Predicting the Side Effects of Drugs using Matrix Factorization on Spontaneous Reports Database

Kohei Fukuto
Osaka University
Tatsuya Takagi
Osaka University
Yu-Shi Tian (yushi-tian@phs.osaka-u.ac.jp)
Osaka University

Research Article

Keywords: Side-effect prediction, Machine learning, Recommender systems, Matrix factorization

Posted Date: July 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-737515/v1

Predicting the Side Effects of Drugs using Matrix Factorization on Spontaneous Reports Database

Kohei Fukuto, MSc[1]; Tatsuya Takagi, PhD[1]; Yu-Shi Tian, PhD[1,*]

[1]Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka 565-0871, Japan

* Corresponding author

ORCID ID

Tatsuya Takagi: 0000-0002-0044-0722

Yu-Shi Tian: 0000-0002-8988-9453

* Corresponding author:

Yu-Shi Tian

Graduate School of Pharmaceutical Sciences, Osaka University

1-6 Yamadaoka, Suita City, Osaka 565-0871, Japan

Phone No: +81-6-6879-8242

Fax No: +81-6-6879-8242

Email Address: yushi-tian@phs.osaka-u.ac.jp

Abstract

Background

Drugs with severe side effects can be threatening to patients and compromise pharmaceutical companies financially. Various computational techniques have been proposed to predict the side effects of drugs, including methods that utilize chemical, biological, and phenotypic features. Among them, matrix factorization (MF), which harnesses the known side effects of different drugs, has shown promising results. However, methods encapsulating all characteristics of side-effect prediction have not been investigated thus far. To this effect, we employed the logistic matrix factorization (Logistic MF) algorithm, i.e., MF modified for implicit feedback data, on a spontaneous reports database to improve the accuracy of side-effect prediction.

Results

A weighting strategy was applied to account for differences in the importance of the drug-side effect pairs. The impact of the cold-start problem and means to tackle it using the attribute-to-feature mapping were also explored. The experimental results demonstrate that the proposed model improved the prediction accuracy by 2.3% and efficiently handled the cold-start problem.

Conclusion

The proposed methodology is envisaged to benefit applications such as warning systems in clinical settings.

Keywords Side-effect prediction, Machine learning, Recommender systems, Matrix factorization

Background

Drugs with severe side effects can be fatal to patients and financially damage pharmaceutical companies. Drug safety information is typically evaluated using data from non-clinical studies and clinical trials. However, due to the limited number of patients and lower diversity in patient population partaking in clinical trials compared to those in actual use, it is fairly common for unknown side effects to be identified after a drug is launched.

Approaches for predicting the side effects of clinical drugs can be broadly divided into using chemical features such as drug structures; biological features such as target proteins, transporters, and enzymes; and phenotypic features such as side effects and therapeutic indications. Past studies have explored algorithms best suited to these approaches, such as using sparse canonical correlation analysis based on the chemical structure of drugs, using canonical correlation analysis and kernel regression based on chemical structures of drugs and target proteins, using logistic regression, naïve Bayes, k-nearest neighbor method, random forest, and support vector machines (SVM) based on chemical, biological, and phenotypic features [1-3]. In the last report, SVM has shown the highest potential, and the phenotypic features are most influential in acquiring predictions [3]. Another report pursued side-effect prediction as a multi-label prediction task and proposed using a k-nearest neighbor-based multi-label learning method [4].

The premise for predicting an unknown side effect based on known side effects (using the phenotypic features) is inspired by recommender systems, commonly utilized in e-Commerce websites to suggest products to users based on their past ratings and behavioral history. To date, predictive pharmacosafety networks (PPNs) constructing a network of drugs and side effects, and matrix factorization (MF), one of the most basic algorithms in recommender systems, have been applied to predict unknown side effects [5-6]. Furthermore, MF regularized by drug or drug and side-effect similarities has also been investigated for similar purposes [7-8].

However, these algorithms do not address several aspects of side-effect prediction. Firstly, the known side effect information is implicit feedback, i.e., if a side effect for a drug has not been reported, then an association between them either does not exist or has not been observed yet. However, MF models are typically designed for explicit feedback data. Secondly, previous

studies have not adequately accounted for the differences in weights among known drug-side effect pairs, apart from Xie and Poleksic [8] where they were all set to 1, and configuring these weights may prove pivotal in improving the prediction results. Finally, recommender systems are known to be afflicted by the cold-start problem, wherein the system is unable to furnish suitable predictions for drugs with very few known side effects, and no precedent has been set for it in side-effect prediction [9-10].

Additionally, previous studies have used Side Effect Resource (SIDER), an aggregated database comprising official documents and package inserts, for model training and evaluation [6-8, 11]. However, the latency in the occurrence of a side effect and updation of pertinent documentation may render the database obsolete for predicting side effects, which typically warrants real-time information. Therefore, we developed a custom dataset for this study derived from the FDA Adverse Event Reporting System (FAERS), a database of spontaneous adverse drug reaction reports maintained by the United States Food and Drug Administration (FDA).

Here, we utilized the logistic matrix factorization (Logistic MF) model [12], a modified MF model with implicit feedback, to predict severe side effects of clinical drugs more effectively based on a custom dataset derived from the FAERS database. We also simulated a cold-start scenario, investigated its impact, and explored attribute-to-feature mapping as a solution [13].

Methods

We downloaded the FAERS database, which stores spontaneous reports from healthcare professionals, patients, and pharmaceutical companies, from 2004 Q1 through 2019 Q2. The *DRUG* and *REAC* tables, in particular, were used to compile drug names and their corresponding side effects. A dataset representing associations between 1,127 drugs and 5,237 side effects, including 68 severe side effects, was created (see Supplementary Information Appendix 1).

Prediction Model

Matrix Factorization

The classic MF algorithm with explicit feedback has been extensively applied to movie rating predictions and other recommender systems. This method along with its variants has also previously been used in side-effect predictions [6-7].

Let m denote the number of drugs and n the number of side effects, then the number of reports for all drug-side effect pairs is represented by the $m \times n$ matrix, $C = (c_{ij})$, where c_{ij} is the number of times drug i is reported as the primary suspect for side effect j. Then, the matrix $A = (a_{ij})$ represents the association of all drug-side effect pairs given by:

$$a_{ij} = \begin{cases} 1, & c_{ij} \geq 3 \\ 0, & c_{ij} < 3 \end{cases}$$

The larger the threshold of occurrence, the more likely it is that drug-side effect associations are overlooked, and the smaller the threshold, the more likely it is that noise in the dataset is labeled as meaningful signals. Thus, we configured the threshold value as 3 for this study in compliance with the conventions in the signal detection field [14].

MF assumes that each drug and side effect has latent factors of dimension k. Let d_i denote the latent factor vector of drug i by and s_j of side effect j, then a_{ij} can be estimated as:

$$\widehat{a_{ij}} = {d_i}^T s_j + b_i + b_j$$

where b_i and b_j are the bias terms for drug i and side effect j respectively [15].

Latent factors are learned by minimizing the squared error as:

$$\min_{D,S} \sum_{(i,j)\in A} \left(a_{ij} - \widehat{a_{ij}}\right)^2 + \lambda \left(\|d_i\|^2 + \|s_j\|^2\right)$$

where D is an $m \times k$ matrix with row i being d_i, and S is an $n \times k$ matrix with row j being s_j. The second term in the loss function is the penalty term for the latent factors to prevent overfitting. λ is the hyperparameter that controls the degree of regularization.

However, this method possesses two shortcomings. First, the number of reported side effects can be regarded as implicit feedback; hence, there is no distinction between the negative

and unobserved examples in A, implying that the corresponding zero entries are potential positive examples. However, the model learns these zero entries as is, thereby reducing its efficiency in predicting the missing side effects. Second, the model does not consider differences in the importance or weight of the associations between the drugs and side effects.

Logistic Matrix Factorization

Logistic MF modifies the MF schema for implicit feedback data [12]. Assuming that the objective variable in the implicit feedback data is binary, Logistic MF employs the sigmoid function, σ, for supplying predictions. Then a_{ij} is computed as follows:

$$\widehat{a_{ij}} = \sigma(d_i^T s_j + b_i + b_j)$$

Latent factors are learned by minimizing the log loss as:

$$\min_{D,S} -\sum_{(i,j)\in D} w_{ij} \left\{a_{ij} \log \widehat{a_{ij}} + \left(1 - a_{ij}\right) \log\left(1 - \widehat{a_{ij}}\right)\right\} + \lambda \left(\|d_i\|^2 + \left\|s_j\right\|^2\right)$$

where w_{ij} corresponds to the weight of each drug-side effect pair.

In [12], $c_{ij} = 1$ is the preconfigured threshold, and $w_{ij} = \alpha\, c_{ij}$ and $w_{ij} = 1 + \alpha \log(1 + c_{ij}/\varepsilon)$ are provided as examples of the weighting functions, where α is a hyperparameter. However, these weighting functions vary depending on the characteristics of the problem. Hence, for this study, we configured $c_{ij} = 3$ and, assuming that the effect of the number of reports on the weights is not linear but grows logarithmically, used the following weighting function:

$$w_{ij} = \begin{cases} 1 + \alpha \log\left(1 + c_{ij}\right), & c_{ij} \geq 3 \\ \beta, & c_{ij} < 3 \end{cases}$$

where β is another hyperparameter used to reduce the impact of negative examples on the overall loss function, to account for the implicit feedback.

Attribute-to-feature mapping

Attribute-to-feature mapping is known to improve the prediction accuracy in cold-start scenarios by learning the mapping function of the user or item attributes to latent factor vectors [13]. In cold-start problems associated with side-effect predictions i.e., adequate information

on side effects for a particular drug is not available causing the model to incorrectly learn the latent factor vector, estimating latent factors from secondary data, like drug structures, may help to improve prediction accuracy.

The k-nearest neighbor and linear mapping algorithms have been previously proposed to map attributes to latent factors, eliciting superior results when the latter algorithm is optimized for the final evaluation metric rather than the squared error, except when the dimension of attributes is extremely high [13]. Here, a linear mapping from attributes to latent factors of drugs was expressed as follows:

$$\widehat{d}_i = M^T desc_i$$

where $desc_i$ is the attribute of drug i, and M is the learnable parameter matrix of the mapping function having the shape of $(n,\ k)$ where n is the dimension of drug attribute and k is the dimension of latent factors.

Conversely, RDKit molecular descriptors [16] and extended-connectivity fingerprints (ECFP) [17] were used as drug attributes. 2048-bit fingerprints generated by the ECFP were reduced to 100 dimensions using kernel principal component analysis (KPCA) [18]. The hyperparameters of KPCA were determined by conducting a grid search on the validation set.

Experiment

Data Preparation

This study attempts to construct MF and Logistic MF models for side-effect prediction and investigate the impact of the cold-start problem. The cold-start scenario was simulated by removing some of the known side effects of drugs used for model evaluation. However, if we randomly split all drug and side effect pairs into training, validation and test set as in the typical evaluation scheme of MF, at least one drug and side effect pair for most drugs will be included in the test set. Thus, removing some of the training pair of these drugs will significantly reduce the amount of training data, resulting in an unrealistic situation. To ensure that the model training does not get affected by the simulation, we adopted a unique data splitting strategy as below.

The dataset leading up to 2015 Q3 was employed for the study. Drugs were randomly split

in half to procure the training and test drugs; 20% of the training drug and side effect pairs were set aside for validation while the rest were used for training, and 40% of test drug and side effect pairs were used for testing while the rest were also used for training. Overall, 70% of all the drug-side effect pairs were used for training, 10% for validation, and 20% for testing. When considering the cold-start situation, only the side effect information in training sets from test drugs was removed. In contrast, the known side effects of training drugs remained the same.

Evaluation metric

The area under the precision-recall curve (PR-AUC) was the primary evaluation metric applied for each side effect. All training data pairs were used to calculate the loss function during training, but the average PR-AUC of severe side effects was used for early stopping. The dataset was partitioned five times and the mean and standard deviation of the evaluation metrics were computed.

Hyperparameter search

A grid search was conducted to locate hyperparameters with the highest evaluation metric in the validation set (Table 1). The experiment was repeated five times and the hyperparameters obtained in the first repetition were fixed for the following cycles. The latent factor parameters were regularized using λ, while α and β were used to adjust the positive and negative example weights. The latent factor dimensionality was fixed at 100 [7]. The number of training epochs was determined by early stopping with the PR-AUC in the validation set. The initial learning rate was set to 0.01 and scheduled to decrease at a fixed value of 0.1 whenever the PR-AUC value dipped in the validation set to avoid local optimal solutions. The Adam optimizer was applied to the loss function [21].

Results and Discussion

Comparing the prediction accuracy

Table 2 highlights the mean and standard deviations of PR-AUCs in the test set for data up to

2015 Q3 for both models. Results for other severe side effects are available in Supplementary Information Appendix 2.

The mean PR-AUC of Logistic MF improved by 2.3% compared to MF. Despite the large standard deviation attributed to a limited number of positive examples in the test set, the sigmoid and weight functions consistently ascertained superior prediction performances. The optimal hyperparameters were $\lambda = 0.005$ for MF and $\lambda = 0.005$, $\alpha = 10$, and $\beta = 0.8$ for Logistic MF.

External validation using future data

The viability and robustness of the proposed model were evaluated using future data from 2015 Q4 onwards. To achieve this, data pairs up to 2015 Q3 were randomly split, of which 10% was used as the validation set. Both models were trained and the model output for drug-side effect pairs with negative labels in the training set i.e., the pairs occurring less than thrice by 2015 Q3, were obtained. The PR-AUCs were then computed using future labels. Table 3 consolidates the results and those of other severe side effects are listed in Supplementary Information Appendix 3.

Results of external validation once again favor our model over MF in predicting side effects more accurately. Please note that the PR-AUC values in tables 2 and 3 cannot be directly compared, owing to the difference in the number of positive examples in the validation schemes, which in turn affects the PR-AUC values. However, the difference in these values is significant, indicating that employing a random split on data generated in a time-series manner may invoke an overly optimistic evaluation of the prediction performance in both models.

Cold-start problem: simulation and results

We simulated a cold-start scenario, i.e., reducing the number of known side effects in the test drugs, and investigated its impact on the prediction performance of the proposed model. We randomly removed training data for a test drug in a defined *test_delete_ratio* and reported the evaluation metrics of the test set at different *test_delete_ratio*s. The weight of the deletion probability was considered based on the number of known side effects. The results are shown

in Fig. 1. The PR-AUC decreased significantly with fewer known side effects, suggesting that the prediction accuracy of our model deteriorated when information about test drugs is insufficient, as may be the case with drugs in the early stages of development or clinical trials.

Effect of attribute-to-feature mapping

We applied the attribute-to-feature mapping to our model, represented by Map-LMF, for the cold-start scenario. The PR-AUCs of the Logistic MF and Map-LMF models for varying numbers of known side effects are highlighted in Table 4.

Predicting the latent factor vectors using ECFP as the drug attribute improved the prediction accuracy under cold start settings. The prediction accuracy of Map-LMF exceeded that of Logistic MF by 2.2% and 7.3% at *test_delete_ratio* = 0.95 and 0.99 with RDKit descriptors, and by 7.2% and 12.4% at *test_delete_ ratio* = 0.95 and 0.99 with ECFP. As previously established, inadequate information on the known side effects of a test drug adversely affected the prediction accuracy. Therefore, the latent factors estimated from the chemical compositions of the drugs evinced better predictions.

Conclusion

Drugs with severe side effects endanger patients and pharmaceutical companies. Therefore, an effective methodology needs to be investigated to predict these side effects and, in turn, ascertain patient safety and efficient drug development. MF has been previously utilized for side-effect prediction. We consolidated the available knowledge on MF and its shortcomings, such as its inability to handle implicit feedback and cold start problems, and identified Logistic MF as an efficient model to meet our objective. Results affirmed that our proposed model improved the overall prediction accuracy by 2.3% and produced superior performance in the cold-start settings using attribute-to-feature mapping by 12.4% at most.

Limitations of the study are as follows. We could not determine whether all drugs from the FAERS database were included in the final dataset during data pre-processing due to incomplete mapping between drug names and their structures. Furthermore, the preconfigured threshold value for forging drug-side effect associations may have overlooked the possibility of

mislabeled drugs due to noises in the spontaneous reports database. In the future, we envisage incorporating the signal detection criteria to extract drug-side effects pairs from the reports database more accurately and find feasible solutions to the other drawbacks identified.

Declarations

Abbreviations

ECFP: extended-connectivity fingerprints; FAERS: the FDA Adverse Event Reporting System; FDA: the United States Food and Drug Administration; KEGG: Kyoto Encyclopedia of Genes and Genomes; KPCA: kernel principal component analysis; MF: matrix factorization; MHLW: the Ministry of Health, Labour and Welfare; PPN: predictive pharmacosafety networks; PR-AUC: the area under the precision-recall curve; SIDER: Side Effect Resource; SVM: support vector machine.

Availability of data and materials

This study analyzed the US FDA FEARS database, which can be obtained from the site of the US FDA. The codes used in the current study are available at https://github.com/ykskks/Matrix-Factorization-for-Drug-Side-Effect-Prediction.

Ethics approval and consent to participate

Not applicable.

Consent to publish

All the authors agree to publish.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by JSPS KAKENHI Grant Number JP15KT0017.

References

[1] Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics. 2011;12:169.

[2] Yamanishi Y, Pauwels E, Kotera M. Drug Side-Effect Prediction Based on the Integration of Chemical and Biological Spaces. J Chem Inf Model. 2012;52:3284–92.

[3] Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen X, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. Journal of the American Medical Informatics Association. 2012;19:e28–35.

[4] Zhang W, Liu F, Luo L, Zhang J. Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinformatics. 2015;16:365.

[5] Cami, A. Arnold, S. Manzi, and B. Reis, "Predicting adverse drug events using pharmacological network models," *Sci. Transl. Med.*, vol. 3, no. 114, 2011, doi: 10.1126/scitranslmed.3002774. 1. Cami A, Arnold A, Manzi S, Reis B. Predicting Adverse Drug Events Using Pharmacological Network Models. Science Translational Medicine. 2011;3:114ra127-114ra127.

[6] Galeano D, Paccanaro A. A Recommender System Approach for Predicting Drug Side Effects. In: 2018 International Joint Conference on Neural Networks (IJCNN). 2018. p. 1–8.

[7] Zhang W, Liu X, Chen Y, Wu W, Wang W, Li X. Feature-derived graph regularized matrix factorization for predicting drug side effects. Neurocomput. 2018;287 C:154–62.

[8] Poleksic A, Xie L. Predicting serious rare adverse reactions of novel chemicals. Bioinformatics. 2018;34:2835–42.

[9] Lam XN, Vu T, Le TD, Duong AD. Addressing cold-start problem in recommendation systems. In: Proceedings of the 2nd international conference on Ubiquitous information management and communication. New York, NY, USA: Association for Computing Machinery; 2008. p. 208–11. doi:10.1145/1352793.1352837.

[10] Lika B, Kolomvatsos K, Hadjiefthymiades S. Facing the cold start problem in recommender systems. Expert Syst Appl. 2014;41:2065–73.

[11] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects.

Nucleic Acids Research. 2016;44:D1075–9.

[12] Johnson C. Logistic matrix factorization for implicit feedback data. Adv. Neural Inf. Process. Syst. 2014;27.

[13] Gantner Z, Drumond L, Freudenthaler C, Rendle S, Schmidt-Thieme L. Learning Attribute-to-Feature Mappings for Cold-Start Recommendations. In: 2010 IEEE International Conference on Data Mining. 2010. p. 176–85.

[14] Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety. 2001;10:483–6.

[15] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems. Computer. 2009;42:30–7.

[16] RDKit. https://www.rdkit.org/. Accessed 20 Jul 2021.

[17] Rogers D, Hahn M. Extended-Connectivity Fingerprints. J Chem Inf Model. 2010;50:742–54.

[18] Schölkopf B, Smola A, Müller K-R. Kernel principal component analysis. In: Gerstner W, Germond A, Hasler M, Nicoud J-D, editors. Artificial Neural Networks — ICANN'97. Berlin, Heidelberg: Springer; 1997. p. 583–8.

Figure and table legends

Table 1 List of hyperparameters and their range in the grid search

	Hyperparameter	*Range*
MF	λ	$[1.0\times 10^{-4},\ 5.0\times 10^{-4},\ 1.0\times 10^{-3},\ 5.0\times 10^{-3},\ 1.0\times 10^{-2}]$
Logistic MF	λ	$[1.0\times 10^{-4},\ 5.0\times 10^{-4},\ 1.0\times 10^{-3},\ 5.0\times 10^{-3},\ 1.0\times 10^{-2}]$
	α	[0, 1, 2, 5, 10, 15]
	β	[0.2, 0.4, 0.6, 0.8, 1.0]

Table 2 Test PR-AUC for MF and Logistic MF

	mean	*SJS*	*LPT*	*NMS*
MF	0.787 ± 0.017	0.876 ± 0.018	0.938 ± 0.022	0.686 ± 0.082
Logistic MF	0.810 ± 0.021	0.864 ± 0.019	0.953 ± 0.009	0.756 ± 0.083

SJS = Stevens-Johnson syndrome, LPT = Low Platelet Count, NMS = Neuroleptic malignant syndrome

Table 3 PR-AUC in the external validation for MF and Logistic MF

	mean	*SJS*	*LPT*	*NMS*
MF	0.289 ± 0.007	0.268 ± 0.022	0.233 ± 0.028	0.204 ± 0.052
Logistic MF	0.299 ± 0.001	0.301 ± 0.009	0.239 ± 0.012	0.244 ± 0.017

Table 4 Test PR-AUC for Logistic MF and Map-LMF with varying number of known side effects

test_delete_ratio	*0.80*	*0.90*	*0.95*	*0.99*
Logistic MF	0.550 ± 0.009	0.377 ± 0.008	0.286 ± 0.011	0.235 ± 0.010
Map-LMF (RDKit)	0.309 ± 0.009	0.310 ± 0.010	0.308 ± 0.010	0.308 ± 0.010
Map-LMF (ECFP)	0.357 ± 0.019	0.357 ± 0.020	0.358 ± 0.020	0.359 ± 0.019

Figures

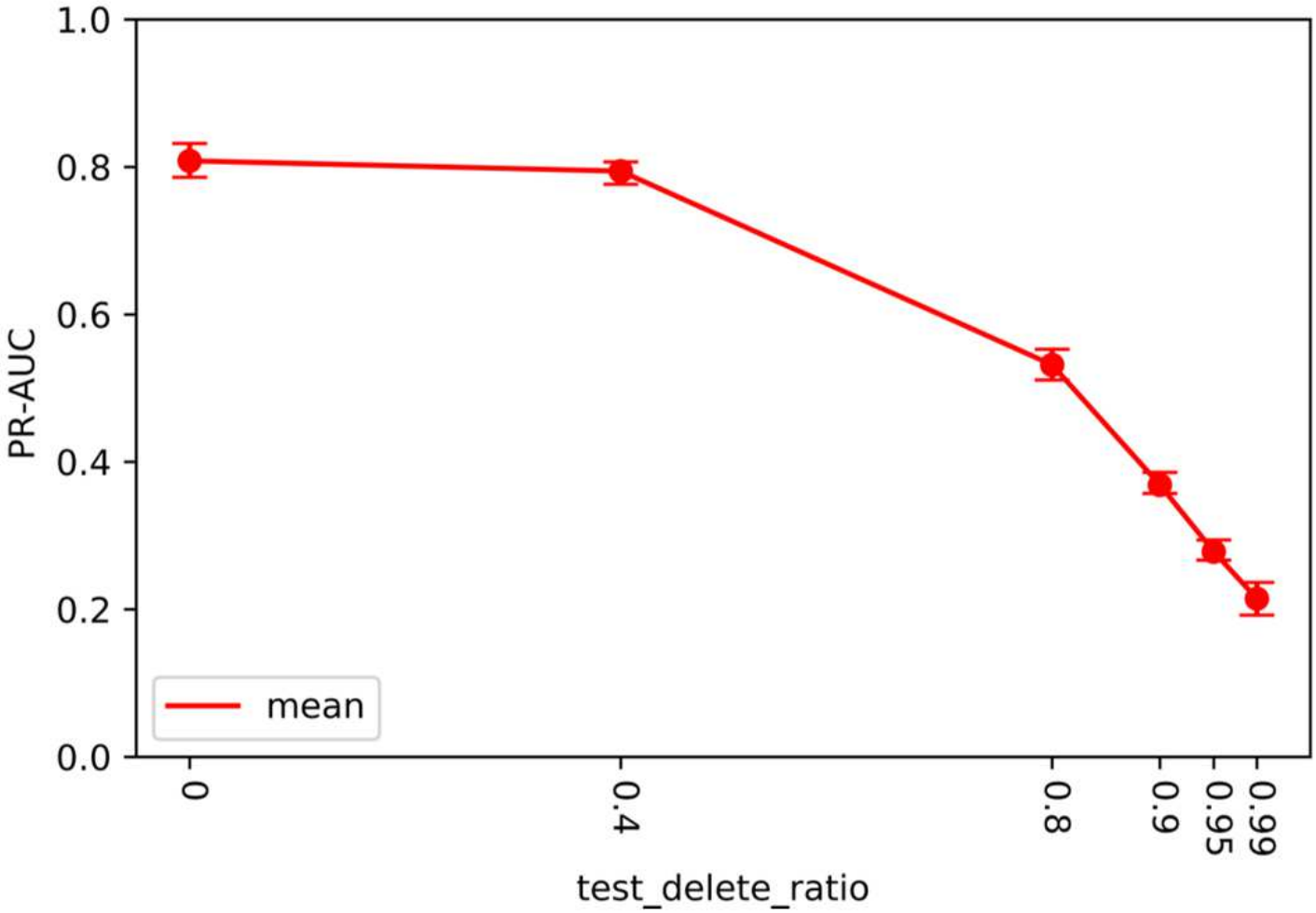

Figure 1

Test PR-AUC with varying number of known side effects

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx