JAMIA Open, 2(3), 2019, 378–385
doi: 10.1093/jamiaopen/ooz022
Advance Access Publication Date: 1 July 2019
Research and Applications

Research and Applications

Inferring new relations between medical entities using literature curated term co-occurrences

Adam Spiro, Jonatan Fernández García, and Chen Yanover

Machine Learning for Healthcare and Life Sciences, Department of Health Informatics, IBM Research, Haifa, Israel

Corresponding Author: Adam Spiro, PhD, Machine Learning for Healthcare and Life Sciences, Department of Health Informatics, IBM Research, University of Haifa Campus, Mount Carmel, Haifa 3498825, Israel (adam.spiro@ibm.com)

Received 15 April 2019; Revised 5 June 2019; Editorial Decision 7 June 2019; Accepted 8 June 2019

ABSTRACT

Objectives: Identifying new relations between medical entities, such as drugs, diseases, and side effects, is typically a resource-intensive task, involving experimentation and clinical trials. The increased availability of related data and curated knowledge enables a computational approach to this task, notably by training models to predict likely relations. Such models rely on meaningful representations of the medical entities being studied. We propose a generic features vector representation that leverages co-occurrences of medical terms, linked with PubMed citations.

Materials and Methods: We demonstrate the usefulness of the proposed representation by inferring two types of relations: a drug causes a side effect and a drug treats an indication. To predict these relations and assess their effectiveness, we applied 2 modeling approaches: multi-task modeling using neural networks and single-task modeling based on gradient boosting machines and logistic regression.

Results: These trained models, which predict either side effects or indications, obtained significantly better results than baseline models that use a single direct co-occurrence feature. The results demonstrate the advantage of a comprehensive representation.

Discussion: Selecting the appropriate representation has an immense impact on the predictive performance of machine learning models. Our proposed representation is powerful, as it spans multiple medical domains and can be used to predict a wide range of relation types.

Conclusion: The discovery of new relations between various medical entities can be translated into meaningful insights, for example, related to drug development or disease understanding. Our representation of medical entities can be used to train models that predict such relations, thus accelerating healthcare-related discoveries.

Key words: machine learning, medical informatics, MeSH headings, literature-based discovery, adverse drug reaction, drug repositioning

BACKGROUND AND SIGNIFICANCE

Medical knowledge can be expressed using semantic relations between entities. These entities may include drugs, diseases, side effects, and proteins, whereas semantic relations may include "causes" (a drug causes a side effect), "indicated" (a drug is indicated for a disease), and "targets" (a drug targets a protein). The discovery of new relations can be translated into meaningful insights. For example, a new "treats" relation between an existing drug and disease entities denotes a potential new indication for the drug. A new "causes" relation between drug and side effect entities, means it might have a side effect that is not yet known. The vast increase in publicly available medical-related data, together with groundbreaking developments in machine learning technologies and processing power, have created an enormous opportunity for computational-based predictions of new relations between medical entities. Using machine learning to predict new relations—for example, a drug's

side effects or indications—has many benefits, including being much more time- and cost-effective than classical methods based on observational and lab experiments.

Addressing a prediction task using machine learning techniques requires both a predictive model and a numerical representation of the input. This numerical representation should ideally capture relevant semantics. Numerous papers published recently focus on predicting new relations between medical entities using various data representations and predictive models. To predict drug side effects, also termed adverse drug reactions (ADRs), earlier works used a representation based on drug chemical structure.[1,2] Other works integrated additional sources such as target proteins of drugs,[3–6] medical literature,[3,7–9] social media,[8,10] and electronic health records.[11] Works predicting new indications for existing drugs, also termed drug repositioning,[12] have used gene expression data[13] or drug side effects[14–16] to represent the drugs. Most of the works on prediction models use classical approaches, such as logistic regression (LR)[17] and kernel-based methods.[3,10] Others have employed additional models such as hidden Markov models[18] and recommender systems.[19] Several recent works use the increasingly popular neural network (NN) models, which have shown promising results.[8,9,11,20]

In this work, we develop a new generic representation scheme that can be utilized by a machine learning model and show its usability by focusing on the 2 examples described above: predicting drug indications and ADRs. We use the co-occurrence of Medical Subject Headings (MeSH) terms to generate numeric representations of each drug. MeSH terms are descriptors that were manually assigned by experts to each article published in PubMed. These currently include over 28 000 terms from 16 categories, including diseases, chemicals and drugs, anatomy, phenomena and processes, and others, with a hierarchical multilayer structure of relevant semantics.

Conceptually, this work is closely related to literature-based discovery (LBD) methods, which seek to infer new knowledge from existing literature in an automated way (see Ref.[21] for review). Typically, these methods use text mining tools to extract terms or concepts, then interpret co-occurrences, potentially semantically constrained, of such entities as relations. To discover new relations, LBD methods either explicitly apply the transitivity notion of Swanson's ABC co-occurrence model,[22] where "A relates to B" and "B relates to C" implies "A relates to C"; or searches for entities with close (under some distance measure) co-occurrence profiles and suggest that these have similar relations. For a review on the different methods used to extract drug safety information from textual resources, see Ref.[23]

MeSH term co-occurrences have been previously used primarily for tasks related to text mining, such as PubMed document search or author name disambiguation.[24–27] MeSH terms have also been used for gene–disease association[28] and drug–drug interaction.[29] Several works have also leveraged MeSH term data for LBD of drug–ADR associations. Winnenburg and Shah[30] used generalized enrichment analysis to examine associations between drugs and ADRs at multiple levels of granularity. Shetty and Dalal[31] developed a statistical document classifier for detecting ADRs based on MeSH terms by filtering irrelevant articles. In another work, Avillach et al.[32] selected specific MeSH terms for 10 drugs and 6 ADRs, and showed that using direct MeSH co-occurrences between these terms can differentiate between true and false drug–ADR relationships. We note that most of these works used small manually selected subsets of terms and simple distance measures or clustering techniques, rather than a generic representation and machine learning framework, as proposed here.

Our approach expands and generalizes the idea of Avillach et al.[32] (referred to hereinafter as the *baseline model*) by exploiting the fact that MeSH term co-occurrences span more than 28 000 terms. We posit that the MeSH co-occurrences-based representation of each drug encompasses the complex relationship of the drug with numerous terms from different domains, rather than just using the direct relationship between the drug and the specific side effects terms. We demonstrate the usefulness of the proposed general representation by training models to predict ADRs or indications using 2 modeling approaches. The first approach uses either gradient boosting machines (GBM) or LR to train an independent predictor for each task (ie, either a specific ADR or indication). The second approach implements a multi-task NN that trains a single model over all tasks (all ADRs or indications). Both approaches perform significantly better than the baseline model. Furthermore, we show that a multi-task NN learner performs slightly better than single-ADR classifiers in the ADR prediction task, suggesting limited but consistent inter-ADR information. The simpler single model LR approach performs slightly better in the drug indications prediction task. This may be due to sparsity and limited available data.

Suggesting new relations between medical entities can be translated into important insights. Our representation provides a more holistic view of an entity's characteristics, which can accelerate the pace of such discoveries.

MATERIALS AND METHODS

MeSH term co-occurrences

The MEDLINE Co-Occurrences files (MRCOC, available at https://ii.nlm.nih.gov/MRCOC.shtml) contain the summary of all MeSH terms that occur together in the citations available from PubMed. PubMed currently comprises more than 29 million citations of biomedical literature from MEDLINE, life science journals, and online books. We downloaded the 2019 summary file and preprocessed it to accumulate data from all the past years covering both major and non-major topics. We extracted a total of 28 320 MeSH terms, 5 095 060 pairs of MeSH terms, and 117 654 465 co-occurrences, which we used to generate the numeric representation.

Extraction of drug indications and ADRs

The Side Effect Resource, also known as the SIDER database,[33] contains public information on drugs, their ADRs, and indications. The database (Version 4.1) contains information on 1430 drugs and 5868 ADRs/indications. In our analysis, we removed ADRs/indications that appear in fewer than 10 drugs, which reduced the total number of ADRs to 1657 and indications to 424. We thus generated 2 Boolean matrices. The first was for the drug–ADR relationship of size 1430 by 1657, which contains 150 412 (6.35%) positive and 2 219 098 (93.65%) negative relationships. The second was for the drug-indications relationship of size 1430 by 424, which contains 10 416 (1.72%) positive and 595 904 (98.28%) negative relationships.

In addition, we used the observational medical outcomes partnership (OMOP) reference dataset,[34] which is a smaller ADR reference standard. It includes 4 ADRs and 182 drugs and contains both positive and negative associations between drugs and ADRs.

Mapping drug IDs to MeSH terms

In the SIDER database, drugs are represented using the PubChem compound identifier (CID).[35] The MeSH co-occurrences data, on the other hand, uses MeSH terms (or MeSH unique IDs) to represent

medical terms, specifically the drug terms. Since our drug feature space uses MeSH terms, we had to map CIDs to MeSH terms to generate the drug representation. To generate a direct mapping between these 2 types of terms, we devised several automatic and semi-automatic heuristics. First, we extracted the drug names from the MeSH Descriptor Files in ASCII, available on the NLM website (https://www.nlm.nih.gov/databases/download/mesh.html). We then extracted the CID compound names from PubChem API and used regular expressions to match the compound names to the MeSH descriptors. Some drugs were mapped to the MeSH Supplementary Concept Records. Although these are not part of the co-occurrences data, they do contain a reference to the nearest MeSH term descriptor, which we used for the actual mapping. This procedure successfully mapped about 90% of the terms. The rest of the drug names were manually curated and mapped to MeSH terms by finding the nearest descriptors using the ATC codes, and the links from SIDER to the PubChem and STITCH[36] databases. Some drug names were mapped to multiple MeSH descriptors, for which the co-occurrences were summed. All 1430 drug names were mapped to one or several MeSH descriptors.

The mapping between the drugs in the OMOP dataset and the MeSH descriptors was done by either direct name matching or manually, as described above.

Generation of drug representation

We mapped each drug D_i $(i = 1, \ldots, 1430)$ to a specific MeSH term for which we generated a numeric representation, used for the learning process. The representation of D_i is a feature vector $V_i = \{V_{ij}\}_{j=1}^{28\,320}$, where $V_{ij} = \text{Co–occur}(D_i, T_j)$ is the number of co-occurrences of drug D_i and MeSH term T_j. Each drug is thus represented by a feature vector of length 28 320 such that the dimensions of the input matrix is 1430 by 28 320 containing 31.75% non-zero elements. The drug representation actually holds the co-occurrence information between the drug and all types of MeSH terms, including other drugs, diseases, ADRs, symptoms, anatomical parts, and biological processes; this offers a more complete picture of the complex semantic relationship between drugs and other medical terms. It is important to note that replacing drugs with a different term group (eg, diseases, symptoms, or therapeutics), generates a corresponding representation that can be used for other generic prediction tasks.

Prediction methods

Term-frequency normalization

The drug representation includes co-occurrence frequencies between each drug and all other MeSH terms. These numbers have to be normalized to account for the variability in the total number of drug occurrences. We implemented 3 normalization methods:[37] maximum term-frequency normalization (Max-TF), in which each coordinate in the representation vector is divided by the maximum value of that vector; Log + Max-TF, in which the logarithm of the terms count is taken followed by Max-TF; and TF–inverse document frequency, which is commonly used in text mining and information retrieval for term normalization.

Machine learning models

To examine the prediction performance of the drug representation on the 2 tasks, we used 3 types of models:

1. Multilayer fully connected NN architecture, implemented using PyTorch.[38] This model gets the drug representation vector of size 28 320 as input. It outputs a vector of size 1657 for the ADRs task and 424 for the indications task, corresponding to the full list of possible ADRs/indications. Using a Sigmoid function for the output layer, we get a probability-like number for each output term representing the strength of its relation to the input drug.
2. GBM, implemented using the LightGBM package.[39] This nonlinear ensemble classifier method uses sequential decision trees, which are considered "weak" classifiers. In each iteration, an additional decision tree is added to improve the prediction obtained so far by the previous trees.
3. LR, implemented using Scikit-learn.[40]

The NN model is a multi-task learning model that performs the prediction for all ADRs or indications simultaneously.[41] This enables it to take advantage of possible interactions between different output variables in its learning process. The latter 2 methods are used to predict only one output term at a time; so for each ADR/indication we trained a different model. We applied probability calibration for the different models,[42] but because the obtained performance was significantly reduced we report below the results without calibration.

Baseline models

We considered 2 baseline models. The first model outputs, for each drug and ADR or indication, the normalized co-occurrence number (we reported results using the Log + Max-TF normalization described above, as it gave the best results among the 3 normalization methods). Similar to 32, we then used these numbers for the classification tasks. The second baseline model is based on disproportionality analysis where the proportional reporting ratio score was calculated as described, for example, in Montastruc et al.[43] For clarity, we report only the results for the best model among the 2 baseline models and note that their performance is typically comparable.

Hyperparameter tuning and model evaluation

We randomly divided the list of drugs into 5 groups and performed a 5-fold cross-validation where 3 folds were used for training, 1 for validation, and 1 for testing. We report the results accumulated over the test groups.

For each of the 3 machine learning models, we used the validation folds to tune and optimize the parameters. We reported the results on the test folds using the best parameters obtained from the validation folds. For all models, we optimized over the 3 term-frequency normalization methods mentioned above, as well as the L2 regularization parameter. For the NN model we also optimized the number and size of the hidden layers. Mean-squared error was used as the loss function. See the Supplementary Information for the final selection of model parameters.

We optimized the models using the precision-recall area under the curve (PRAUC) as a performance measurement. The reported results of the 3 models use the hyperparameters that achieved the highest PRAUC scores on the validation folds. See the "Discussion" section for the rationale behind using the PRAUC score as the performance measurement. When reporting PRAUC scores we also report standard deviation calculated over the 5-folds. Micro- and macro-averaging yielded similar results thus we report only the macro-averaging.

To capture another aspect of the model performance we calculated the top K ranking results. As opposed to PRAUC, ranking does not depend on global thresholds, rather it uses the top sorted scores. We generated 2 types of top rankings; the first relates to each drug separately and counts how many of the top K predicted ADRs/indications are labeled positive. The second relates to each ADR/indication term separately and counts how many of the top K scores for this term belong to drugs that have a positive relationship with that term.

RESULTS

We hypothesized that repeated references to medical terms alongside drug names in scientific literature correspond to drug characteristics. Therefore, we propose to leverage Medical Subject Headings (MeSH) term co-occurrences as a multifaceted representation of drugs; statistical analysis of the suggested representation is provided below. To demonstrate its usability, we trained models to predict ADRs as well as indications from drug MeSH term co-occurrences, and compared the predictive power of single- and multi-task models.

MeSH term representation

Overall, the MeSH vocabulary includes 28 320 terms, organized in a multi-level hierarchy. Table 1 illustrates the per-category term distribution. Notably, 32.7% of MeSH terms are classified as "Chemicals and Drugs" and 15.3% as "Diseases" (note that terms can belong to multiple categories). Table 2 illustrates the drugs subcategory distribution (again, terms can belong to multiple categories). Examining the drug representation, we see that most drugs are classified as either "Organic Chemicals" (38.3%) or "Heterocyclic Compounds" (29.1%). Overall, 91% of drugs co-appear with 1000 or more terms and 40% coincide with more than 5000 terms [median number of terms 3800, interquartile range (IQR) 1969–7837; Figure 1, left, shows the entire distribution]. Conversely, almost 6000 MeSH terms co-appear with $\leq$50 drugs and only 8% co-occur with most drugs (median number of drugs: 192, IQR 65–413; Figure 1, right, shows the entire distribution). Examples of MeSH terms that co-appear with many drugs include some generic terms such as "Humans" and "Rats", but most are drug-related terms such as "Dose–Response Relationship, Drug", "Treatment Outcome", and "Drug Therapy, Combination".

Prediction models

To demonstrate the usefulness of the proposed representation in inferring new relations between medical entities, we focus here on predicting relations of 2 types: (1) a drug causes an ADR and (2) a drug treats an indication. The filtered SIDER data (see "Materials and Methods" section for details) includes 1657 ADRs with a median of 34 drugs causing an ADR (IQR 17–92); and only 424 indications, each shared by a median of 18 drugs (IQR 13–27).

We trained 3 types of classification models: per-task LR and GBM applied, separately, for each ADR and indication; and a multi-task NN applied, collectively, on all ADRs or indications. We report the performance of these models using 3 measures: the average precision within the top-K drugs, over all ADR or indication prediction models; the PRAUC (see "Discussion" section for the rationale behind this choice); and the average precision within the top-K ADRs or indications, over all drugs. The former measure assesses each model separately. The latter 2 measures combine predictions across multiple single-task models; thus, they are potentially more sensitive to cross-model lack of calibration. We also report a baseline performance, as described in the "Materials and Methods" section.

We first assessed single-task performance, measured as the mean precision within the top K-drugs, for each relation type (Figure 2). The performance of all 3 models for both relation types is significantly higher than the baseline model (see Supplementary Figure S1), demonstrating the utility of our representation. It is interesting to note for both ADR and indication predictions that the performance of LR only slightly decreases across all values of K, while GBM and NN performance deteriorate more significantly with K. Consequently, GBM and NN, which initially outperform LR, obtain reduced precision on higher values of K.

Next, we evaluated model performance across all tasks. Figure 3 depicts the precision recall curves of the 3 trained model types in predicting ADRs and indications. Among the 2 single-task algorithms, LR consistently performs better than GBM. Both algorithms outperform the multi-task NN in predicting indications but not ADRs. Focusing on the high-precision range, LR and NN obtain similar performance, while GBM lags behind. Notably, the baseline model (see "Materials and Methods" section) achieved a PRAUC score of 0.12 ± 0.01 in the ADR task and 0.20 ± 0.07 in the indication task, significantly lower than the 3 machine learning models (see Figure 3).

Finally, we zoom in on the top-K predicted ADRs and indications for each drug (Figure 4). Consistent with model performance in the high-precision range (Figure 3, insets), the multi-ADR NN model slightly outperforms the 2 types of single-ADR models and LR obtains higher average precision than GBM. LR is somewhat higher in the top-ranked indications per drug. Again, all models significantly outperform the baseline (see Supplementary Figure S1).

In summary, none of the 3 models significantly outperform the other 2, and selecting the best model depends on the goals of the task. The NN model performs slightly better in the high precision range in both ADR and indication prediction tasks (Figure 3) and in ranking ADRs for each drug (Figure 4, left). However, the LR model has a higher PRAUC value in the indication task and favorable and more stable results most top-ranking comparisons (Figure 2, Figure 3, right, and Figure 4, right), especially for higher K values.

To further examine the usability of MeSH co-occurrences data for predicting ADRs and to demonstrate the advantage of using machine learning methods for that task, we used the OMOP dataset,[34] which includes data on 4 ADRs as described in Table 3. Since there are only 4 ADRs and little overlap between the different ADR-drug data we ran only the single-task models LR and GBM and not the multi-task NN model. As shown in Table 3 both these models outperform the baseline performance.

Partial representation performance

To explore the contribution of the different feature types (see Table 1) we used the SIDER database and calculated the overall PRAUC scores using different features subsets as shown in Table 4. Note that we divided the diseases-related MeSH terms into 2 groups: signs and symptoms (consisting of terms with the MeSH prefix code of C23.888) and all other disease terms. The reason is that the signs and symptoms MeSH terms contain most of SIDER's ADRs and indication terms corresponding to direct co-occurrences between features and labels. We calculated the results in Table 4 using the LR model. Results for the GBM and NN models show similar trends and thus are not reported. It can be seen that using subsets of the features achieve similar results to using all features (except for signs and symptoms, which includes a relatively small number of features), indicating redundancy in the feature space.

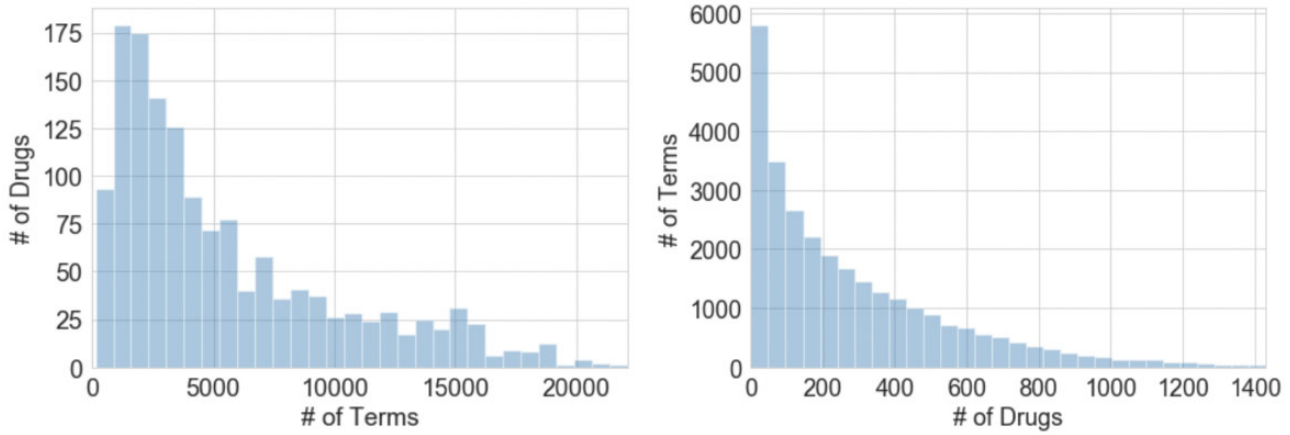

Figure 1. Drug and MeSH term co-occurrences; the number of drugs that co-appear with any number of MeSH terms (left) as well as the number of MeSH terms that coincide with different number of drugs (right). MeSH: Medical Subject Headings.

Table 1. MeSH term count per category

Category	# Terms	% Terms
Chemicals and drugs	10 086	32.7
Diseases	4711	15.3
Organisms	3678	11.9
Analytical, diagnostic and therapeutic techniques, and equipment	2885	9.4
Phenomena and processes	2227	7.2
Anatomy	1802	5.8
Healthcare	1687	5.5
Psychiatry and psychology	1042	3.4
Anthropology, education, sociology, and social phenomena	573	1.9
Technology, industry, and agriculture	565	1.8
Disciplines and occupations	403	1.3
Geographicals	385	1.2
Information science	338	1.1
Named groups	260	0.8
Humanities	171	0.6

Note. The full tree structure can be found in the NLM MeSH website (https://meshb.nlm.nih.gov/treeView).

Abbreviation: MeSH: Medical Subject Headings.

Table 2. Drug count per MeSH term sub-category

Sub-category	# Drugs	% Drugs
Organic chemicals	1228	38.3
Heterocyclic compounds	933	29.1
Polycyclic compounds	257	8.0
Amino acids, peptides, and proteins	245	7.6
Inorganic chemicals	164	5.1
Hormones, hormone substitutes, and hormone antagonists	84	2.6
Nucleic acids, nucleotides, and nucleosides	71	2.2
Carbohydrates	67	2.1
Lipids	64	2.0
Biological factors	40	1.2
Pharmaceutical preparations	20	0.6

Note. Categories with less than 20 drugs are omitted.

Abbreviation: MeSH: Medical Subject Headings.

DISCUSSION

We presented a drug representation generated using MeSH co-occurrence data and showed its potential by using it to predict both ADRs and drug indications without any prior drug information. The representation spans many term categories, as defined by the MeSH descriptor hierarchy, which provides a comprehensive picture of the relationship between all available MeSH terms. MeSH indexing is the task of manually assigning relevant MeSH terms to biomedical literature. Although the indexing procedure is currently relatively slow and expensive, it is a carefully reviewed and high-quality process that has been shown to be robust and highly informative.[24–29,44] We take advantage of this fact and use it, for the first time as far as we know, to generate a general-purpose drug representation that can be used by a machine learning algorithm. We emphasize that our method does not rely on the availability of existing knowledge (but rather uses only co-occurrences) to predict the entire ADRs or indications labeling, enabling a "cold start" prediction. This is in contrast to other methods (eg, Refs [9,19]) which delete a subset of ADRs for each drug, and then use the other drug's known ADRs to predict the deleted ones.

One of the big challenges in medical data analysis is the use of multiple terminologies and standards across different knowledge-bases and databases.[45] Consequently, combining multiple data sources that use different terminology standards poses significant challenges. In our case, we had to map drug CIDs (used by SIDER) to MeSH terms, which currently do not have a direct structured mapping. To this end, we used heuristics that include automatic and semi-automatic procedures, as described in the "Materials and Methods" section. This mapping is provided as a Supplementary File, for the benefit of the entire community.

We reported results using PRAUC score, as it is a more appropriate measure when high precision is more important than high sensitivity. In the case of ADR prediction, and even more so for indication prediction, we focused on obtaining a high true positive rate because it is important that the positive predictions be correct with a high probability. As opposed to PRAUC, receiver operating characteristics (ROC) AUC takes into account true negatives; in our case, the data are very much biased toward easy-to-predict negative examples. This makes the ROCAUC score inappropriate as a performance measurement (see also Discussion in Ref [46]). To demonstrate this point, we doubled the number of ADRs/indications by syntheti-

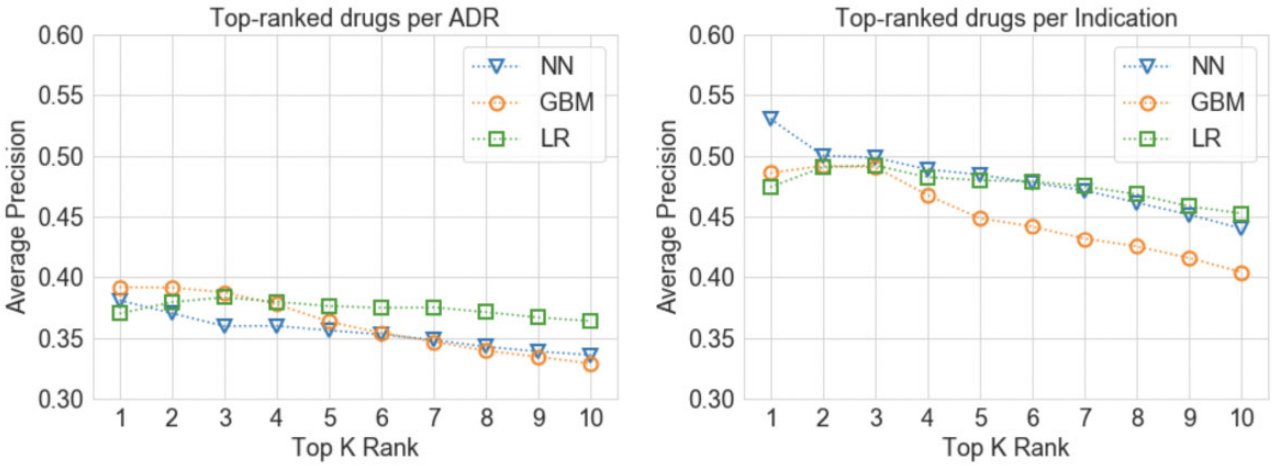

Figure 2. Task-specific performance. Markers indicate the average per-ADR (left) and per-indication (right) precision within the top-K ranked drugs (*x*-axis). ADR: adverse drug reaction.

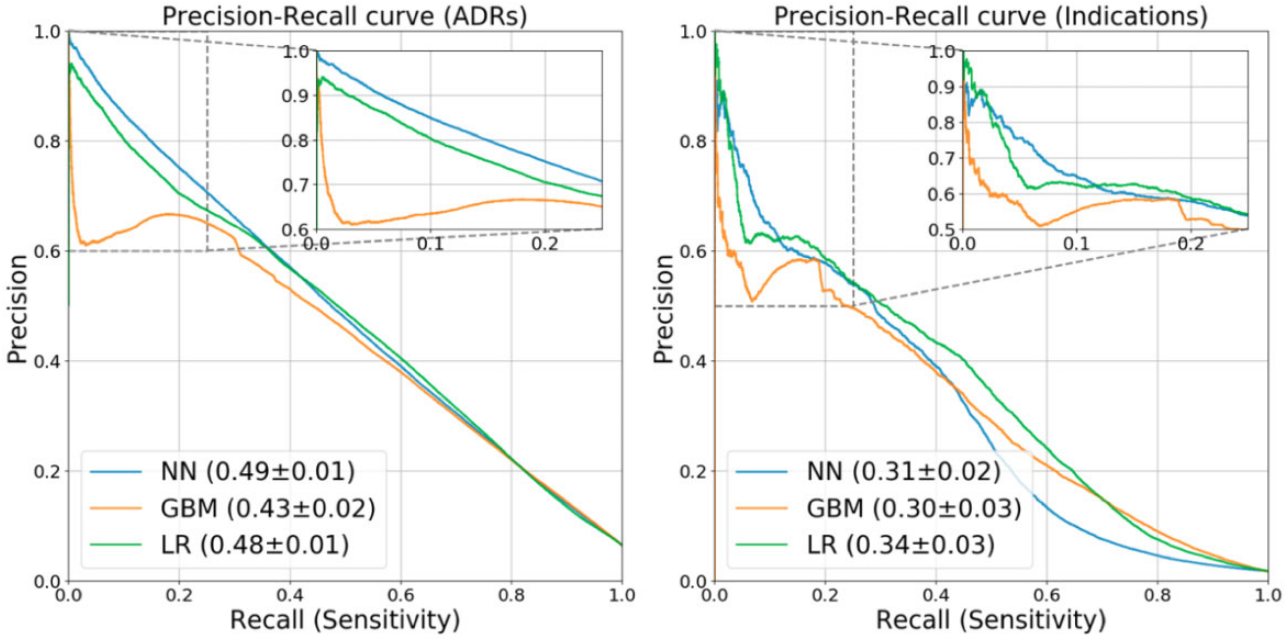

Figure 3. Overall prediction accuracy. Precision-recall curves plotted for the 3 trained model types in predicting ADRs (left) and indications (right); the PRAUC is shown in parentheses. The inset zooms in on the high precision range. ADR: adverse drug reaction; PRAUC: precision-recall area under the curve.

cally adding negative mock ADR/indications examples with low prediction scores. This caused the ROCAUC score of the NN to go up from 0.88 to 0.94 in the ADR prediction task and from 0.84 to 0.92 in the drug indications prediction task. In both cases, the PRAUC did not change, since it is not affected by true negatives. We also note that the ROCAUC score of the GBM model is slightly higher than the other 2 models, unlike what is depicted from the above PRAUC-based analysis. This is possibly due to better performance in the high sensitivity range. For completeness, we report the ROCAUC results in Supplementary Table S1.

We focused on 2 important tasks: prediction of ADRs and drug indications. ADRs represent the fourth leading cause of death in the United States, with an economic impact of more than $30B annually.[47] Predicting drug indications can potentially reduce the many years and enormous costs of drug development.[48] An important advantage of our method is that it can be applied to additional tasks for other groups of medical entities and relations between them, by using a similar representation scheme on the input data. It can also be used for unsupervised tasks such as symptoms classification or hierarchical clustering of diseases. We plan to explore these extensions in future work.

CONCLUSION

Selecting the data representation has a critical impact on the predictive power of machine learning models. We showed that a relatively simple representation scheme, based on medical term co-occurrences, can be effectively used for various prediction tasks related to drug development, namely prediction of ADRs and drug indications. We compared 2 modeling approaches: multi-task modeling based on NN and a combination of single-task modeling

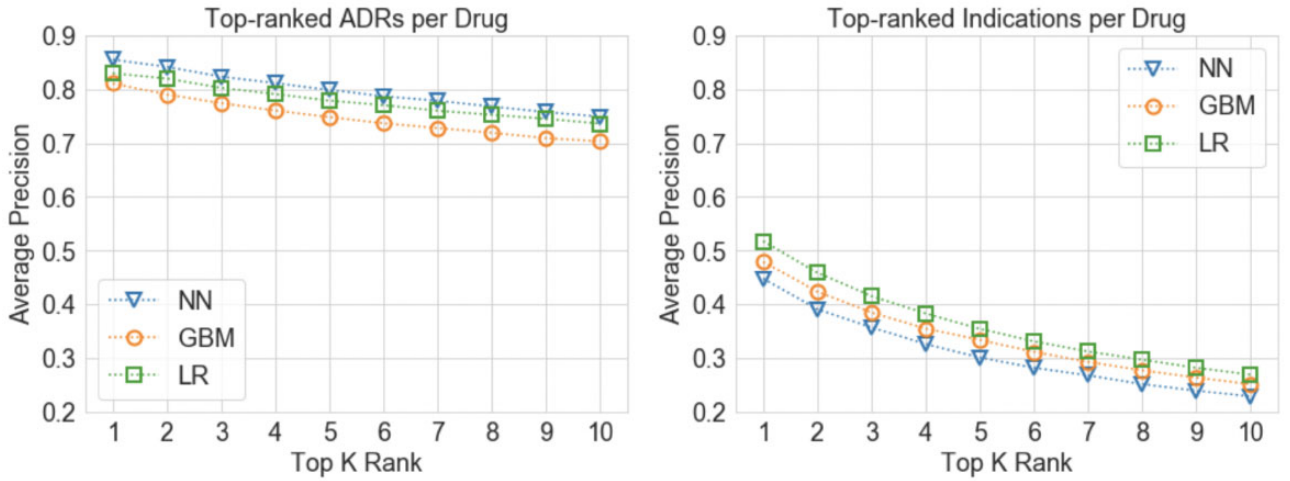

Figure 4. Per-drug performance. Markers indicate the average per-drug precision within the top-K ranked ADRs (left) and indications (right). ADR: adverse drug reaction.

Table 3. Model performance for the OMOP data

ADR	# of drugs with positive and negative relation	LR	GBM	Baseline
Acute kidney injury	24 positive, 63 negative	0.83 ± 0.08	0.81 ± 0.10	0.50 ± 0.22
Chemical and drug induced liver injury	81 positive, 36 negative	0.91 ± 0.04	0.94 ± 0.02	0.85 ± 0.08
Myocardial infarction	36 positive, 65 negative	0.66 ± 0.17	0.59 ± 0.24	0.51 ± 0.16
Gastrointestinal hemorrhage	24 positive, 67 negative	0.77 ± 0.27	0.76 ± 0.13	0.56 ± 0.24

Note. For each ADR, the table shows the number of drugs in the OMOP data (mapped to MeSH) with positive relation (drug causes the ADR) and negative relation (drug does not cause the ADR) and the PRAUC scores for all single-task models.

Abbreviations: ADR: adverse drug reaction; GBM: gradient boosting machines; LR: logistic regression; MeSH: Medical Subject Headings; OMOP: observational medical outcomes partnership; PRAUC: precision-recall area under the curve.

Table 4. PRAUC scores using subsets of the features

Features types used	# of features	PRAUC score ADRs	PRAUC score indications
All features	28 320	0.48 ± 0.01	0.34 ± 0.03
All except drugs and diseases	13 523	0.48 ± 0.01	0.32 ± 0.03
Drugs	10 086	0.47 ± 0.01	0.32 ± 0.03
Diseases without signs and symptoms	4344	0.47 ± 0.01	0.35 ± 0.03
Signs and symptoms	367	0.46 ± 0.01	0.27 ± 0.03
Baseline	1	0.12 ± 0.01	0.20 ± 0.07

Note. For each feature group, the table shows the number of features and the corresponding PRAUC scores for the ADRs and indications tasks.

Abbreviations: ADR: adverse drug reaction; PRAUC: precision-recall area under the curve.

based on either GBM or LR. All approaches achieved comparable results with marginal differences, depending on the specific task, but all outperformed the baseline model. Our suggested representation is relevant for a broad spectrum of prediction tasks that can be expressed using a generic semantic graph of various medical entities and relations.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

We wish to thank the researchers from the Machine Learning for Healthcare and Life Sciences team, IBM Research – Haifa, for fruitful discussions. We would also like to thank the anonymous reviewers for their helpful remarks and comments.

CONTRIBUTORS

AS conceived and led the study, drafted the manuscript, and conducted data analysis. AS and JFG collected and curated the data, generated descriptive statistics, and executed the data analysis. CY helped with the study design, edited the manuscript, conducted literature review, and added critical discussion points. All authors revised and approved the final manuscript.

FUNDING

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement number 766214, by providing funding to JFG.

Conflict of interest statement. None declared.

REFERENCES

1. Atias N, Sharan R. An algorithmic framework for predicting side effects of drugs. *J Comput Biol* 2011; 18 (3): 207–18.
2. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics* 2011; 12 (1): 169.
3. 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: 872.
4. Mizutani S, Pauwels E, Stoven V, *et al*. Relating drug-protein interaction network with drug side effects. *Bioinformatics* 2012; 28 (18): i522–8.
5. Zheng Y, Peng H, Ghosh S, *et al*. Inverse similarity and reliable negative samples for drug side-effect prediction. *BMC Bioinformatics* 2019; 19; (Suppl 13):554.
6. Luo H, Fokoue-Nkoutche A, Singh N, *et al*. Molecular docking for prediction and interpretation of adverse drug reactions. *Comb Chem High Throughput Screen* 2018; 21: 314–22.
7. 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* 2018; 25 (10): 1339–50.
8. P Tafti A, Badger J, LaRose E, *et al*. Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR Med Inform* 2017; 5 (4): e51.
9. Wang C-S, Lin P-J, Cheng C-L, *et al*. Detecting potential adverse drug reactions using a deep neural network model. *J Med Internet Res* 2019; 21 (2): e11016.
10. Liu J, Zhao S, Zhang X. An ensemble method for extracting adverse drug events from social media. *Artif Intell Med* 2016; 70: 62–76.
11. Chu J, Dong W, He K, *et al*. Using neural attention networks to detect adverse medical events from electronic health records. *J Biomed Inform* 2018; 87: 118–30.
12. Li J, Zheng S, Chen B, *et al*. A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016; 17 (1): 2–12.
13. Sirota M, Dudley JT, Kim J, *et al*. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med* 2011; 3 (96): 96ra77.
14. Gottlieb A, Stein GY, Ruppin E, *et al*. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2014; 7 (1): 496.
15. Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. *PLoS One* 2011; 6 (12): e28025.
16. Zhang P, Wang F, Hu J, *et al*. Exploring the relationship between drug side-effects and therapeutic indications. *AMIA Annu Symp Proc* 2013; 2013: 1568–77.
17. LaBute MX, Zhang X, Lenderman J, *et al*. Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PLoS One* 2014; 9 (9): e106298.
18. Sampathkumar H, Chen X, Luo B. Mining adverse drug reactions from online healthcare forums using hidden Markov model. *BMC Med Inform Decis Mak* 2014; 14: 91.
19. Zhang W, Zou H, Luo L, *et al*. Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* 2016; 173: 979–87.
20. Dey S, Luo H, Fokoue A, *et al*. Predicting adverse drug reactions through interpretable deep learning framework. *BMC Bioinformatics* 2018; 19(Suppl 21):476.
21. Henry S, McInnes BT. Literature based discovery: models, methods, and trends. *J Biomed Inform* 2017; 74: 20–32.
22. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif Intel* 1997; 91 (2): 183–203.
23. 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.
24. Smalheiser NR, Bonifield G. Two similarity metrics for medical subject headings (MeSH): an aid to biomedical text mining and author name disambiguation. *J Biomed Discov Collab* 2016; 7: e1.
25. Kastrin A, Rindflesch T, Hristovski D. Link prediction on a network of co-occurring MeSH terms: towards literature-based discovery. *Methods Inf Med* 2016; 55: 340–6.
26. Theodosiou T, Vizirianakis IS, Angelis L, *et al*. MeSHy: mining unanticipated PubMed information using frequencies of occurrences and concurrences of MeSH terms. *J Biomed Informat* 2011; 44 (6): 919–26.
27. Kim S, Yeganova L, Wilbur WJ. *Meshable*: searching PubMed abstracts by utilizing MeSH and MeSH-derived topical terms. *Bioinformatics* 2016; 32 (19): 3044–6.
28. Zhou J, Fu B. The research on gene-disease association based on text-mining of PubMed. *BMC Bioinformatics* 2018; 19; 19(1):37.
29. Lu Y, Figler B, Huang H, *et al*. Characterization of the mechanism of drug-drug interactions from PubMed using MeSH terms. *Plos One* 2017; 12 (4): e0173548.
30. Winnenburg R, Shah NH. Generalized enrichment analysis improves the detection of adverse drug events from the biomedical literature. *BMC Bioinformatics* 2016; 17; 17:250.
31. Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. *J Am Med Inform Assoc* 2011; 18 (5): 668–74.
32. Avillach P, Dufour J-C, Diallo G, *et al*. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU–ADR project. *J Am Med Inform Assoc* 2013; 20 (3): 446–52.
33. Kuhn M, Letunic I, Jensen LJ, *et al*. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016; 44 (D1): D1075–9.
34. Ryan PB, Schuemie MJ, Welebob E, *et al*. Defining a reference set to support methodological research in drug safety. *Drug Saf* 2013; 36 (Suppl 1): S33–47.
35. Kim S, Thiessen PA, Bolton EE, *et al*. PubChem substance and compound databases. *Nucleic Acids Res* 2016; 44 (D1): D1202–13.
36. Kuhn M, von Mering C, Campillos M, *et al*. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res* 2007; 36 (Database): D684–8.
37. Singhal A, Buckley C, Mitra M. Pivoted document length normalization. *SIGIR Forum* 2017; 51 (2):176–84.
38. Paszke A, Gross S, Chintala S, *et al*. Automatic Differentiation in Pytorch. 2017. Long Beach, California, USA :Autodiff Workshop
39. Ke G, Meng Q, Finley T, *et al*. LightGBM: a highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 2017; 3146–54.
40. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–30.
41. Ruder S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv: 170605098* 2017.
42. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning 2005 Aug 7. *ACM* 2005; 625–32.
43. Montastruc J, Sommet A, Bagheri H, *et al*. Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database. *Br J Clin Pharmacol* 2011; 72 (6): 905–8.
44. Baumann N. How to use the medical subject headings (MeSH). *Int J Clin Pract* 2016; 70 (2): 171–4.
45. Saitwal H, Qing D, Jones S, *et al*. Cross-terminology mapping challenges: a demonstration using medication terminological systems. *J Biomed Informat* 2012; 45 (4): 613–25.
46. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10 (3): e0118432.
47. Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. *J Pharmacol Pharmacother* 2013; 4 (5): 73–77.
48. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ* 2016; 47: 20–33.