Published in final edited form as:

Drug Saf. 2020 January ; 43(1): 67–77. doi:10.1007/s40264-019-00872-9.

Complementing Observational Signal with Literature-derived Distributed Representations for Post-marketing Drug Surveillance

Justin Mower¹, Trevor Cohen², Devika Subramanian¹

¹Rice University, Department of Computer Science, Houston TX, 77018

²University of Washington, Biomedical Informatics and Medical Education, Seattle WA, 98195

Abstract

Introduction: On account of well-documented limitations of data collected by spontaneous reporting systems (SRS), such as bias and under-reporting, a number of authors have evaluated the utility of other data sources for the purpose of pharmacovigilance, including the biomedical literature. Previous work has demonstrated the utility of literature-derived distributed representations (concept embeddings) with machine learning for the purpose of drug side-effect prediction. In terms of data sources, these methods are complementary, observing drug safety from two different perspectives (knowledge extracted from the literature and statistics from SRS data). However, the combined utility of these pharmacovigilance methods has yet to be evaluated.

Objective: This research investigates the utility of directly or indirectly combining observational signal from SRS with literature-derived distributed representations into a single feature vector or in an ensemble approach for downstream machine learning (logistic regression).

Methods: Leveraging a recently developed representation scheme, concept embeddings were generated from relational connections extracted from the literature and composed to represent drug and associated adverse reactions, as defined by two reference standards of positive (likely causal) and negative (no causal evidence) pairs. Embeddings were presented with and without common measures of observational signal from SRS sources to logistic regressors, and performance was evaluated with the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) metric.

Results: ROC AUC performance with these composite models improves up to $\approx 20\%$ over SRS-based disproportionality metrics alone and exceeds the best prior results reported in the literature when models leverage both sources of information.

Conclusions: Results from this study support the hypothesis that knowledge extracted from the literature can enhance the performance of SRS-based methods (and vice versa). Across reference sets, using literature and SRS information together performed better than using either source

Corresponding Author: Justin Mower, 713-348-2596, jmm21@rice.edu.

Conflicts of Interest:

Justin Mower, Trevor Cohen, and Devika Subramanian have no conflicts of interest relevant to the content of this study.

¹Further information on the reference standards used in this study can be found in Section 2.2.

alone, providing strong support for the complementary nature of these approaches to post-marketing drug surveillance.

1. Introduction

Nearly half of the US population interacts with a prescription drug every single month, and approximately three quarters of all doctor's visits result in pharmaceutical intervention[1-4]. Unfortunately, this prevalence of therapeutic drug use also carries with it a prevalence of adverse drug reactions (ADRs). Common in both outpatient and hospital settings[2,4-7], nonoptimal medication therapy was estimated to cost over 500 billion USD in 2016, approximately 16% of all healthcare expenditures that year[8]. These costs derive from increased patient morbidity and mortality and associated secondary treatments[9]. While many of these ADR risks are assessed by clinicians in a risk-benefit analysis for patient health, some adverse reactions remain undiscovered while being prescribed on-market in clinical practice. A recent study, following newly approved drugs between 2001 and 2010, found that 1 in 3 drugs required some sort of post-market safety update, such as a label warning or withdrawal, in just the few years following their release[10]. While monitoring drugs for safety is a cornerstone of modern pharmaceutical practice, there clearly remains a tremendous opportunity to assist in the monitoring and detection of on-market ADRs[11].

Traditionally, supervision of drugs after they have been released to market has occurred through monitoring of drug safety incidents submitted to Spontaneous Reporting Systems (SRS), such as the FDA Adverse Event Reporting System (FAERS). Although SRS systems carry bias and other reporting issues[12-14], they remain a heavily utilized and primary source of information for pharmacovigilance activities. These systems are also large aggregates of information[15], making exhaustive manual processing intractable. Consequently, methods have been developed which mine this aggregate for safety signals[16-18]. In general, these methods derive their estimates from 2x2 contingency tables, with the Proportional Reporting Ratio (PRR) and Reporting Odds Ratio (ROR) [16,19] being two prominent examples. These ratios compare the observed counts to expected counts (in the contingency table) and allow for quantification of the additional odds (or risk) of the drug and event compared to general background noise. These ratios can be thought of as disproportionality measures (DPMs) in this sense, as the ratios are a measure of disproportional reporting of two concepts together against expected odds and background.

After a signal is identified, experts are still required to manually analyze this information for plausible causality and relative risk[20,21]. During this process, the biomedical literature is often consulted for information which can plausibly link (in terms of a causal mechanistic explanation) a putative ADR signal[22]. From this, downstream regulatory action can be taken in terms of market relabeling or withdrawal, if necessary. On account of the growing and already expansive literature space, methods have been proposed to facilitate literature-based pharmacovigilance by automatically mining the literature for possible connections between drugs and potential adverse reactions[23-25]. Often, statistics derived from explicit co-occurrence of concepts are leveraged[24,26-28]. However, emerging side-effects may not explicitly co-occur with causative drugs in the literature. Rather, these entities may instead be connected *implicitly* via a third concept. This idea of implicit connections underlies the

field known as Literature-based Discovery (LBD)[29-35]. A number of authors have applied LBD methods to identify drug-reaction connections[23,36-40]. Some of this work goes beyond co-occurrence based statistics to consider the nature and directionality of the relational connections between concepts, as extracted using Natural Language Processing (NLP)[36-38,40,41]. Such relational connections are crucial in causality assessment and are a unique advantage to the literature as compared with observational data alone. In previous work, we have shown that modeling relational connections between concepts in the literature can be leveraged by machine learning methods to classify drug-reaction associations with state-of-the art performance on multiple reference standards[37,39,40]. Furthermore, these trained models can generalize to previously unseen drugs[39,40]. However, inferences drawn from machine learning models of the literature alone are unlikely to justify regulatory action in the absence of observational data to support them and the hypothesis that these models can be used to enhance SRS-based methods (e.g. DPMs) remains untested. In the current paper we address this issue by combining literature-derived distributed representations and statistical metrics derived from SRS data in a unified machine-learning approach for predicting ADRs.

2. Methods

The methods deployed in the current work leverage an extensive body of prior work that demonstrates that distributed vector representations (embeddings) of concept-relation-concept triplets (or semantic predications) extracted from the biomedical literature by NLP provide the means to infer helpful and harmful effects of drugs[37,39,40,42-45]. In this paper we extend this work by combining these representations with statistics from observational data, as input to a logistic regression classifier for predicting ADRs.

2.1 Distributed Representation of Literature-derived Relationships

SemRep is a widely-used NLP system that extracts concept-relation-concept triples – known as *semantic predications* – from the biomedical literature[46]. SemMedDB is a large repository of such extracted predications maintained and disseminated by the U.S. National Library of Medicine[47]. In order to generate literature-derived, relationship-based representations for biomedical concepts, SemMedDB version 31 without anaphora resolution was downloaded, containing 96,363,098 predications (i.e. concept-relationship-concept triplets) extracted from 28,429,379 citations encompassing data through June 30, 2018. For experiments done in this analysis, information was excluded based on the publication year to match the years for which information was available in the generation of the reference sets¹ (predications extracted from citations up to the end of 2012) and to more closely facilitate contemporaneous comparison to the previously-derived disproportionality measures² (Q3 2011 and 2015 June 2012). From these predications, 8000 dimensional binary vectors were generated with the Semantic Vectors Java package using the Embedding of Semantic Predications (ESP) architecture[38,48-50]. In this analysis, a stop list was not utilized. Training occurred over 5 epochs with concept and predication subsampling

Data Sharing:

Data and code used in this study is available at: <https://github.com/jusger/FAERLitComplement>.

²Further information on the disproportionality measures used in this study can be found in Section 2.3.

thresholds of 10^{-5} and 10^{-7} respectively, discarding concepts that occur more than 10^6 times, and five negatively sampled concepts of the same semantic type as the object per observed predication.

In brief, the objective with ESP is to generate embeddings for concepts that encode the semantics of those concepts. If two concepts relate to similar things in similar ways, this should be reflected by high vector similarity. For example, ibuprofen and diclofenac should have similar vector representations, as both are nonsteroidal anti-inflammatory drugs (NSAIDs) that treat pain. For binary vectors as used in this analysis, vector similarity is computed via the Non-negative Normalized Hamming Distance (NNHD), which measures the number of differing bits in two binary vectors, normalized by the dimensionality of the vectors. In order to encode compositionality, these binary vectors can be joined with exclusive OR (XOR, represented by $\otimes$), a self-invertible operation that yields a third “bound” vector product. For example, to generate a bound, semantic representation of ibuprofen from a single predication, ibuprofen-treats-pain, ‘treats’ and ‘pain’ vectors would be bound, such that the vector for ‘ibuprofen’ = (treats $\otimes$ pain). In practice, many predications exist for concepts, and the superposition of these predications can be embedded by adding bound vectors for a concept together. Continuing with the ibuprofen example, the vector for ‘ibuprofen’ would be equal to (treats $\otimes$ pain) + (inhibits $\otimes$ inflammation). Due to the self-invertible nature of the XOR operation, this embedded bound representation can be interrogated, whereby ibuprofen $\otimes$ treats would yield representations of ‘pain’ and ‘inflammation’ as the nearest neighbor vectors by NNHD. An interesting consequence of this procedure is that applying XOR to the bound product of two composite embeddings (such as ibuprofen $\otimes$ arthritis) results in a vector encoding the superposition of relational connections between them (e.g. treats $\otimes$ causes + inhibits $\otimes$ associated-with). Intuitively, the resulting vector products will be similar for pairs of entities that relate to one another in similar ways (relational similarity). They will also be similar for pairs of entities with similar properties (attributional similarity), because rofecoxib $\otimes$ myocardial_infarction will be similar to celecoxib $\otimes$ myocardial_infarction if these drugs have similar properties (such as ‘isa cox_2 inhibitor’). For further explication we refer the interested reader to Cohen et al 2017[38].

2.2 Reference Sets

To facilitate comparing performance across methods, the reference standard developed by Ryan et al in 2013 for the Observational Medical Outcomes Partnership was utilized (henceforth the OMOP set)[51]. This set comprises 165 positive and 234 negative controls of drug-health-outcome-of-interest (HOI) assertions across 4 HOIs: acute liver injury, acute myocardial infarction, acute kidney injury, and upper gastrointestinal bleeding. Additionally, a wide range of drug classes are in the set, including (but not limited to) NSAIDs, antibiotics, antivirals, and antipsychotics. Literature-derived vector representations trained on semantic predications extracted from published articles through the end of 2012 produced concept vectors that resolved 394 of the 399 total examples. That is to say, it was possible to identify a concept vector representation for both the drug and associated HOI in all but 5 pairs with a string match or lexical lookup (e.g. niacin – nicotinic acid). For disproportionality analysis, precalculated DPMs are reported for 380 of the examples by

Harpaz et al 2013 and we directly used these reported DPMs[52]. Empirical Bayesian Geometric Mean (EBGM), Proportional Reporting Ratio (PRR), Reporting Odds Ratio (ROR) and EBGM Non-Stratified metrics, with 95% confidence lower bounds, were extracted from the Harpaz paper to facilitate this analysis[52].

The reference standard developed in 2012 by Coloma et al for the Exploring and Understanding Adverse Drug Reactions by Integrative Data Mining initiative was also utilized (henceforth the EU-ADR set)[53]. This set comprises 44 positive and 50 negative controls across 10 HOIs: bullous eruptions; acute renal failure; anaphylactic shock; acute myocardial infarction; rhabdomyolysis; aplastic anaemia / pancytopenia; neutropenia / agranulocytosis; cardiac valve fibrosis; acute liver injury; and upper gastrointestinal bleeding. As in the OMOP set, a wide range of drug classes are represented: NSAIDs, antidepressants, antihypertensives, antibiotics and more. In literature-derived vector representations encompassing the same information through the end of 2012 as in the OMOP set, 93 of the 94 examples were resolved. For disproportionality measures, we extracted from Banda et al 2016, pre-calculated PRR and ROR values along with their 95% confidence interval upper and lower bounds for 91 of the 94 examples[54].

Since DPMs were available for 380 and 91 examples for the OMOP and EU-ADR reference sets respectively and since literature-derived vectors were available for every example for which DPMs were available, we constrained both reference sets to include only the examples for which DPMs could be evaluated.

2.3 Disproportionality Measures

For this research, the EBGM (stratified and unstratified), PRR, and ROR metrics were used for the OMOP set, and the PRR and ROR were used for the EU-ADR set³. With accompanying lower 95 percentile bounds for the OMOP metrics and upper and lower 95 percentile bounds for the EU-ADR metrics, each set comprises 8 (4 metrics plus 4 matching lower bounds) and 6 features respectively (2 metrics plus 2 matching lower bounds and 2 matching upper bounds). As previously mentioned, DPMs were derived from Harpaz et al for the OMOP reference set and Banda et al resource for the EU-ADR reference set[52,54]. Harpaz et al used FAERS data through Q3 2011, and deduplicated reports, corrected terminological errors, standardized and normalized drug names, and loaded the data into the Empirica Signal v7.3 system[55] to generate DPMs for the OMOP drug-event pairs. Similarly, Banda et al present a unified and normalized version of the FDA Adverse Reporting System (FAERS) through June 2015, and calculate DPMs for co-occurring concepts[54]. Notably, the Harpaz resource was developed in documenting the performance of DPMs on the OMOP reference set specifically, whereas the Banda resource was generated by calculating DPMs for all drug-level concepts (mapped to the RxNorm vocabulary) with all events (mapped to SNOMED-CT) that occurred in the FAERS system. To derive DPMs for the EU-ADR reference standard pairs from these pre-calculated tables, a direct string match or lexical lookup was manually performed in the same fashion as was done for the literature representations.

³The Banda resource does not report the EBGM measure.

2.4 Regularized Logistic Regression Analysis

Two model paradigms were evaluated, both utilizing L1 regularized logistic regression (LR). In addition to matching previously reported research, regularized logistic regression was chosen to enforce a sparse combination of features and to resist overfitting; with 8000+ feature columns and a limited number of training examples, unregularized models may be susceptible to overfitting and brittleness. In concatenated models, EBGM, PRR, and ROR disproportionality features were concatenated to the literature-derived binary vectors, resulting in a new vector of $8000 + 8$ or 6 dimensions (OMOP and EU-ADR, respectively), and used as the input to a L1 LR model. These models were evaluated by calculating the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) at various cutoffs of training examples with 5-fold stratified cross-validation (a learning curve). 5-fold cross-validation is first used to split the data into training and testing folds, and then n training samples are selected as a subset from the training set for training, where n is the training example threshold. For context, additional learning curves were generated for models using the literature alone (8000 dimension vectors) and disproportionality measures alone (8 or 6 dimension vectors). Learning curves were selected as performance modalities for these models on the basis that inputs (e.g. DPM and/or literature) may exhibit differential preference for training set sizes. Learning curves represent many ROC AUCs across a range of training set sizes, lending additional comparative context over a single train / test ratio for each of the modalities (DPM only, literature only, and combined).

Second, an ensemble model was developed by training an L1 LR model on the literature-derived vectors and the disproportionality measures separately (in contrast to the combined model, which trained a single LR model on the concatenation of those features). Each component LR model was trained in a stratified 5-fold cross-validation configuration, and then a weighted average of output predictions was calculated, such that the contribution of the literature towards the final prediction could be modulated. Once a final averaged prediction was calculated, examples were rank ordered and used to calculate ROC AUC metrics. For context, the baseline performance leveraging only the disproportionality measures was utilized. This evaluation focuses on how shifting contribution of data sources (as opposed to shifting the number of training examples) affects model performance, a scenario which might be realistically encountered if models were to see adoption as part of a pharmacovigilance system already leveraging DPMs.

An overview schematic of these models can be seen in Figure 1.

2.4 Computational Considerations

Except for the generation of literature vectors via the Semantic Vectors software package, all analysis was accomplished in a Python 3.7 environment via the Anaconda distribution[56]. For the machine learning analysis, the scikit-learn package was utilized, including learning curve and cross-validation helper functions[57]. Plotting was done utilizing the matplotlib plotting package[58]. A Github repository is available containing data files, accompanying Jupyter notebook[59], and Python utility code that can be used to recapitulate the results contained here (available at <https://github.com/jusger/FAERLitComplement>).

3. Results

Learning curves for the combined model, literature only, and disproportionality only modes can be seen in Figures 2 and 3. Across all models, incorporating information from both the disproportionality and literature data sources resulted in the highest performance, with average precision of >0.95 for training sizes of $\geq 65\%$ of the total data set, a nearly 20% increase over the best performing DPMs-based model. Compared to literature-based models, the combination model resulted in a modest but consistent $\approx 2\%$ absolute increase in mean performance. The literature-only model outperforms the DPMs-only models at all training thresholds tested for the OMOP reference set, with closest comparative performance observed when leveraging only a few training examples (the proximal aspect of the learning curve). For the EU-ADR reference set, the DPMs-only based model outperforms combined models until $\approx 25\%$ or more of the examples are used for training. In either case, both reference sets exhibit separation of DPMs-only based models and literature-only models at 50 training examples or more. Of note, the EU-ADR space is sampled at the same rate as the OMOP reference set (10 thresholds), but with fewer numbers of total examples. Consequently, the EU-ADR leverages fewer examples and each threshold contributes a smaller absolute addition of training examples (though equivalent relative contribution). Models are reported with 95% confidence intervals (± 1.96 times the standard error), calculated by standard error of the mean across 100 cross-validation runs. The EU-ADR results exhibit wider confidence bands and more overall variability than the OMOP set.

Next, the ensemble performance using five-fold cross-validation can be seen in Figures 4 and 5. The best performance of the thresholds tested was at near equal contributions of the outputs of both models (that is, both literature and disproportionality models were weighted equally; a true average). At a weighted contribution of 20% literature and 80% disproportionality, an overall performance uplift of $\approx 10\%$ over utilizing disproportionality-based models alone is observed for both reference sets. Overall ensemble model performance is comparable to the direct concatenation of feature vectors in combined models, with average precision values of ≈ 0.97 and ≈ 0.93 at a near equal ($\approx 50/50$) weighting threshold (OMOP and EU-ADR respectively). As with the learning curves, 95% confidence intervals are reported as calculated by standard error of the mean across 100 cross-validation runs, and the EU-ADR set again shows more variability and wider confidence bands.

4. Discussion

This study set out to examine the extent to which literature-derived distributed representations and statistical metrics derived from SRS data could be unified in a machine-learning approach for predicting ADRs. These results document improvement in model performance when incorporating literature and SRS information over using either alone, which strongly supports the hypothesis that these methods provide complementary perspectives pertinent to the evaluation of drug-HOI relationships. While these improvements were consistent across both reference standards, there were some differences noted, in that models trained on the EU-ADR exhibited increased variability and wider confidence intervals. These results may not be entirely unexpected, however, as the EU-

ADR set comprises more side-effects with fewer examples per side-effect than the OMOP set, and significantly fewer examples overall. In leveraging the literature, a greater representation of drugs in a given side-effect context seems especially important for model performance, likely as a result of the unsupervised nature of the embeddings simultaneously including information extraneous to a particular task (as these representations are trained to represent the entirety of a concept's connections, not just those pertinent to a given task). Fine-tuning this representation for a particular task, in this case ADR classification, requires some breadth of example space from which to learn, and in the EU-ADR reference standard, a limited example space likely limits performance of literature-based models. In contrast, DPMs should be relatively unperturbed by limited numbers of training examples, a finding supported in the comparatively flat performance in the learning curves. Of note, the combined model (which leverages both literature and DPM information) in the case of the EU-ADR does not match the performance of DPM only models when extremely few numbers of training examples are chosen, likely on account of the large parameter space offered by the literature vectors in this context. When more training examples are included, more fine-tuning is achieved through regularization of less task-specific information in the literature parameter space, which may overwhelm the DPM measures in extremely example sparse training regimens seen with the EU-ADR. Importantly, the DPM measures leveraged here should also be reference set agnostic, and the literature representations used have been previously shown to hold promise for generalization across reference standards and out of set examples[40]. Their combination provides two independent sources of information (literature and SRS) for models to draw from, both with generalization potential. Their combination results in better performance than either individual source alone, and exceeds previously reported performance on these sets.

These experiments show that disproportionality measures can be directly complemented by distributed representations of literature-derived semantic predications. This result is not unexpected, as previous multi-modal signal integration studies have shown increased performance when considering multiple data sources[23,28,60]. However, results here outperform previous multi-modal studies that draw from not only the literature and FAERS reports, but product labels, web logs, and/or electronic health records[23,60]. Although both reference sets presented differing sets of DPMs, an appreciable and significant increase was seen with both reference standards when these SRS-based features were presented. Moreover, both models could be incorporated into an ensemble configuration that produced consistently and significantly better cross-validation performance at every range of literature-contribution. To our knowledge, this ensemble configuration is a novel approach in this domain, with previous modalities only examining feature concatenated combined models. An ensemble approach, however, provides a functional and easy transition for integration into existing ranking methodologies, as literature information can be incorporated such that the vast majority of the signal used for classification comes from SRS data and still improve performance. To the extent that a regulatory entity wished to incorporate a controlled contribution of literature-derived information, the method provided here is effective, so long as probability estimates can be established for DPMs (as LR was utilized here).

This work does have some notable limitations. As the methods described here operate in a supervised manner, labeled example pairs are required for training and our models perform better when more training examples are available. Additionally, labeled examples used here are derived, in part, from information in the literature and positive pairs may have significant documentation in both the literature and SRS sources on account of the selection criterion (e.g. example pairs such as ibuprofen-gastrointestinal bleed), which screened on sources such as case reports, product labels, Tisdale review, and presence in an SRS database. This could potentially limit the generalization of these methods to emerging drug-HOI pairs that are not well documented. To simulate such a scenario, literature-derived models were deliberately deprived of any citation in which the drug-HOI pairs in the reference sets co-occurred; performance of literature models remained strong despite this severe constraint (see Supplementary Figures 1 and 2). In contrast, disproportionality ratios, which depend on co-occurrence of concepts, cannot function under such constraints, as co-occurrence of drug-HOI pairs is completely ablated. Consequently, for a drug-HOI connection which exists only latently, the literature-based methods here provide a more robust platform for detection. In either case, if a drug has extremely limited information, it can be difficult both for literature and disproportionality measures to generate sufficient signal, as there may not be any or very limited reports concerning the drug at all, or the reaction with the drug.

With these considerations in mind, several potential avenues exist to integrate these methods into the existing pharmacovigilance landscape. First, signals identified by disproportionality measures that also possess significant literature-derived signal could be fast-tracked for manual review, since such signals are less likely to be false positives. Conversely, existing drugs could be run against HOIs for which many training examples exist (such as in the generalization analysis of [40]) and fast-tracked for review when signal reaches a lower-than-otherwise minimum threshold for investigation, potentially mitigating false negatives and/or improving time-to-detection. For example, Evans et al used a minimum PRR of 2.0, a minimum number of cases of 3, and a Chi-square value of 4 to warrant further investigation[16]. If literature-derived models already suspected a plausible connection could exist, then perhaps a lowering of the PRR threshold could allow (earlier) detection. This study's results support these modalities, since combined models indicate the literature and DPMs present complementing views. Finally, either the combined or ensemble models presented here could be used to generate first-line rankings of drugs for further investigation. The ensemble model seems the easiest for transition, as the vast majority of weight could be given to the existing mode of using DPMs and modulated as further retrospective analysis on larger numbers of validated drug-HOI pairs warranted. In any of these scenarios, this work presents avenues to augment existing pharmacovigilance DPMs with literature information in adverse health condition contexts for which literature models can be trained on validated example pairs.

5. Conclusion

In this paper, we evaluated the utility of combining distributed representations of knowledge extracted from the biomedical literature with disproportionality metrics derived from adverse event report data in a unified machine learning model. Across two reference sets, results with the combined model were better than those obtained using either method alone,

providing strong support for the complementary nature of these approaches to post-marketing drug surveillance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding:

This work was supported by US National Library of Medicine grant (R01 LM011563).

Bibliography

1. National Center for Health Statistics. Health, United States, 2016: With Chartbook on Long-term Trends in Health [Internet]. Hyattsville, MD; 2017 Available from: <https://www.cdc.gov/nchs/data/health-us/2016.pdf>
2. Center for Disease Control and Prevention. National Hospital Ambulatory Medical Care Survey: 2011 Outpatient Department Summary Tables [Internet]. 2012 Available from: https://www.cdc.gov/nchs/data/ahcd/nhamcs_outpatient/2011_opd_web_tables.pdf
3. Hing E, Rui P, Palso K. National Ambulatory Medical Care Survey: 2013 State and National Summary Tables [Internet]. 2014 Available from: http://www.cdc.gov/nchs/ahcd/ahcd_products.htm
4. Rui P, Kang K, Albert M. National Hospital Ambulatory Medical Care Survey: 2013 Emergency Department Summary Tables [Internet]. 2014 Available from: http://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2013_ed_web_tables.pdf.
5. 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;14:125. [PubMed: 24620750]
6. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA. 1998;279:1200–5. [PubMed: 9555760]
7. Bourgeois FT, Shannon MW, Valim C, Mandl KD. Adverse Drug Events in the Outpatient Setting: An 11-Year National Analysis. Pharmacoepidemiol Drug Saf. 2010;19:901–10. [PubMed: 20623513]
8. Watanabe JH, McInnis T, Hirsch JD. Cost of Prescription Drug-Related Morbidity and Mortality. Ann Pharmacother. 2018;1060028018765159.
9. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. JAMA. 1997;277:301–6. [PubMed: 9002492]
10. Downing NS, Shah ND, Aminawung JA, Pease AM, Zeitoun J-D, Krumholz HM, et al. Postmarket Safety Events Among Novel Therapeutics Approved by the US Food and Drug Administration Between 2001 and 2010. JAMA. 2017;317:1854–63. [PubMed: 28492899]
11. World Health Organization. The importance of pharmacovigilance. 2002; Available from: <http://apps.who.int/iris/bitstream/10665/42493/1/a75646.pdf>
12. Pariente A, Gregoire F, Fourrier-Reglat A, Haramburu F, Moore N. Impact of safety alerts on measures of disproportionality in spontaneous reporting databases: the notoriety bias. Drug Saf. 2007;30:891–8. [PubMed: 17867726]
13. Stephenson WP, Hauben M. Data mining for signals in spontaneous reporting databases: proceed with caution. Pharmacoepidemiol Drug Saf. 2007;16:359–65. [PubMed: 17019675]
14. Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, et al. Bayesian methods in pharmacovigilance. Oxf Univ Press 2011;23:29.
15. Center for Drug Evaluation and Research. FDA Adverse Events Reporting System (FAERS) - Reports Received and Reports Entered into FAERS by Year [Internet]. [cited 2017 Jul 16].

Available from: <https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070434.htm>

16. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf.* 2001;10:483–6. [PubMed: 11828828]
17. Li Y, Ryan PB, Wei Y, Friedman C. A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. *Drug Saf.* 2015;38:895–908. [PubMed: 26153397]
18. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf.* 2009;18:427–436. [PubMed: 19358225]
19. Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol Drug Saf.* 2004;13:519–23. [PubMed: 15317031]
20. Meyboom RH, Hekster YA, Egberts AC, Gribnau FW, Edwards IR. Causal or casual? The role of causality assessment in pharmacovigilance. *Drug Saf.* 1997;17:374–89. [PubMed: 9429837]
21. Naidu RP. Causality assessment: A brief insight into practices in pharmaceutical industry. *Perspect Clin Res.* 2013;4:233–6. [PubMed: 24312892]
22. Center for Drug Evaluation and Research. Questions and Answers on FDA's Adverse Event Reporting System (FAERS) [Internet]. 2016 [cited 2017 Jul 19]. Available from: <https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugs/effects/>
23. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform.* 2017;66:72–81. [PubMed: 27993747]
24. Winnenbun R, Sorbello A, Ripple A, Harpaz R, Tonning J, Szarfman A, et al. Leveraging MEDLINE indexing for pharmacovigilance-Inherent limitations and mitigation strategies. *J Biomed Inform.* 2015;100:425–435.
25. Collaborative TKB workgroup of the OHDS and I (OHDSI). Large-scale adverse effects related to treatment evidence standardization (LAERTES): an open scalable system for linking pharmacovigilance evidence sources with clinical data. *J Biomed Semant.* 2017;8:11.
26. Winnenbun R, Shah NH. Generalized enrichment analysis improves the detection of adverse drug events from the biomedical literature. *BMC Bioinformatics.* 2016;17:250. [PubMed: 27333889]
27. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Saf.* 2014;37:777–790. [PubMed: 25151493]
28. Xu R, Wang Q. Large-scale combining signals from both biomedical literature and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection. *BMC Bioinformatics.* 2014;15:17. [PubMed: 24428898]
29. Ahlers CB, Hristovski D, Kilicoglu H, Rindflesch TC. Using the Literature-Based Discovery Paradigm to Investigate Drug Mechanisms. *AMIA Annu Symp Proc.* 2007;2007:6–10.
30. Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. 1998; Available from: <https://deepblue.lib.umich.edu/handle/2027.42/34255>
31. Henry S, McInnes BT. Literature Based Discovery: Models, methods, and trends. *J Biomed Inform.* 2017;74:20–32. [PubMed: 28838802]
32. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. *AMIA.* 2006.
33. Smalheiser NR. Literature-based discovery: Beyond the ABCs. *J Am Soc Inf Sci Technol.* 63:218–24.
34. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med.* 1986;30:7–18. [PubMed: 3797213]
35. Swanson DR, Smalheiser NR. Undiscovered Public Knowledge: A Ten-Year Update. *KDD* [Internet]. 1996 p. 295–298. Available from: <https://ocs.aaai.org/Papers/KDD/1996/KDD96-051.pdf>
36. Hristovski D, Burgun-Parenthoine A, Avillach P, Rindflesch TC. Towards using literature-based discovery to explain drug adverse effects. 24th Int Conf Eur Fed Med Inform Qual Life Qual Inf

- MIE [Internet]. 2012 Available from: <http://person.hst.aau.dk/ska/mie2012/AllPresentations/422.pdf>
37. Shang N, Xu H, Rindflesch TC, Cohen T. Identifying plausible adverse drug reactions using knowledge extracted from the literature. *J Biomed Inform.* 2014;52:293–310. [PubMed: 25046831]
 38. Cohen T, Widdows D. Embedding of semantic predications. *J Biomed Inform.* 2017;68:150–166. [PubMed: 28284761]
 39. Mower J, Subramanian D, Shang N, Cohen T. Classification-by-Analogy: Using Vector Representations of Implicit Relationships to Identify Plausibly Causal Drug/Side-effect Relationships. *AMIA Annu Symp Proc.* 2017;2016:1940–9. [PubMed: 28269953]
 40. Mower J, Subramanian D, Cohen T. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. *J Am Med Inform Assoc [Internet].* 2018 [cited 2018 Sep 26]; Available from: <https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocy077/5052182>
 41. Mower Justin. *Compositional Relation-based Learning (CoRL): A General-Purpose Method to Leverage Literature-Derived Relationships Applied to Pharmacovigilance.* [Houston]: Baylor College of Medicine; 2018.
 42. Cohen T, Widdows D, Schvaneveldt RW, Davies P, Rindflesch TC. Discovering discovery patterns with predication-based semantic indexing. *J Biomed Inform.* 2012;45:1049–1065. [PubMed: 22841748]
 43. Cohen T, Widdows D, Schvaneveldt R, Rindflesch TC. Finding Schizophrenia's Prozac Emergent Relational Similarity in Predication Space. *Quantum Interact [Internet].* Springer, Berlin, Heidelberg; 2011 [cited 2017 Oct 12]. p. 48–59. Available from: https://link.springer.com/chapter/10.1007/978-3-642-24971-6_6
 44. Cohen T, Widdows D, De Vine L, Schvaneveldt R, Rindflesch TC. Many paths lead to discovery: analogical retrieval of cancer therapies *Int Symp Quantum Interact.* Springer; 2012 p. 90–101.
 45. Cohen T, Widdows D, Stephan C, Zinner R, Kim J, Rindflesch T, et al. Predicting High-Throughput Screening Results With Scalable Literature-Based Discovery Methods. *CPT Pharmacomet Syst Pharmacol.* 2014;3:1–9.
 46. 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;36:462–477. [PubMed: 14759819]
 47. Kilicoglu H, Shin D, Fiszman M, Rosembat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinforma Oxf Engl.* 2012;28:3158–60.
 48. Widdows D, Ferraro K. *Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application.* LREC. Citeseer; 2008.
 49. Widdows D, Cohen T. The semantic vectors package: New algorithms and public tools for distributional semantics. *2010 IEEE Fourth Int Conf Semantic Comput IEEE;* 2010 p. 9–15.
 50. *Semantic Vectors [Internet].* 2019 [cited 2019 Jun 10]. Available from: <https://github.com/semanticvectors/semanticvectors>
 51. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. *Drug Saf.* 2013;36:33–47.
 52. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System. *Clin Pharmacol Ther.* 2013;93:539–546. [PubMed: 23571771]
 53. Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, Pariente A, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. *Drug Saf.* 2013;36:13–23. [PubMed: 23315292]
 54. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data.* 2016;3.
 55. Oracle Health Sciences. *Empirica Signal [Internet].* Oracle; Available from: <http://www.oracle.com/us/products/applications/health-sciences/safety/empirica-signal/index.html>
 56. Continuum Analytics. *Anaconda Python Distribution [Internet].* Available from: <https://www.anaconda.com/>

57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12:2825–30.
58. Hunter JD. Matplotlib: A 2D graphics environment *Comput Sci Eng.* 2007;9:90–95.
59. Kluyver T, Ragan-Kelley B, Perez F, Granger BE, Bussonnier M, Frederic J, et al. Jupyter Notebooks-a publishing format for reproducible computational workflows *ELPUB.* 2016 p. 87–90.
60. Harpaz R, DuMouchel W, Schuemie M, Bodenreider O, Friedman C, Horvitz E, et al. Toward Multimodal Signal Detection of Adverse Drug Reactions. *J Biomed Inform.* 2017;

Key Points

Embeddings derived from relational connections in the biomedical literature can be used alongside observational data for enhanced drug safety monitoring.

Developed models can be tuned to weight the relative contribution of literature or observational data, allowing end users to rely on either source of information at their discretion.

Even a modest contribution from the literature (20% relative weight) results in a 10-15% increase in performance relative to using observational data alone.

Fig. 1.

Experimental schematic overview. In literature only, DPM only, and combined models, a single logistic regression model is cross-validated (evaluated on multiple non-overlapping train/test splits) for each respective input data configuration. For combined models, input data is the concatenation of binary vectors from the literature and DPMs from SRS. In contrast, the ensemble method trains two logistic regression models - one on each source of information independent of the other. The probability estimates from each cross-validated logistic regressor for each drug-HOI example are then averaged together, with the weight of that average favoring either source of information or being a true average. In all cases, the resulting performance on the testing sets over 100 cross-validation runs are plotted for performance comparison.

Fig. 2.

A learning curve comparison of ROC AUC performance at varying levels of training examples drawn from the OMOP reference set. Curves are drawn with 95% confidence interval bands calculated via standard error of the mean across 100 random stratified 5-fold cross-validation runs. The best performing model across all ranges is the combined model, with $\approx 20\%$ increase over the DPM only model and $\approx 2\%$ increase over the closely following literature only model. For this reference set, DPM model performance always trails that of the literature only or combined models. All models appear to plateau in performance after approximately 150 examples.

Fig. 3.

A learning curve comparison of ROC AUC performance at varying levels of training examples drawn from the EU-ADR reference set. Curves are drawn with 95% confidence interval bands calculated via standard error of the mean over 100 random stratified 5-fold cross-validation runs. The combined model outperforms the literature only model across the entire range of training example thresholds, but the DPMs only model outperforms both until approximately 28 and 45 training examples, respectively, for the combined and literature only models. DPM only models increase slightly and quickly plateau, while literature inclusive models show a steady increase in performance with increasing numbers of examples. At the distal end of the curve, the combined model represents a $\approx 12\%$ increase over DPM models alone and a $\approx 5\%$ increase over literature only models.

Fig. 4.

Weighted ensemble model ROC AUC for the OMOP reference set. Each point is the mean performance across 100 random stratified 5-fold cross-validation runs at the given literature contribution (e.g. where each prediction is calculated as the literature model's output probability times the given contribution plus the DPM model's probability times 1-minus the given contribution), with 95% confidence interval bands calculated via standard error of the mean. At all weightings sampled, including literature information results in a significant increase in mean ROC AUC compared to the baseline DPM method alone. Best performance occurs at an approximately equal weighting of information sources, with a $\approx 16\%$ performance improvement over the DPM baseline.

Fig. 5.

Weighted ensemble model ROC AUC for the EU-ADR reference set. Each point is the mean performance across 100 random stratified 5-fold cross-validation runs at the given literature contribution (e.g. where each prediction is calculated as the literature model's output probability times the given contribution plus the DPM model's probability times 1-minus the given contribution), with 95% confidence interval bands calculated via standard error of the mean. At all weightings sampled, including literature information results in a significant increase in mean ROC AUC compared to the baseline DPM method alone. Best performance is achieved at an approximately equal contribution of information sources, with a $\approx 16\%$ performance improvement over the DPM baseline.