Adverse Drug Reaction Prediction with Symbolic Latent Dirichlet Allocation

Cao Xiao,* Ping Zhang,* W. Art Chaowalitwongse,† Jianying Hu,* Fei Wang¶
*IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
† University of Arkansas, Fayetteville, AK 72701
¶ Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, New York, NY 10065

Abstract

Adverse drug reaction (ADR) is a major burden for patients and healthcare industry. It usually causes preventable hospitalizations and deaths, while associated with a huge amount of cost. Traditional preclinical *in vitro* safety profiling and clinical safety trials are restricted in terms of small scale, long duration, huge financial costs and limited statistical significance. The availability of large amounts of drug and ADR data potentially allows ADR predictions during the drugs' early preclinical stage with data analytics methods to inform more targeted clinical safety tests. Despite their initial success, existing methods have trade-offs among interpretability, predictive power and efficiency. This urges us to explore methods that could have all these strengths and provide practical solutions for real world ADR predictions. We cast the ADR-drug relation structure into a three-layer hierarchical Bayesian model. We interpret each ADR as a symbolic word and apply latent Dirichlet allocation (LDA) to learn topics that may represent certain biochemical mechanism that relates ADRs with drug structures. Based on LDA, we designed an equivalent regularization term to incorporate the hierarchical ADR domain knowledge. Finally, we developed a mixed input model leveraging a fast collapsed Gibbs sampling method that the complexity of each iteration of Gibbs sampling proportional only to the number of positive ADRs. Experiments on real world data show our models achieved higher prediction accuracy and shorter running time than the state-of-the-art alternatives.

1 Introduction

An adverse drug reaction (ADR) is a harmful or unpleasant reaction caused by taking a medication. ADRs become a huge public health concern as they result in $100,000$ fatalities per year and incur morbidity and mortality related cost $\sim \$177$ billion annually (Giacomini et al. 2007). Traditional ADR detection strategies include preclinical *in vitro* safety profiling and clinical drug safety trials, however, both are restricted in terms of small scale, long duration, huge financial costs and limited statistical significance (Whitebread et al. 2005).

Meanwhile, the availability of large amount of drug and ADR data provides a unique opportunity to address the challenges with predictive modeling methods. Accurate prediction of potential ADRs at the early stage of drug development cycle could recommend targeted safety tests and thus help reduce time and financial costs in drug safety trials. As a proof-of-concept, (Bender, Scheiber, and Glick 2007) explored the chemical space of drugs and established its correlation for ADR prediction. (Scheiber, Jenkins, and Sukuru 2009) presented a global analysis that identified chemical substructures associated with ADRs. However, given millions of marketed drugs or experimental lead compounds and thousands of candidate ADRs, accurate prediction of ADRs at early stage still remains a challenging task.

State-of-the-art The aforementioned challenges motivate a series of works that apply data analytics methods to seek co-occurring patterns between drugs and ADRs. From the perspective of representation learning, CCA (Hotelling 1936) was applied to identify projections of drug features and ADRs that are maximally correlated (Liu et al. 2012). Kernel CCA is also adopted in a similar fashion. (Pauwels, Stoven, and Yamanishi 2011) treated k-nearest neighbors (Altman 1992) results from different similarity measures as kernels and developed a sparse canonical correlation analysis (CCA) method to predict high-dimensional ADR profiles of drug molecules based on drug structures. While (Yamanishi, Pauwels, and Kotera 2012) developed a multiple kernel regression method that integrates drug and biological features to predict ADRs. On the other hand, from the perspective of relational modeling, lasso (Tibshirani 1994) was a popular method for ADR prediction. For example, (Caster 2007) built a multivariate lasso framework to simultaneously treat all drugs as predictors for the presence of the ADRs and showed good prediction results. In addition, inductive matrix-completion methods were also developed, but they often need to be combined with drug-target interactions (Li et al. 2015) or other side information.

Challenges Despite their initial success, each of existing data analytics method requires some trade-offs among performance, interpretability and efficiency. For example, the canonical variates in CCA or sparse CCA do not have particular meaning other than linear transformation. It may also be sensitive to noise or collinearity in the data (McCune 1997). In our case, the drug features are highly correlated, thus we observed poor generalization power and results in poor prediction performance on target drugs. Nonlinear CCA or other kernel based methods further lack interpretability since it is difficult to explain the meaning of a nonlinear combination of the entire set of variables. Lasso, though has better prediction quality, would require an external loop to iterate on each response variable dimension, and need to determine the penalty factor for each dimension in learning.

Our Approach To address the arising challenges, we noticed that topic models such as LDA (Blei, Ng, and Jordan 2003), though originally developed to characterize semantic relations, could be transferred to model the symbolic relations between drugs and ADRs. For a LDA-like model, we constructed its "drug document" as a mixture of "ADR topics", where a topic consists of a set of words (ADR terms) that frequently occur together across the drug documents. We considered drug structures as features to correlate structural information to drug topics for ADR prediction. Further, we provided a variation of the base model to add an equivalent regularization penalty that could incorporate domain knowledge to enhance interpretability. Finally, we proposed a mixed input (drug features and ADR terms) model that fully leverages the fast collapsed Gibbs sampling method to speed up the learning and still preserves high accuracy. Our main contributions are:

- We developed three LDA-like models that achieved higher accuracy than the state-of-the-art alternatives. Their performance demonstrated that LDA could successfully uncover the probabilistic patterns among ADR topics, and showed the applicability of this three-level hierarchical Bayesian approach. In general, Bayesian approach avoids overfitting the data, and is especially useful on small datasets.
- The LDA-like models could generate interpretable results such that each topic is explained as a probability distribution over ADRs, while each drug document can be understood as a mixture of topics. Each topic may represent certain biochemical mechanism that relates ADRs with drug structures. The regularized version further enhanced the interpretability with domain knowledge incorporated.
- Empirical results from real world data showed that the mixed input model runs significantly faster, with the complexity of each iteration of Gibbs sampling proportional to the number of positive ADRs instead of total candidate ADRs (Porteous et al. 2008). It manages to achieve top accuracy, and hence is a substantial contribution towards scalable ADR prediction.

The rest of the paper is organized as follows. In section 2, we talk about the building blocks of our models. Then we introduce the developed models in Section 3, and evaluate it with real world data in Section 4. We also discuss results in Section 4.4. Last, we conclude our work and highlight future directions in Section 5.

2 Background

Drug structure features and LDA methods are the two major building blocks of this work. Therefore, in this section, we briefly introduce why we choose them and how we use them in this work.

2.1 Drug Structure Features

Drug structure features (i.e. chemical fingerprints) are structural descriptors of drugs. In our study, we generate drug structure features with the extended-connectivity fingerprints with diameter 6 (ECFP6) (Rogers and Hahn 2010), a technique developed specifically for structure-activity modeling, by using "rcdk" package (Guha 2007). The features are hashed binary vectors of 1,024-bit length, of which each bit encodes the presence or absence of a substructure in a drug molecule, allowing better representation in structural similarity. We use drug structure features to correlate target drug to training drugs and thus to infer the ADR topics for the target drug.

2.2 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) is a probabilistic topic modeling method that aims at finding concise descriptions for a data collection. Originally proposed in the context of text document modeling, LDA discovers latent semantic topics in large collections of text data. Each discovered topic is characterized by its own particular distribution over words. Each document is then characterized as a random mixture of topics indicating the proportion of time the document spends on each topic. This random mixture of topics is essentially the "concise descriptions". LDA not only expresses the semantic content of a document in a concise manner, but also gives us an interpretable approach for describing documents quantitatively via how similar the corresponding topic mixtures are.

In recent years, LDA has been extended to be a general topic discovering framework in numerous domains, including object recognition (Cao and Li 2007), spam filtering (Biro et al. 2009) and web mining (Mei et al. 2006). In the domain of medical informatics, there has been a few previous work that link LDA to semantic drug label mining (Bisgin et al. 2011; Paul and Dredze 2012), however, there has never been such a work that designs LDA based models to uncover the symbolic relations underpinning drugs and ADRs.

3 Method

Now we are ready to introduce the three LDA-like models. Our task here is to predict potential ADRs from thousands of candidate ADRs for a target drug. Such a task is challenging due to the huge amount of ADR terms as response. Some ADRs would co-occur to a drug is due to the common underlying mechanism the ADRs share, which could be understood as a "ADR topic". Therefore, it is plausible to apply LDA to achieve dimension reduction via considering each drug as a document: a mixture of "ADR topics", where a topic consists of a set of ADR terms that frequently occur together across the drug documents. To make predictions, we firstly identify ADR topic distributions that characterize the target drug, then we build a predictive model to relate drug structure features to the ADR topic distributions. After that, we could predict the ADRs associated with the drug through its topic distribution. We adopt the following training and prediction procedure. And we also list the notations used in the models in Table 1.

```
procedure THE TRAINING PROCEDURE
    $\{Doc^1, \dots, Doc^D\}$ ← drug document
    $K$ ← number of ADR topics
    train LDA($\{Doc^1, \dots, Doc^D\}, K$)
    $\beta$ ← ADR distribution of ADR topics
    $\Theta$ ← ADR topic distribution
    $\vec{X} = \{\vec{x}_d\}$ ← drug structure features
    train $\theta_d \sim \vec{x}_d$ with lasso.
end procedure
```

```
procedure THE PREDICTION PROCEDURE
    $\vec{x}'_d$ ← drug structure features for target drug $d'$
    predict topic distribution $\theta_{d'}$ with lasso.
    $\beta$ ← ADR distribution of ADR topics
    predict ADRs using $\beta$ and $\theta_{d'}$ with LDA

end procedure
```

3.1 The Base Model

The LDA concept mapping directly generates the base model. The graphical structure is illustrated as in Figure 1 and its generative process is described in the procedure.

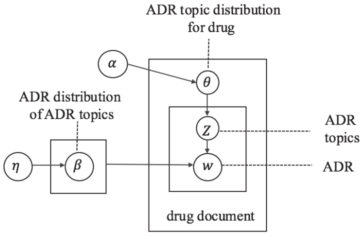

Figure 1: The base model

Here we consider each drug as a document that defines a probability distribution on topics, and each topic defines a probability distribution on ADRs. The ADRs follow a multinomial distribution with one probability vector per topic. The conditional probability distribution for the ADR ω under ADR topic k is given by Formula 1.

$$p(w_n = \omega | z_n = k; \beta_{1:K}) = \beta_k(\omega) \qquad (1)$$

In addition, the topics themselves follow a multinomial distribution $Multinomial(\theta)$, where θ is a Dirichlet variable. The marginal distribution of ADR topic is given by Formula 2.

$$p(z_n = k | \theta) = \theta(k). \qquad (2)$$

where the Dirichlet variable θ has a marginal density of the form in Formula 3.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1} \qquad (3)$$

Table 1: Mathematical Notation

Notation	Description
D	# of drugs
N	# of candidate ADRs
K	# of topics
$w_{d,n}$	binary variable for drug d has ADR n
$\vec{x}_d = \{x_d^1, \cdots, x_d^p\}$	drug features for drug d
θ_d	ADR topic distribution for drug d
$z_{d,n}$	topic assignment for $w_{d,n}$
β_k	ADR distribution of topic k
α, η	Dirichlet/beta parameters of the Multinomial/Bernoulli distribution

```
procedure THE GENERATION PROCESS OF LDA FOR DRUG DOCUMENT
    $d$ ← index of drug
    $D$ ← number of drugs in the corpus
    for $d \in [1, \dots, D]$ do
        draw a topic mixture $\theta_d$ such that $p(\theta|\alpha) = Dirichlet(\alpha)$.
    end for
    $n$ ← index of ADR
    $N$ ← number of candidate ADRs in the ADR corpus
    for $n \in [1, \dots, N]$ do
        draw a topic $z_n \sim Multinomial(\theta)$.
        draw a word $w_n$ from $p(w_n|z_n, \beta)$, a multinomial conditioned on topic $z_n$
    end for
end procedure
```

Our objective is to estimate the parameters of the ADR distribution per ADR topic as well as the topic distribution per drug, with Formula 4 being the objective function.

$$\ell(\theta_{1:D}, \beta_{1:K}; \alpha, \eta) = \sum_{k=1}^{K} \log p(\beta_k|\eta) + \sum_{d=1}^{D} \log p(\theta_d, w_d | \alpha, \beta_{1:K}) \qquad (4)$$

Here the density function $p(\beta_k|\eta)$ for β_k is of the form as in Formula 5.

$$p(\beta_k|\eta) = \frac{\Gamma(\sum_{n=1}^{N} \eta_n)}{\prod_{n=1}^{N} \Gamma(\eta_n)} \beta_{k1}^{\eta_1 - 1} \cdots \beta_{kN}^{\eta_N - 1} \qquad (5)$$

While the marginal distribution of drug d is as Formula 6.

$$p(\theta_d, w_d | \alpha, \beta_{1:K}) = p(\theta_d|\alpha) \cdot \prod_{n=1}^{N} \left(\sum_{z_{d,n}} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{1:K}) \right) \qquad (6)$$

Given the formulation above, from existing drug and ADR relations, we train a LDA model to generate ADR topics, where the topic distribution for drug d is denoted as θ_d. In prediction, we predict θ_d for a target drug d' given drug structure features $\vec{x}'_d = \{x'^1_d, \cdots, x'^p_d\}$, and then generate the distribution of ADRs with $\theta_{d'}$.

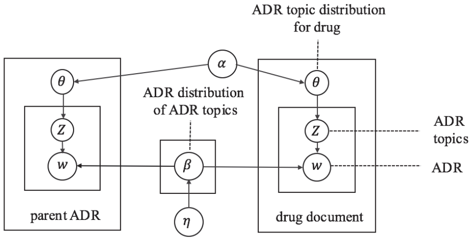
ADR topic distribution for drug
α
θ
θ
ADR distribution of ADR topics
Z
Z
ADR topics
w
β
w
ADR
parent ADR
η
drug document

Figure 2: The regularized model

3.2 Regularization to Incorporate Domain Knowledge

To incorporate domain knowledge, such as the available hierarchical ADR ontology systems (e.g. ADReCS (Cai et al. 2015), and WHO-ART(WHO-ART 2015)), we extend the base LDA-like model with a hierarchical penalty term. These ontology systems are multi-level ADR hierarchy trees, where concepts described by higher level nodes ($par(\cdot)$) are the abstract versions of their child nodes. To faithfully reflect such knowledge, we introduce a hierarchical penalty term that gives out a penalty when two ADR terms have the same topic but different ADR parents.

Here we denote w as ADR terms, z as the ADR-to-topic assignment. We also set superscript h to indicate the hierarchy layer from bottom layer coded "1" to top layer coded "T", and let γ_h be penalty coefficient. Our objective is to minimize the following penalized negative log-likelihood function as in Formula 7.

$$-\ell(\theta_{1:D}, \beta_{1:K}; \alpha, \eta) + \tag{7}$$
$$\sum_{h=2}^{T} \left(\gamma_h \sum_{d} \sum_{n,n'} \mathbf{1}_{par^{(h)}(w^{(1)}_{d,n})!=par^{(h)}(w^{(1)}_{d,n'})} \mathbf{1}_{z^{(1)}_{d,n}=z^{(1)}_{d,n'}} \right)$$

Dummy Document However, the objective is not easy to optimize. Thus we introduce an equivalent way to incorporate the hierarchy penalty by including dummy documents for each parent ADR node in the training process. The dummy document represents a parent ADR node consisting of all its child ADRs at the bottom layer. The existence of dummy documents will enforce higher probability of assigning child ADRs into the same group. The penalty weight can be equivalently tuned by adjusting number of dummy documents or word counts in the LDA model. A simple verification has been performed to train LDA only on dummy documents and we observed that ADRs under lower hierarchy are grouped into one topic, while only parent node at a higher hierarchy consists more than one topic. The graphical structure of the regularized model is illustrated as in Figure 2.

Mathematically, the objective of the regularized model can be re-formulated as follows. Denote $\{D+1, \ldots, D'\}$ as dummy documents, we want to minimize the following penalized negative log-likelihood function.

$$-\ell_{ext}(\theta_{1:D'}, \beta_{1:K}; \alpha, \eta) = -\ell(\theta_{1:D}, \beta_{1:K}; \alpha, \eta) -$$
$$\gamma \sum_{d=D+1}^{D'} \log p(\theta_d, w_d | \alpha, \beta_{1:K}) \tag{8}$$

For non-hierarchical information, we can still add penalty terms like Laplacian matrices to induce certain ADRs to be grouped together.

3.3 Mixed Input Model

In real life, biomedical data such as drug and ADR data are high dimensional. Certain method (e.g. lasso) that requires training and prediction on each ADR will require an external loop to iterate over ADR dimension to determine hyperparameters. It will take a long time to complete even on a medium size dataset. To alleviate it, we further modify the base LDA with mixed inputs. Instead of learning the hidden topics separately, we treat both ADRs and drug structure features as words for a drug document so that each learned topic will potentially contain a subset of both ADRs and drug features. Figure 3 illustrates the graphical structure of the mixed input design.

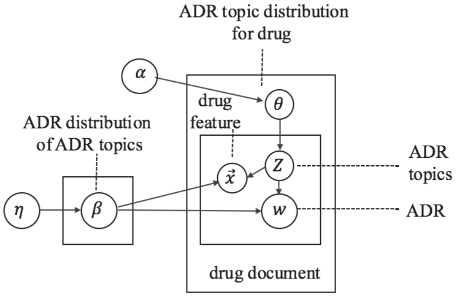
ADR topic distribution for drug
α
drug feature
θ
ADR distribution of ADR topics
$\vec{x}$
Z
ADR topics
η
β
w
ADR
drug document

Figure 3: The mixed input model

The mixed input model adopts a generative approach and hence significantly speeds up learning with a proper training method. (Porteous et al. 2008) proposed a fast collapsed Gibbs sampling method with the complexity of each iteration of Gibbs sampling proportional to the number of positive ADRs instead of total candidate ADRs. The learning process for the mixed input model effectively partitions ADRs and drug features into separate groups. The partitioning could be interpreted as each group represents certain mechanism that could separate the cause (features) and effect (ADRs) out from other mechanisms.

Specifically, we revise the objective function for the mixed input model as follows. Denote the number of ADRs as N and the number of features as p, the revised marginal distribution of drug d will be as in Formula 9.

$$p(\theta_d, \tilde{w}_d | \alpha, \beta_{1:K}) = p(\theta_d|\alpha) \cdot$$
$$\prod_{n=1}^{N+p} \left(\sum_{z_{d,n}} p(z_{d,n}|\theta_d) p(\tilde{w}_{d,n}|z_{d,n}, \beta_{1:K}) \right) \tag{9}$$

where $\tilde{w}_d = [w_d, \vec{x}_d]$. To make prediction for a target drug d', we still need to determine the topic distribution $\theta_{d'}$ based on the drug structure features $\vec{x}_{d'}$ and observed drug corpus. What we do is to set $w_{d'} = 0$ and co-train it with the corpus, then make inference based on $\theta_{d'}$.

3.4 Inference in the Model via Gibbs Sampling

For parameter estimation, we adopt the fast collapsed Gibbs sampling algorithm. The objective is to evaluate the posterior distribution $p(z|w)$. From the posterior distribution, we can compute the statistics such as θ_d and β_k. We calculate the posterior distribution $p(z|w)$ using the Bayes rule.

$$p(z|w) = \frac{p(w,z)}{\sum_z p(w,z)},$$

where the joint distribution $p(w,z) = p(w|z)p(z)$ could be obtained by integrating out β and θ separately from $p(w|z)$ and $p(z)$. Here the posterior $p(w|z)$ is obtained from Dirichlet prior $p(\beta|\eta)$ and multinomial likelihood $p(w|z,\beta)$ such that:

$$p(w|z) = \left(\frac{\Gamma(N\eta)}{\Gamma(\eta)^N}\right)^K \prod_{k=1}^{K} \frac{\prod_w \Gamma(n_k^{(w)} + \eta)}{\Gamma(n_k^{(\cdot)} + N\eta)} \tag{10}$$

where $n_k^{(w)}$ refers to the number of times ADR(word) w has been assigned to a ADR topic k. Likewise, $p(z)$ is given by Dirichlet prior $p(\theta|\alpha)$ and multinomial likelihood $p(z|\theta)$ such that:

$$p(z) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^D \prod_{d=1}^{D} \frac{\prod_k \Gamma(n_k^{(d)} + \eta)}{\Gamma(n_{\cdot}^{(d)} + K\eta)} \tag{11}$$

where $n_k^{(d)}$ refers to the number of times a word from document d has been assigned to topic k.

In addition, for the task of prediction, the predictive probability is given by the following formula.

$$\begin{aligned} p(w_{d,n}) &= \sum_k p(w_{d,n}|z_{d,n} = k)p(z_{d,n} = k) \\ &= \sum_k (\beta_k(w_{d,n}) + \eta)(\theta_d(k) + \alpha) \end{aligned}$$

4 Experiment

4.1 Data Sources

We used ADReCS database (Cai et al. 2015) in evaluation. The drug-ADR information of ADReCS was mainly extracted from the drug labels in the DailyMed[1], a website managed by the U.S. National Library of Medicine (NLM) to provide comprehensive information about marketed drugs. ADReCS adopted a four level ADR hierarchy tree with the System Organ Class, the High level Group Term, the High Level Term and the Preferred Term. As the hierarchy goes to higher level, the ADR terms become more abstract and generic in describing medical concept. We preprocessed the data to filter out ADR with fewer than 17 associations (due to R lasso package cannot handle the case of too few positive samples. 17 happens to be minimum number not incur error during cross-validation loops). We did the similar preprocessing on drug data. In the experiment, we have 996 ADRs and 1000 drugs, each with a 1024-dimension feature vector.

Table 2: Performance Comparison

Model	AUROC (± SD)	PR-AUC (± SD)
lasso	0.830 ± 0.063	0.420 ± 0.188
CCA	0.659 ± 0.119	0.245 ± 0.176
LDA (basic)	0.836 ± 0.055	0.452 ± 0.190
LDA (regularized)	**0.840 ± 0.055**	**0.453 ± 0.188**
LDA (mixed input)	0.838 ± 0.054	0.445 ± 0.184

Table 3: Comparison of Running Time

Model	Running Time (hours)
lasso	10.0 ∼ 12.0
CCA	1.0 ∼ 1.2
LDA (basic)	7.8 ∼ 8.0
LDA (regularized)	7.8 ∼ 8.0
LDA (mixed input)	**0.4 ∼ 0.5**

4.2 Evaluation Method

We performed 20-fold cross validation to evaluate our models against baseline methods including lasso and CCA. Specifically, in each iteration, 95% of the training drug was used to construct the models and the remaining 5% of the drug was used for performance testing. We also tune the number of topics K from 20 to 140 with 20 per increment, and select one that gives the best performance. In all cases, we have $K = 100$. The measures we use include area under the receiver operating characteristic curve (ROC-AUC or AUROC) and area under the precision recall curve (PR-AUC).

4.3 Results

We processed the data using Python package "pandas (McKinney 2015)", as well as evaluated the algorithms using R packages "glmnet (Friedman, Hastie, and Tibshirani 2010)", "cca (González et al. 2008)", and "lda (Chang 2015)", respectively. Table 2, Figure 4 and 5 compare the validation ROC-AUC (AUROC) and PR-AUC, while in Table 3 we also compare the algorithm running time. The results show that on average, our models significantly outperform baseline in ROC-AUC and PR-AUC. Among them, the regularized model that incorporate domain knowledge and the mixed input model are among top performers. The mixed input model also has advantage in fast running time. Lasso, though has decent prediction performance, runs almost half day. Therefore, from empirical results, it is easy to see our models, especially the mixed input one has substantial improvement in all perspectives.

4.4 Case Study

We also use case study to demonstrate the performance and interpretability of the models. Taking Meloxicam, a nonsteroidal anti-inflammatory drug (NSAID) as an example. Our models outperform baselines (see Figure 6 and 7) and

[1] https://dailymed.nlm.nih.gov/

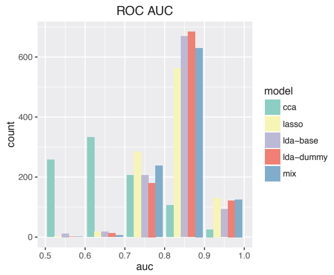

Figure 4: ROC AUC across all drugs.

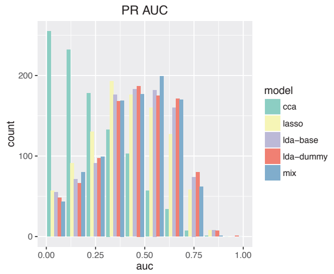

Figure 5: PR AUC across all drugs.

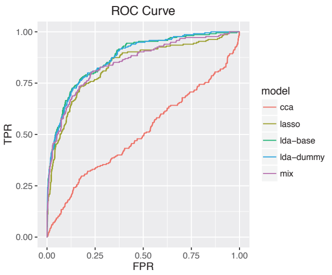

Figure 6: ROC AUC for Meloxicam

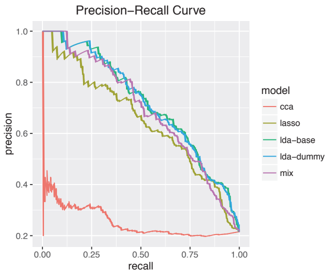

Figure 7: PR AUC for Meloxicam

correctly predict most of its major known ADRs. Figure 8 shows the selected topics and ADRs with top likelihood. The true positive, false positive and false negative predictions are shaded with different patterns, the ADR terms are listed with their ADReCS hierarchical indexes, and the dotted boxes group ADRs together according to the hierarchy in ADReCS. The selected ADR topics could be mapped to ADR hierarchy in the ADReCS ontology database. For example, gastrointestinal signs and symptoms is a mixture of nausea, vomiting and abdominal pain.

From the false positive predictions, we discover that diarrhea, although not labeled as one ADR of Meloxicam, has been predicted as a potential one due to being a sibling of many ADRs of Meloxicam. Since many known ADRs in gastrointestinal disorders are close related to it, it is reasonable to believe that diaherra could be of the ADRs of Meloxicam. In addition, real world evidence from online forums such as (WedMD 2016) and (Peoplespharmacy 2016) also indicate that patients took Meloxicam suffered from severe diarrhea.

One example of false negative prediction is Mydriasis, i.e. dilation of the pupil of the eye. Our algorithm did not predict this as an ADR of Meloxicam. First of all, the condition is quite rare. Secondly, there could be confounding factors such that the ADR could be caused by a often co-prescribed drug. Or as a NSAID, Moxicam could be a medication for patients undergone ocular surgeries, in which case, the surgeries themselves could be confounding factors.

These analysis suggest that incorporating of real-world evidence could be a direction to improve the medical ontology system and bring the drug safety studies into a better iteration. Moreover, the identification of drug-ADR relation sometimes is misled by confounding factors, while in the future we could adding causal inference to distinguish causation of an ADR from a correlation between drugs and ADRs, thus to remove confounding factors.

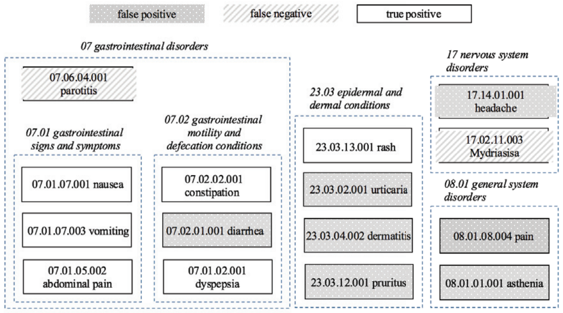

Figure 8: Selected false positive and false negative ADR predictions for drug Meloxicam.

5 Conclusion and Future Work

In this paper, we presented three LDA-based models for ADR prediction. The approach learns a hidden topic layer that may relate to biochemical mechanisms that link drug structures to ADRs. Moreover, the mixed input model has the best combination of prediction performance and training efficiency. Experiments show that all models have better prediction accuracy than baselines. Furthermore, we analyze the positive and false negative predictions based on a sample drug and the results point to some future directions including 1) improve medical ontology system with real world evidence, and 2) removing confounding factors with causal inference. Both directions and the link between topics and biomedical mechanisms will be studied in our future work.

6 Acknowledgment

Fei Wang is partially supported by National Science Foundation under Grant Number IIS-1650723.

References

Altman, N. S. 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* 46(3):175–185.

Bender, A.; Scheiber, J.; and Glick, M. 2007. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* 2:861–73.

Biro, I.; Siklosi, D.; Szabo, J.; and Benczur, A. 2009. Linked latent dirichlet allocation in web spam filtering. In *Proceedings of the 5th Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb).*

Bisgin, H.; Liu, Z.; Fang, H.; Xu, X.; and Tong, W. 2011. Mining fda drug labels using an unsupervised learning technique - topic modeling. *BMC Bioinformatics.* 12:s11.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Cai, M.; Xu, Q.; Pan, Y.; Pan, W.; Ji, N.; Li, Y.; Jin, H.; Liu, K.; and Ji, Z. 2015. Adrecs: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. *Nucleic Acids Research* 43(1):D907–13.

Cao, L., and Li, F. 2007. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV 2007).*

Caster, O. 2007. Mining the who drug safety database using lasso logistic regression.

Chang, J. 2015. Package 'lda'.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22.

Giacomini, K.; Krauss, R.; Roden, D.; Eichelbaum, M.; and Hayden, M. 2007. When good drugs go bad. *Nature* 446:975–977.

González, I.; Déjean, S.; Martin, P.; and Baccini, A. 2008. Cca: An r package to extend canonical correlation analysis. *Journal of Statistical Software* 23(1):1–14.

Guha, R. 2007. Chemical informatics functionality. *R. Journal of Statistical Software.* 6:18.

Hotelling, H. 1936. Relations Between Two Sets of Variates. *Biometrika* 28:321–377.

Li, R.; Dong, Y.; Kuang, Q.; and Wu, Y. 2015. Inductive matrix completion for predicting adverse drug reactions integrating drug–target interactions. *Chemometrics and Intelligent Laboratory Systems* 144.

Liu, M.; Wu, Y.; Chen, Y.; Sun, J.; Zhao, Z.; Chen, X.-w.; Matheny, M. E.; and Xu, H. 2012. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association* 19(e1):e28–e35.

McCune, B. 1997. Influence of noisy environmental data on canonical correspondence analysis. *Ecology* 78(8):2617–2623.

McKinney, W. 2015. pandas: a foundational python library for data analysis and statistics.

Mei, Q.; Liu, C.; Su, C.; and Zhai, c. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th Int. World Wide Web Conference (WWW'06).*

Paul, M., and Dredze, M. 2012. Experimenting with drugs (and topic models): Multi-dimensional exploration of recreational drug discussions. *AAAI Technical Report FS-12-05 Information Retrieval and Knowledge Discovery in Biomedical Text.*

Pauwels, E.; Stoven, V.; and Yamanishi, Y. 2011. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics* 12(1):1–13.

Peoplespharmacy. 2016. http://www.peoplespharmacy.com /2013/01/31/meloxicam-mobic-side-effects-complications/.

Porteous, I.; Newman, D.; Ihler, A.; Asuncion, A.; Smyth, P.; and Welling, M. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, 569–577.

Rogers, D., and Hahn, M. 2010. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50:742–754.

Scheiber, J.; Jenkins, J.; and Sukuru, S. 2009. Mapping adverse drug reactions in chemical space. *Journal of Medicinal Chemistry* 52:3103–7.

Tibshirani, R. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

WedMD. 2016. http://www.webmd.com/drugs/drugreview-911-Meloxicam+Oral.aspx.

Whitebread, S.; Hamon, J.; Bojanic, D.; and Urban, L. 2005. In vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discovery Today* 10(21).

WHO-ART. 2015. The WHO Adverse Reaction Terminology. *http://www.umc-products.com/graphics/28010.pdf.*

Yamanishi, Y.; Pauwels, E.; and Kotera, M. 2012. Drug side-effect prediction based on the integration of chemical and biological spaces. *Journal of Chemical Information and Modeling* 52(12):3284–3292.