is distance based). LR was chosen as a parametric linear model that scales comfortably to large data sets. LOO was chosen to generate results comparable to other research on these standards, and S5F was chosen as a more challenging CV configuration for comparison to LOO and previous work. kNN was deployed with 1, 2, 5, and 10 nearest neighbors. For LR, L1 regularization was utilized with default parameters. To assess performance, F1 scores and receiver operating characteristic (ROC)

---

1 As XOR is its own inverse, the vector representation of "PAIN" cancels out from the bound product (PAIN)$\otimes$(PAIN)$\otimes$(TREATS)$\otimes$(CAUSES), leaving (TREATS)$\otimes$(CAUSES).

Figure 1. Example schematic of binding, bundling, and composition of representational vectors. In the top pane, random instantiation of context embeddings is shown. In the middle pane, binding (pairwise exclusive OR) and bundling (majority rule with ties split at random) of predicates and concepts relating to ibuprofen is depicted, resulting in a semantic vector for ibuprofen. In the bottom pane, a composite representation of the concept pair ibuprofen/arthritis is created using the same binding operator (as it is its own inverse) with the semantic vector for arthritis. The result is a vector approximating the representation of the relational pathways that link these concepts together, which in turn serves as the input vector for downstream machine learning applications. Gray boxes indicate a tie split at random (with a 0.5 probability of 1) when bundling. In this example, collisions between concepts occur in lower dimensions (where two vector embeddings have the same representation for different concepts). In practice and at high dimensions, random splitting of ties and collisions are exceedingly unlikely to occur, and concepts (and their relational pathways) are distinct.

Figure 2. Cross-validation (CV) configurations. Training sets are illustrated in white with black text. Held-out test sets are shown in black with white text. In within-set configurations, one of five (S5F) folds is illustrated.

area under the curve (AUC) metrics were computed on held-out validation sets, both within (within-set) and across (across-set) reference sets (Figure 2), as well as within the union of the two sets. For overlapping drug/ADE pairs in combined CV and across-set CV, seven duplicate pairs were removed before CV.

Visualization

To achieve a low-dimensional approximation of the data set for visual interrogation, t-distributed stochastic neighbor embedding (tSNE)[77] was used with a learning rate of 200.0 and perplexity of 30. Pairs in this low-dimensional space were labeled according to the ADE they were composed with and according to their ground-truth assertion in their reference set of origin.

Generalization

For generalization assessment, a list of drugs was downloaded from the side-effect database SIDER, version 4.1, containing 1430 drugs.[78,79] Drug/ADE pair representations for each of the drugs resident in our vector spaces were composed for each of the four OMOP set ADEs, which we selected as cues for prediction because the OMOP set provides sufficient positive and negative examples to derive a robust model for each of them. Pairs included in the OMOP set were removed, as were pairs contained in the high-performance subset of the MEDication Indication resource (MEDI) database,[80] to prevent inadvertent recovery of therapeutic relationships. After removal of MEDI indications and reference set pairs, the final number of unique drug/ADE pairs derived from SIDER for MI, GIB, KI, and LI were 1138, 1186, 1150, and 1155, respectively. After training a LR model on the full OMOP set (with identical configuration to CV experiments), we rank-ordered its predictions on the SIDER-derived test set. The top 10 predictions for each ADE were then manually evaluated by searching FDA and/or United Kingdom (UK)/European Medicines Agency (EMA) drug labels. Additionally, for every drug/ADE pair, we mined the extracted label information contained in SIDER to assess whether highly ranked predictions

Figure 3. Schematic overview. Input data from SemMedDB are processed and then encoded into a distributed vector space as described in **Unsupervised Pre-training**. Composite vectors for Drug/ADE pairs are then generated from this vector store as described in **Generation of Composite Feature Vectors**. These vectors are visualized with tSNE as described in **Visualization**. The composite vectors are analyzed by labeling them as positive or negative according to the ground-truth assertion in the respective reference standards, and then machine learning is deployed as described in **Training and Cross-validation** and **Generalization**.

from our models were more likely to be mentioned in drug labels (to the extent the NLP-derived information available in SIDER is accurate) than lower ranked predictions. For this mining, a dictionary of several synonyms for each of the four ADEs (full list in Supplementary Appendix) was used to determine if SIDER had mined an association between a given drug/ADE pair. However, and as noted previously, drug terms were not expanded. Figure 3 provides a visual overview of the current research.

The code and data required to reproduce these experiments is available at https://github.com/jusger/ADEClassifier-RepLearnML.

RESULTS

Cross-validation performance

The results of our experiments across CV configurations are shown in Table 1 (F1 scores). Table 2 presents ROC AUC and F1 metrics for both the OMOP and EU-ADR reference standards for ESP-LR LOO and PSI-LR LOO configurations alongside results from prior research.[38,63]

Comparison between ESP and PSI

ESP-based models perform better than PSI-based models in LR configurations. However, this is not the case with kNN configurations, a finding consistent with previous research.[64] With ESP, LR models improve upon kNN performance in all comparisons between them (Table 1),[2] providing the best overall performance. Across models, cross-set LR performance was lower than within-set CV performance but was best preserved with ESP-based models. In examination of Table 2, PSI-LR has the highest AUC for GIB results on the

OMOP set, but ESP generally performs more consistently and with higher performance than PSI-based models across ADEs and reference standards, improving up to 0.11 in AUC over PSI on the EU-ADR reference set.

Comparison with prior methods

As shown in Table 2, ESP-based LR generally performs better on this set than the best results reported using generalized enrichment analysis (GEA).[38] GEA is of interest as a point of comparison, as it also leverages the biomedical literature, but differs in methodology. Winnenburg and Shah utilized GEA to detect signal from MEDLINE indexed information using terminological expansion at varying levels of abstraction to increase signal strength by mapping drugs and ADEs to related concepts.[38] On an ADE-by-ADE basis, ESP improves performance over GEA on MI and KI AUCs (0.765 to 0.979 and 0.929 to 0.947, respectively). Additionally, the best overall F1 score for any individual GEA model (that is, with all side effects at the same level of terminological expansion) reported by Winnenburg and Shah is 0.8 on the OMOP reference standard.[38] In contrast, ESP-based LR models attain a 0.901 F1, a 12.5% improvement.

Recent research presented by Voss et al.[37] provides another point of comparison. Their method utilized supervised machine learning (regularized linear regression), with classifiers trained on a range of manually engineered features integrated from multiple sources, including the biomedical literature, assertions extracted from it with SemRep, FAERS data, and pharmaceutical product labels.[37] These authors report AUCs for the full OMOP set only (without per ADE results), with a best overall AUC of 0.94 (compared to ESP-LR's 0.96 AUC), and no F-metrics reported. When rounded to the same precision, Voss et al.[37] and ESP present identical AUCs (0.92) for the EU-ADR reference set. Voss et al.[37] also present AUCs for subsets of SemMedDB information, which have greatly diminished performance (0.57-0.59 AUC) when compared to ESP or PSI models (0.809-0.960 AUC).

---

2   Although not shown, ESP models of lower dimensionality perform similarly to results reported in previous work examining PSI[63] and ESP,[64] with PSI requiring higher dimensionality than ESP to retain its performance.

Table 1. Cross-validation Performance (F1 scores). Results from LOO and S5F CV configurations are shown. OMOP is presented in internal CV in the first section, followed by EU-ADR, and finally the combined grouping of OMOP with EU-ADR, in which one set is used for training and the left out set for testing. Results presented throughout the table are the average +/− 2 times the standard deviation over 100 runs with random assignment to CV partitions on each run. The best results for each CV configuration are shown in **boldface**.

OMOP

| Model	ESP S5F F1-Score	ESP LOO F1-Score	PSI S5F F1-Score	PSI LOO F1-Score
kNN 1	0.839 +/− 0.017	0.852	0.846 +/− 0.018	0.855
kNN 2	0.848 +/− 0.024	0.869	0.875 +/− 0.021	0.890
kNN 5	0.793 +/− 0.022	0.804	0.840 +/− 0.021	0.852
kNN 10	0.766 +/− 0.023	0.771	0.821 +/− 0.020	0.835
Logistic Regression	**0.895 +/− 0.020**	**0.901 +/− 0.012**	0.835 +/− 0.035	0.848 +/− 0.013
EU-ADR				
kNN 1	0.658 +/− 0.070	0.660	0.730 +/− 0.056	0.760
kNN 2	0.620 +/− 0.085	0.675	0.620 +/− 0.086	0.667
kNN 5	0.550 +/− 0.081	0.587	0.618 +/− 0.092	0.704
kNN 10	0.486 +/− 0.116	0.200	0.491 +/− 0.135	0.203
Logistic Regression	**0.834 +/− 0.066**	**0.841 +/− 0.017**	0.662 +/− 0.098	0.745 +/− 0.028
EU-ADR + OMOP (Combined Internal)				
kNN 1	0.798 +/− 0.020	0.804	0.814 +/− 0.022	0.827
kNN 2	0.810 +/− 0.026	0.835	0.821 +/− 0.026	0.832
kNN 5	0.753 +/− 0.024	0.768	0.790 +/− 0.021	0.807
kNN 10	0.725 +/− 0.023	0.735	0.780 +/− 0.023	0.784
Logistic Regression	**0.886 +/− 0.021**	**0.911 +/− 0.009**	0.812 +/− 0.028	0.788 +/− 0.030
Train Set	**Test Set**	**Vector Base**	**F1 Score**	**Model**
OMOP	EU-ADR	ESP	**0.721 +/− 0.049**	LR
EU-ADR	OMOP	ESP	**0.626 +/− 0.018**	LR
OMOP	EU-ADR	PSI	0.331 +/− 0.059	LR
EU-ADR	OMOP	PSI	0.521 +/− 0.020	LR

Table 2. Receiver operating characteristic area under the curve (AUC) and F1 comparisons across OMOP and EU-ADR reference sets. For GEA, three abstraction (eg term expansion) levels are given, where higher values indicate more term expansion. GEA covers ≈95% of reference drug/ADE pairs. For Voss et al., the combined performance of nine predictive features is shown alongside performance for individual predictive features of clinical trial (CT) and case report (CR) subsets of SemMedDB information. Voss et al. covers ≈80% of drug/ADE pairs. ESP and PSI are presented in logistic regression leave one out cross-validation configurations, showing the average +/- 2 times the standard deviation over 100 runs. ESP/PSI models cover ≈99% of drug/ADE pairs. Shaded cells indicate results were not reported. The best results for each metric are shown in **boldface**. *indicates results as reported in previous work.[37,38]

	OMOP						EU-ADR	
	MI AUC	GIB AUC	LI AUC	KI AUC	Overall AUC	Overall F1	Overall AUC	Overall F1
GEA 4.5-7*	0.765	0.887	0.906	0.929		0.76		
GEA 7-10*	0.692	0.972	**0.93**	0.845		0.80		
GEA 1.5-5*						0.70		
Voss et al. Combined*					0.94		**0.92**	
Voss et al. SemMedDB CT*					0.58		0.57	
Voss et al. SemMedDB CR*					0.58		0.59	
ESP-LR LOO	**0.979 +/- 0.004**	0.934 +/- 0.008	0.920 +/- 0.005	**0.947 +/- 0.005**	**0.960 +/- 0.002**	**0.901 +/- 0.012**	**0.918 +/- 0.006**	**0.841 +/- 0.016**
PSI-LR LOO	0.960 +/- 0.008	**0.978 +/- 0.007**	0.825 +/- 0.015	0.945 +/- 0.008	0.946 +/- 0.004	0.848 +/- 0.013	0.809 +/- 0.015	0.742 +/- 0.025

Visualization of composite feature vectors

A tSNE plot for ESP-derived composite representations of drug/ADE pairs in the OMOP and EU-ADR reference standards is shown in Figure 4. Separation in the reduced dimensional space appears to first occur based on side effect, and within ADE specific clusters, there is some localization of ground-truth positive pairs (dark/saturated) versus ground-truth negative (light/pastel) pairs. The EU reference standard also shows clusters specific to side effects, with EU-ADR clusters for conserved ADEs co-localizing, while ADEs unique to the EU-ADR reference occupy disparate regions.
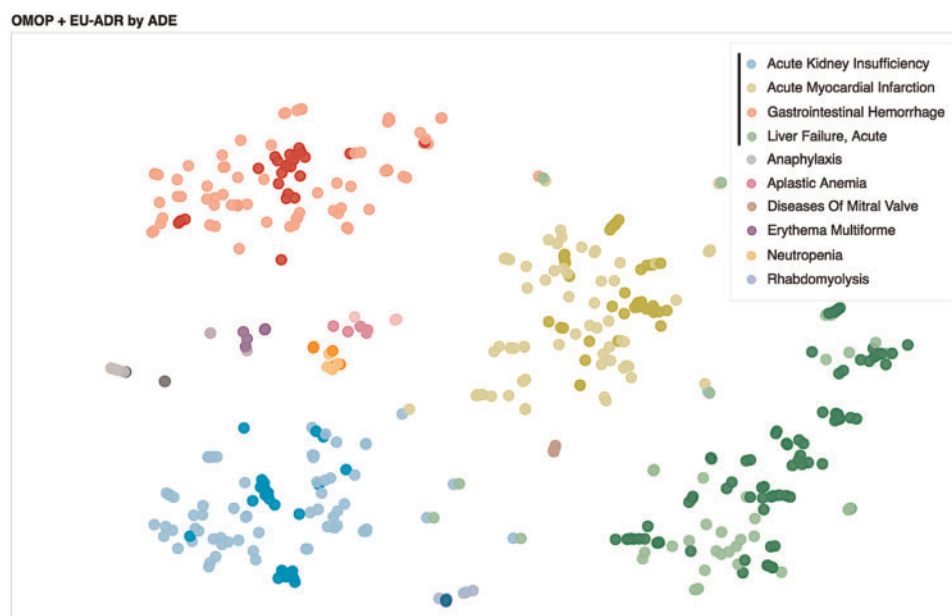
Figure 4. A tSNE plot of the compositional drug/ADE pair embeddings generated from the unsupervised pretraining step with ESP. Conserved ADE examples between the EU-ADR and OMOP reference standards (indicated by black legend bar) localize together in their respective ADE spaces. Despite the highly compressed representation, some delineation between positive (dark/saturated glyphs) and negative (light/pastel glyphs) spaces can be seen.

Generalization to unseen drugs

The top 10 rank ordered LR-ESP predictions for approximately one thousand previously unseen drugs from the SIDER database for each of the four ADEs are shown in Table 3, with label information and additional comments from manual review. URLs for labels consulted for each drug can be found in the Supplementary Appendix.

Support was found for 37 of the 40 top-ranked predictions, including the high-profile association between rofecoxib and MI. This corresponds to a mean precision at k = 10 of 0.925 across ADEs. Two of the three remaining predictions were related to the side effect in question: isosorbide dinitrate is typically used to treat coronary artery disease, but abrupt cessation can cause myocardial infarction in physically dependent patients; and although label information was unavailable for oral Suprofen, literature evidence does exist supporting an association with GIB.[83] In the case of amlodipine, while no label information was present, disproportionality measures on FAERS data (as seen in Table 3) meet criteria for further investigation specified by Evans et al.[82] when defining the proportional reporting ratio (PRR), a statistical measure adopted by the FDA to aid PV.[84]

Figure 5 shows a comparison of top-ranked predictions versus low-ranked predictions for all four OMOP set side effects at various ranks in terms of label support, as found via matching to NLP-mined product label information contained in the SIDER database. Seventy-two of the top 100 ranked drug/ADE pairs had label support; seven of the bottom-ranked 100 drug/ADE pairs had label support. This trend is diminished as more drugs are considered, with 469 out of the top 1000 drug/ADE pairs having label support compared with 107 drug/ADE pairs in the bottom 1000. Generally, the higher a drug/ADE pair is ranked by our method, the greater the chance SIDER will contain label information connecting that drug to that ADE.

DISCUSSION
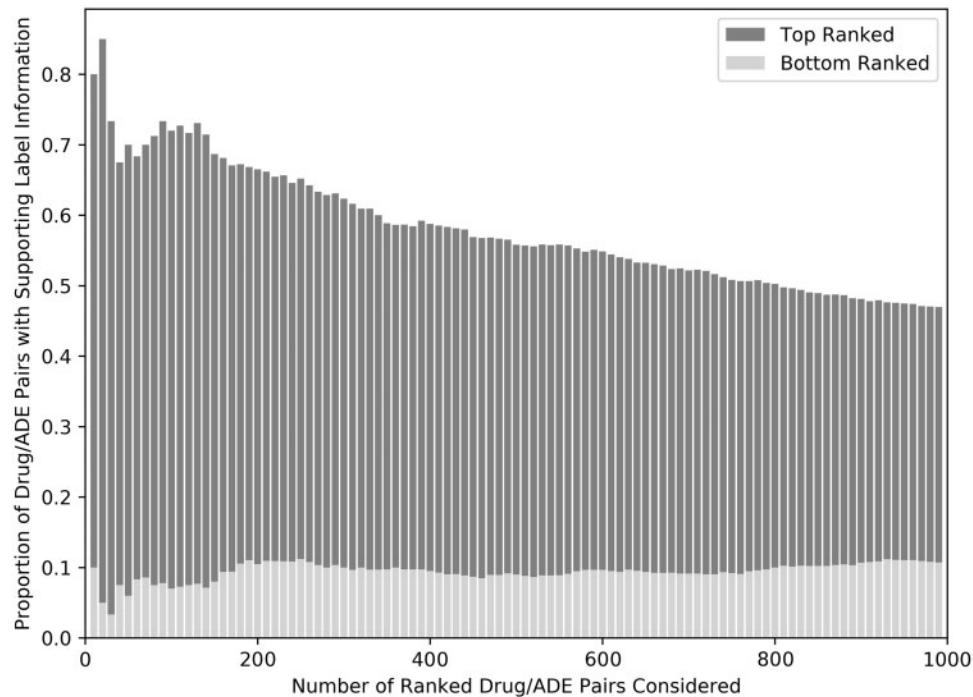
Advantages over existing co-occurrence methods

When compared to existing methods, such as those presented by Winnenburg and Shah[38] and Voss et al.,[37] ESP- and PSI-based models presented here have several advantages. With respect to performance, our results set the state of the art on the OMOP reference standard, and are equal to those reported by Voss et al.[37] on the EU-ADR standard. Furthermore, in contrast to previously published methodologies (such as those described in[37,38]) that operate on explicit drug/ADE co-occurrence events, our method presented here does not require co-occurrence for drug/ADE pairs (eg no direct co-occurrence is required in SemMedDB to generate performant models). Rather, the distributed representations upon which our models depend carry information concerning drug mechanisms and disease pathophysiology (among other constituents), information that can be leveraged for downstream supervised machine learning. Consequently, our methods may be better positioned to detect emerging side effects, which have yet to be described in detail in the literature.

Additionally, our approach does not require terminological expansion on account of the representational pretraining offered by ESP/PSI. As similar concepts have similar vectors, there is no need for expansion or cross-linking of concepts (eg mapping drugs to their active ingredients). In GEA, this expansion plays a pivotal role, as in order to achieve optimal performance, an optimal degree of term expansion abstraction for each side effect must be identified (a process that requires labeled training data).[38] This tuning is important, as there is not a consistently best performing level of abstraction across the OMOP reference standard for GEA. In contrast, ESP-based L1 logistic regression models are trained using labeled training data and MEDLINE-indexed information, but without recourse to term expansion. This becomes especially important for coverage and signal enhancement. For example, results in Table 2 are not strictly comparable, as only a subset of around 80% of each

Table 3. Rank-ordered predictions derived from training on the full OMOP set and testing on a list of unseen drugs derived from the SIDER resource. For drugs with readily available information found in the FDA label, only the FDA label information was considered. For drugs without availability in the United States, UK/EMA label information was assessed. Suprofen was discontinued, and label information was unavailable for qualitative analysis; comments are speculative. Only amlodipine did not have support on the label for the predicted ADE (kidney injury).

Myocardial Infarction

Drug	Product Label	Comments
Naproxen	On FDA Label	
Ibuprofen	On FDA Label	
Hydralazine	On FDA Label	In "overdosage" section; myocardial ischemia leading to myocardial infarction; angina pectoris / tachycardia in "adverse reactions" section
Isosorbide Dinitrate	Not on label as adverse effect	Usually used to treat angina pectoris due to coronary artery disease; warning for those with MI or congestive heart failure to avoid tachycardia and hypotension; abrupt cessation of nitrates causes acute MI in those with physical dependence
Rofecoxib	On FDA Label	Withdrawn from market in 2004 over concerns of acute MI
Etoricoxib	On UK/EMA Label	Not available in United States
Diclofenac	On FDA Label	
Tenoxicam	On UK/EMA Label	Not available in United States
Meloxicam	On FDA Label	
Mefenamic Acid	On FDA Label	

Gastrointestinal Bleed

Drug	Product Label	Comments
Tenoxicam	On UK/EMA Label	Not available in United States
Rofecoxib	On FDA Label	
Etoricoxib	On UK/EMA Label	Not available in United States
Diclofenac	On FDA Label	
Aspirin	On FDA Label	
Celecoxib	On FDA Label	
Mefenamic Acid	On FDA Label	
Parecoxib	On UK/EMA Label	Not available in United States
Acenocoumarol	On FDA Label	
Suprofen	N/A	Discontinued; oral tablet may have caused GIB similar to other NSAIDs; ophthalmic solution unlikely

Liver Injury

Drug	Product Label	Comments
Pravastatin	On FDA Label	
Atorvastatin	On FDA Label	
Fluvastatin	On FDA Label	
Pentoxifylline	On FDA Label	
Lovastatin	On FDA Label	
Simvastatin	On FDA Label	
Pirfenidone	On FDA Label	Elevated enzyme levels
Ticlopidine	On FDA Label	
Sorafenib	On FDA Label	
Rosuvastatin	On FDA Label	

Kidney Injury

Drug	Product Label	Comments
Tenoxicam	On UK/EMA Label	Not available in United States
Flurbiprofen	On FDA Label	
Quinapril	On FDA Label	
Nabumetone	On FDA Label	
Amlodipine	Not on label as adverse effect	Only connection via FDA label is that of affecting urine output; additionally, no contraindication with renal impairment; however, calculating metrics on FAERS data through 2018-01-12 yields PRR = 3.11, X^2 Yates = 3633.48, p<.0001, and 2704 reported cases,[81] meeting criteria for further investigation[82]
Rofecoxib	On FDA Label	
Etoricoxib	On UK/EMA Label	Not available in United States
Benazepril	On FDA Label	No discontinuation of the product, but elevated levels of serum creatinine and blood urea nitrogen
Perindopril	On FDA Label	
Cilazapril	On UK/EMA Label	Not available in United States

Figure 5. Comparison of the proportion of highest- and lowest-ranked drug/ADE pairs for label support in SIDER, as indicated by a match between a drug and a small list of ADE terms (see Supplementary Appendix) in ADE label information extracted from the SIDER database. Dark bars denote the proportion of top-ranked predictions that have support, and light bars denote the proportion of bottom (lowest)-ranked predictions that have label support. For example, the left-most bar indicates both the proportion of the top 10 ranked predictions that have label support (dark bar), and the proportion of the bottom 10 predictions that have label support (light bar). Moving left to right in the figure, the number of ranked pairs considered increases from 10 up to 1000 top- and bottom-ranked drug/ADE pairs in increments of 10. In total, the graph represents the top ≈20% (1000 of 4629 total drug/ADE predictions) and bottom ≈20% of all predictions.

reference standard was available in Voss et al.,[37] and 94 % to 95% in the GEA analysis (depending on mapping level),[38] compared to ≈99% for our methods. Not only does our method perform better on the OMOP reference standard, but we also maintain greater coverage, as we do not require sizeable direct associations for detectable signal. The capacity for accurate prediction without direct co-occurrence is further indicated by the stark difference in performance between Voss et al.'s use of SemMedDB-derived information and our models. That ESP-based models perform better overall than GEA (with a substantial improvement on MI-related side effects in particular) and match or exceed the performance documented by Voss et al.[37] using SemMedDB features supports the hypothesis that considering implicit relationships can enhance the performance of literature-based PVmethods.

ESP and PSI with machine learning

Although previous research showed that with the simple algorithm of kNN classification, PSI performed better on this classification task than did ESP, additional machine learning approaches had not been evaluated using ESP prior to the current research.[64] While our results with kNN mirror those reported previously, with L1 LR, ESP demonstrates significantly increased performance on the majority of OMOP ADEs, overall on OMOP, and on the EU-ADR standard. At times, the improvement is as much as 14%. This advantage may be due to ESP's enhanced capacity for similarity-based inference relative to PSI.[64] With more consistent and better overall performance, our findings support the hypothesis that ESP offers advantages over PSI as a basis for supervised machine learning.

Of note, OMOP results are better than EU-ADR results for both PSI and ESP models. We suspect this is likely due to a smaller ADE space (four in OMOP versus 10 in EU-ADR) and more examples per ADE for the OMOP reference set. This may also explain the larger degradation in performance of kNN at larger k in EU-ADR results relative to OMOP results. Such results suggest performance is contingent upon availability of sufficient numbers of training examples for each side effect of interest, further evidenced by diminished performance when training and testing are split across reference standards with only partially overlapping side effects.

Visualization and generalization

Although relative cluster size, density, and inter-cluster distances are not especially meaningful in tSNE diagrams, clusters themselves are likely to represent underlying data set structure.[85] When examining the tSNE plot for drug/ADE pairs for the OMOP and EU-ADR reference sets, the intra-ADE clustering of positive examples versus negative examples explains the utility of these compositional distributed representations as a basis for supervised machine learning with simple algorithms—in many cases, it is possible to discern a likely classification boundary, even with reduction to two dimensions. This observation, together with the clustering by side effect, explain the reduction in performance when attempting to generalize to previously unseen ADEs, as these classification boundaries would be located within ADE-specific clusters. In contrast, as both OMOP and EU-ADR drug/ADE pairs colocalize for synonymous ADEs, this tSNE plot does support the hypothesis that trained models may generalize to previously unseen drugs paired with previously seen ADEs.

With this in mind, our generalization analysis looked only at the four ADEs in the larger OMOP reference standard. On qualitative assessment, results appear very promising, with ≈93% of the top-ranked drug/ADE pairs having some form of label support. Furthermore, in the case of amlodipine for KI, there is some indication that this may be a previously unrecognized side effect, as though label information is absent, and the association is consistent with results from a disproportionality analysis of FAERS data. In addition, a coarse-grained quantitative analysis of the proportion of predictions at different ranks that correspond to drug/ADE relationships asserted in the NLP-derived SIDER database showed a 10-fold increase for the top 100 ranked predictions as compared with the bottom 100 ranked predictions. While this suggests a considerably lower precision at k = 100 (of 0.72) than our manually evaluated precision at k = 10 (of ≈0.93), there is some indication that ostensible false positive relationships (ie relationships not in SIDER that are highly ranked) may constitute side effects missing from SIDER on account of NLP errors. For example, in the case of hydralazine, our mining of SIDER for a link to MI returns false, yet in the qualitative assessment, information can be found that strongly links hydralazine to MI. Others may be as-yet unrecognized side effects, as suggested by qualitative analysis in the case of amlodipine/KI. These findings support the hypothesis that trained models can generalize to unseen drugs when adequate training data for an ADE are available.

Limitations

The most prominent limitations to this work exist in the generalization analysis. The qualitative analysis covers only a small portion of drugs queried, and the coarse-grained quantitative analysis of mining SIDER-extracted label information is challenged by limitations in recall and precision for the NLP that generated the information in SIDER, and by our ability to mine such assertions, which required a small amount of terminological mapping (the Supplementary Appendix contains the set of terms queried for each ADE). Additionally, 337 drugs from SIDER did not have a direct string match in our vector stores, and required manual mapping, which resolved all but 138 (≈9.7%). Using SIDER as a point of comparison in this way requires the very terminological mapping and expansion that we seek to mitigate or obviate with our methods here. As such, we still see tremendous value for terminological mapping and abstraction methodologies to aid and guide further research, and permit integration of observational data sources with our methods as they evolve.

Additionally, a number of therapeutic indications were removed from consideration during the generalization task; as the mechanisms of drugs in treating or causing a particular effect may overlap, it seems likely that our models will at times recover therapeutic indications instead of side effects. As these entities can be readily and automatically removed using existing reference stores in a PV pipeline, such as the MEDI resource (as done here), we consider this a minor limitation. Finally, as with other supervised machine learning approaches, additional labeled training examples are likely to increase scope and generalization performance across reference sets and to unseen pairs. However, manual curation of these examples would require significant, continued human effort in this domain.[61]

Future work

An important direction for future work concerns the evaluation of our methods using the time-delimited reference standard provided by Harpaz et al.,[86] which will permit assessment of their performance for emerging side effects;[87,88] estimation of their impact on public health (manifesting as earlier ADE detection); and evaluation of the hypothesis that leveraging implicit relationships permits earlier detection of drug/ADE relationships than is possible with methods requiring explicit drug/ADE co-occurrence. Expanding our models with additional data sources, such as spontaneous reporting data, is another area left for future work. Additionally, it may be the case that incorporating therapeutic indications as negative examples in the training set eliminates the need for post-process removal of indications using a reference such as the MEDI resource, a direction we have yet to explore.

CONCLUSIONS

CV performance utilizing the approaches presented here exceeds that reported previously, even accounting for methodologies incorporating information from the literature, SRS, drug product labels, and/or additional sources, such as those used by Voss et al. and Winnenburg and Shah.[37,38] These results indicate that ESP-derived representations provide a basis for robust performance without terminological expansion, with advantages over our previous approach (PSI) as a basis for machine learning, given a suitable supervised learning algorithm. While performance is influenced by the availability of examples to develop a robust model for each ADE, trained models can generalize to previously unseen drugs, as indicated by the evidence supporting predictions for the four ADEs in the OMOP set. As these methods leverage implicit relationships, we view them as complementary to existing approaches based on explicit co-occurrence in the literature and other data sources such as FAERS. Of note, our methods produce state-of-the-art performance on two widely used reference standards utilizing literature-derived relational information only. It seems likely that their integration as a component of an ensemble of PV signal detection methods would further improve performance, as has been the case in prior evaluations of multimodal signal integration.[27,37,89]

CONTRIBUTORS

All authors meet the guidelines as established by the ICMJE for authorship. Justin Mower is the primary author and was responsible for the majority of the analysis and writing. Revision, approval, and guidance in the design of experiments and writing were given by co-authors Devika Subramanian and Trevor Cohen.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Ryan PB, Schuemie MJ, Welebob E, *et al*. Defining a reference set to support methodological research in drug safety. *Drug Saf* 2013; 36 (S1): 33–47.

2. Meyboom RH, Hekster YA, Egberts AC, *et al*. Causal or casual? The role of causality assessment in pharmacovigilance. *Drug Saf* 1997; 17 (6): 374–89.

3. Harpaz R, Callahan A, Tamang S, *et al*. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Saf* 2014; 37 (10): 777–90.

4. Swanson DR, Smalheiser NR. Undiscovered public knowledge: a ten-year update. In: KDD. 1996: 295–8. https://ocs.aaai.org/Papers/KDD/1996/KDD96-051.pdf. Accessed July 13, 2017.

5. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005; 6 (1): 57–71.

6. National Center for Health Statistics. Health, United States, 2016: With Chartbook on Long-term Trends in Health. Hyattsville, MD; 2017. https://www.cdc.gov/nchs/data/hus/hus16.pdf. Accessed July 10, 2017.

7. Hing E, Rui P, Palso K. National Ambulatory Medical Care Survey: 2013 State and National Summary Tables. 2014. http://www.cdc.gov/nchs/ahcd/ahcd_products.htm. Accessed July 10, 2017.

8. Center for Disease Control and Prevention. National Hospital Ambulatory Medical Care Survey: 2011 Outpatient Department Summary Tables. 2012. https://www.cdc.gov/nchs/data/ahcd/nhamcs_outpatient/2011_opd_web_tables.pdf. Accessed July 10, 2017.

9. Rui P, Kang K, Albert M. *National Hospital Ambulatory Medical Care Survey: 2013 Emergency Department Summary Tables*. 2014. http://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2013_ed_web_tables.pdf. Accessed July 10, 2017.

10. Watanabe JH, McInnis T, Hirsch JD. Cost of prescription drug-related morbidity and mortality. *Ann Pharmacother* 2018; 1060028018765159.

11. Stausberg J. International prevalence of adverse drug events in hospitals: an analysis of routine data from England, Germany, and the USA. *BMC Health Serv Res* 2014; 14 (1): 125.

12. Bourgeois FT, Shannon MW, Valim C, *et al*. Adverse drug events in the outpatient setting: an 11-year national analysis. *Pharmacoepidemiol Drug Saf* 2010; 19 (9): 901–10.

13. Coloma PM, Trifirò G, Patadia V, *et al*. Postmarketing safety surveillance: where does signal detection using electronic healthcare records fit into the big picture? *Drug Saf* 2013; 36: 183–97.

14. FDA. Postmarket Drug Safety Information for Patients and Providers—Information for Healthcare Professionals: Valdecoxib (Marketed as Bextra). https://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm124649.htm Accessed July 11, 2017.

15. Ray WA, Griffin MR, Stein CM. Cardiovascular toxicity of Valdecoxib. *N Engl J Med* 2004; 351 (26): 2767.

16. Downing NS, Shah ND, Aminawung JA, *et al*. Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010. *JAMA* 2017; 317 (18): 1854–63.

17. Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. *J Pharmacol Pharmacother* 2013; 4 (5): 73–7.

18. World Health Organization. The Importance of Pharmacovigilance. 2002. http://apps.who.int/iris/bitstream/10665/42493/1/a75646.pdf. Accessed July 10, 2017.

19. Center for Drug Evaluation and Research. Questions and Answers on FDA's Adverse Event Reporting System (FAERS). 2016. https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugeffects/ Accessed July 19, 2017.

20. Center for Drug Evaluation and Research. FDA Adverse Events Reporting System (FAERS)—Reports Received and Reports Entered into FAERS by Year. https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070434.htm Accessed July 16, 2017.

21. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions: a systematic review. *Drug Saf* 2006; 29 (5): 385–96.

22. Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Determinants of under-reporting of adverse drug reactions: a systematic review. *Drug Saf* 2009; 32 (1): 19–31.

23. Sakaeda T, Tamon A, Kadoyama K, *et al*. Data mining of the public version of the FDA adverse event reporting system. *Int J Med Sci* 2013; 10 (7): 796–803.

24. Pariente A, Gregoire F, Fourrier-Reglat A, *et al*. Impact of safety alerts on measures of disproportionality in spontaneous reporting databases: the notoriety bias. *Drug Saf* 2007; 30 (10): 891–8.

25. Naidu RP. Causality assessment: a brief insight into practices in pharmaceutical industry. *Perspect Clin Res* 2013; 4 (4): 233–6.

26. Harpaz R, DuMouchel W, LePendu P, *et al*. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clin Pharmacol Ther* 2013; 93 (6): 539–46.

27. Li Y, Ryan PB, Wei Y, *et al*. A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. *Drug Saf* 2015; 38 (10): 895–908.

28. Natsiavas P, Koutkias V, Maglaveras N. Exploring the capacity of open, linked data sources to assess adverse drug reaction signals. In: Semantic Web applications and tools for life sciences (SWAT4LS) International Conference, held at Cambridge, England Dec. 7-10th 2015. 2015: 224–6.

29. Koutkias VG, Jaulent M-C. Computational approaches for pharmacovigilance signal detection: toward integrated and semantically enriched frameworks. *Drug Saf* 2015; 38 (3): 219–232.

30. Food and Drug Administration. Guidance for Industry: Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment. Rockville, MD: Food and Drug Administration; 2005.

31. European Medicines Agency. Good Pharmacovigilance Practices. http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000345.jsp Accessed July 20, 2017.

32. Zweigenbaum P, Demner-Fushman D, Yu H, *et al*. Frontiers of biomedical text mining: current progress. *Brief Bioinform* 2007; 8 (5): 358–75.

33. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. *Nat Rev Genet* 2012; 13 (12): 829–39.

34. Cohen T, Schvaneveldt R, Widdows D. Reflective random indexing and indirect inference: a scalable method for discovery of implicit connections. *J Biomed Inform* 2010; 43 (2): 240–56.

35. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif Intell* 1997; 91 (2): 183–203.

36. Henry S, McInnes BT. Literature based discovery: models, methods, and trends. *J Biomed Inform* 2017; 74: 20–32.

37. Voss EA, Boyce RD, Ryan PB, *et al*. Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform* 2017; 66: 72–81.

38. Winnenburg R, Shah NH. Generalized enrichment analysis improves the detection of adverse drug events from the biomedical literature. *BMC Bioinformatics* 2016; 17 (1): 250.

39. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986; 30 (1): 7–18.

40. Swanson DR. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med* 1988; 31 (4): 526–57.

41. 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 (2): 158–64.

42. Hristovski D, Burgun-Parenthoine A, Avillach P, *et al*. Towards using literature-based discovery to explain drug adverse effects. In: *24th International Conference of the European Federation for Medical Informatics Quality of Life through Quality of Information*. MIE. 2012. http://person.hst.aau.dk/ska/mie2012/AllPresentations/422.pdf. Accessed July 18, 2017.

43. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003; 36 (6): 462–77.

44. Hristovski D, Friedman C, Rindflesch TC, *et al*. Exploiting semantic relations for literature-based discovery. In: AMIA Annual Symposium Proceedings, 2006, held in Washington, DC, USA Nov. 11–15.

45. Song D, Bruza P, Cole R. Concept Learning and Information Inferencing on a High-Dimensional Semantic Space. 2004. http://oro.open.ac.uk/35506/. Accessed October12, 2017.

46. Gordon MD, Dumais S. Using Latent Semantic Indexing for Literature Based Discovery. 1998. https://deepblue.lib.umich.edu/handle/2027.42/34255. Accessed October12, 2017.

47. Cohen T, Widdows D, Schvaneveldt RW, et al. Discovering discovery patterns with predication-based semantic indexing. *J Biomed Inform* 2012; 45 (6): 1049–65.

48. Lever J, Gakkhar S, Gottlieb M, et al. A collaborative filtering based approach to biomedical knowledge discovery. *Bioinformatics*. doi:10.1093/bioinformatics/btx613.

49. Ahlers CB, Hristovski D, Kilicoglu H, et al. Using the literature-based discovery paradigm to investigate drug mechanisms. *AMIA Annu Symp Proc* 2007; 2007: 6–10.

50. Zhang R, Adam TJ, Simon G, et al. Mining biomedical literature to explore interactions between cancer drugs and dietary supplements. *AMIA Jt Summits Transl Sci Proc* 2015; 2015: 69–73.

51. Cohen T, Widdows D, De Vine L, et al. Many paths lead to discovery: analogical retrieval of cancer therapies. In: *6th International Symposium, QI 2012, Paris, France, June 27-29, 2012, Revised Selected Papers*. Paris, France: Springer; 2012: 90–101.

52. Cohen T, Widdows D, Schvaneveldt RW, et al. Discovery at a distance: farther journeys in predication space. In: *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*. IEEE. 2012: 218–25. http://ieeexplore.ieee.org/abstract/document/6470307/. Accessed July 18, 2017.

53. Shang N, Xu H, Rindflesch TC, et al. Identifying plausible adverse drug reactions using knowledge extracted from the literature. *J Biomed Inform* 2014; 52: 293–310.

54. Liu M, Wu Y, Chen Y, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc* 2012; 19 (e1): e28–35.

55. Caster O, Sandberg L, Bergvall T, et al. vigiRank for statistical signal detection in pharmacovigilance: first results from prospective real-world use. *Pharmacoepidemiol Drug Saf* 2017; 26 (8): 1006–10.

56. Huang L-C, Wu X, Chen JY. Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. *Proteomics* 2013; 13 (2): 313–24.

57. Jamal S, Goyal S, Shanker A, et al. Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models. *Sci Rep* 2017; 7 (1): 872.

58. Bengio Y, Courville AC, Vincent P. Unsupervised feature learning and deep learning: a review and new perspectives. *CoRR Abs12065538* 2012; 1. https://pdfs.semanticscholar.org/f8c8/619ea7d68e604e40b814b40-c72888a755e95.pdf. Accessed October 12, 2017.

59. Erhan D, Bengio Y, Courville A, et al. Why does unsupervised pretraining help deep learning? *J Mach Learn Res* 2010; 11: 625–60.

60. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 2013; 35 (8): 1798–828.

61. Sun C, Shrivastava A, Singh S, et al. Revisiting unreasonable effectiveness of data in deep learning era. *ArXiv170702968 Cs*. 2017. http://arxiv.org/abs/1707.02968. Accessed July 19, 2017.

62. Khodak M, Risteski A, Fellbaum C, et al. Extending and improving wordnet via unsupervised word embeddings. *ArXiv170500217 Cs*. 2017. http://arxiv.org/abs/1705.00217.

63. Mower J, Subramanian D, Shang N, et al. Classification-by-analogy: using vector representations of implicit relationships to identify plausibly causal drug/side-effect relationships. *AMIA Annu Symp Proc* 2016; 2016: 1940–9.

64. Cohen T, Widdows D. Embedding of semantic predications. *J Biomed Inform* 2017; 68: 150–66.

65. Cohen T, Schvaneveldt RW, Rindflesch TC. Predication-based semantic indexing: permutations as a means to encode predications in semantic

66. Widdows D, Cohen T. Reasoning with vectors: a continuous model for fast robust inference. *Log J IGPL Interest Group Pure Appl Log* 2015; 23 (2): 141–73.

67. Cohen T, Widdows D, Schvaneveldt R, et al. Finding Schizophrenia's prozac emergent relational similarity in predication space. In: *Quantum Interaction*. Berlin, Heidelberg: Springer; 2011: 48–59. doi:10.1007/978-3-642-24971-6_6.

68. Kilicoglu H, Shin D, Fiszman M, et al. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinforma Oxf Engl* 2012; 28 (23): 3158–60.

69. Widdows D, Ferraro K. Semantic vectors: a scalable open source package and online technology management application. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco.

70. Kanerva P. Binary spatter-coding of ordered K-tuples. *Artif Neural Networks—ICANN 96*, Bochum, Germany: Artificial Neural Networks - ICANN 96 (Springer). 1996: 869–73.

71. Gayler RW, Wales R. Connections, Binding, Unification and Analogical Promiscuity. 1998. http://cogprints.org/500. Accessed July 10, 2017.

72. Plate TA. *Holographic Reduced Representation: Distributed Representation for cognitive structures*, Chicago, IL: University of Chicago Press. 2003.

73. Rachkovskij DA, Kussul EM. Binding and normalization of binary sparse distributed representations by context-dependent thinning. *Neural Comput* 2001; 13 (2): 411–52.

74. Coloma PM, Avillach P, Salvo F, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. *Drug Saf* 2013; 36 (1): 13–23.

75. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *Front Neuroinform* 2014; 8: 2825–30.

76. Continuum Analytics. Anaconda Python Distribution. https://www.anaconda.com/. Accessed October 12, 2017.

77. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; 9: 2579–605.

78. Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010; 6: 343.

79. Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016; 44 (D1): D1075–79. Accessed October 12, 2017.

80. Wei W-Q, Cronin RM, Xu H, et al. Development and evaluation of an ensemble resource linking medications to their indications. *J Am Med Inform Assoc* 2013; 20 (5): 954–61.

81. Böhm R, Hehn L, von Herdegen T, et al. OpenVigil FDA—Inspection of U.S. American Adverse Drug Events Pharmacovigilance Data and Novel Clinical Applications. *PLoS One* 2016; 11 (6): e0157753.

82. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001; 10 (6): 483–6.

83. Porro GB, Pace F. 5 Ulcerogenic drugs and upper gastrointestinal bleeding. *Baillières Clin Gastroenterol* 1988; 2 (2): 309–27.

84. Duggirala HJ, Tonning JM, Smith E, et al. Use of data mining at the food and drug administration. *J Am Med Inform Assoc* 2016; 23 (2): 428–34.

85. Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. *Distill* 2016; 1: e2.

86. Harpaz R, Odgers D, Gaskin G, et al. A time-indexed reference standard of adverse drug reactions. *Sci Data* 2014; 1: 140043.

87. Norén GN, Caster O, Juhlin K, et al. Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. *Drug Saf* 2014; 37 (9): 655–9.

88. Harpaz R, DuMouchel W, Shah NH. Comment on: "Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance". *Drug Saf* 2015; 38 (1): 113–4.

89. Harpaz R, DuMouchel W, Schuemie M, et al. Toward multimodal signal detection of adverse drug reactions. *J Biomed Inform* 2017; doi:10.1016/j.jbi.2017.10.013.