Unsupervised Discovery of Drug Side-Effects from Heterogeneous Data Sources

Fenglong Ma
SUNY Buffalo
Buffalo, NY, USA
fenglong@buffalo.edu

Chuishi Meng
SUNY Buffalo
Buffalo, NY, USA
chuishim@buffalo.edu

Houping Xiao
SUNY Buffalo
Buffalo, NY, USA
houpingx@buffalo.edu

Qi Li
SUNY Buffalo
Buffalo, NY, USA
qli22@buffalo.edu

Jing Gao, Lu Su
SUNY Buffalo
Buffalo, NY, USA
jing@buffalo.edu,lusu@buffalo.edu

Aidong Zhang
SUNY Buffalo
Buffalo, NY, USA
azhang@buffalo.edu

ABSTRACT

Drug side-effects become a worldwide public health concern, which are the fourth leading cause of death in the United States. Pharmaceutical industry has paid tremendous effort to identify drug side-effects during the drug development. However, it is impossible and impractical to identify all of them. Fortunately, drug side-effects can also be reported on heterogeneous platforms (i.e., data sources), such as FDA Adverse Event Reporting System and various online communities. However, existing supervised and semi-supervised approaches are not practical as annotating labels are expensive in the medical field. In this paper, we propose a novel and effective unsupervised model Sifter to automatically discover drug side-effects. Sifter enhances the estimation on drug side-effects by learning from various online platforms and measuring platform-level and user-level quality simultaneously. In this way, Sifter demonstrates better performance compared with existing approaches in terms of correctly identifying drug side-effects. Experimental results on five real-world datasets show that Sifter can significantly improve the performance of identifying side-effects compared with the state-of-the-art approaches.

CCS CONCEPTS

•Information systems → Data mining; •Applied computing → Health informatics;

KEYWORDS

Healthcare informatics, truth discovery, probabilistic graphical model, drug side-effects

1 INTRODUCTION

Drug side-effects or adverse drug events (ADEs), defined as harmful or unpleasant reactions resulted from drug related medical interventions, are a worldwide public health concern. According to the report from the U.S. Department of Health and Human Services in 2014[1], drug side-effects caused about one third of hospital adverse events and 280,000 hospital admissions on average annually. Moreover, serious drug side-effects are the fourth leading cause of death in the U.S., resulting in about 100,000 deaths per year [8]. Thus, it is imperative to discover unreported side-effects.

The U.S. Food and Drug Administration uses the Adverse Event Reporting System (FAERS)[2] to monitor post-market usage of drugs in order to discover unknown drug side-effects. It is a voluntary system where doctors, patients, and pharmacists report unpleasant reactions. However, a significant portion of the ADEs have not been reported by FAERS [11]. Fortunately, drug side-effects are also studied, recorded and discussed in various unofficial platforms (i.e., data sources), such as biomedical literatures, clinical documents, electronic health records, and online healthcare forums. From these sources, we can discover drug side-effects not reported by FAERS.

Several supervised and semi-supervised approaches have been proposed to detect drug side-effects from the aforementioned platforms [1, 4, 9, 12, 13, 25, 40]. For semi-supervised and supervised methods, classifiers are trained based on the features and labels related to the drugs [13, 25]. However, in the medical field, it is very expensive to annotate sufficient data to train a good classifier. Thus, a practical problem in drug side-effect discovery is *how to automatically identify drug side-effects in an unsupervised way.*

The simplest way to identify drug side-effects is to set a threshold for the number of reported ADEs so that the side-effects will be recognized if it exceeds this threshold. However, this approach fails to take reporters' (e.g., users of a medical forum) reliability degree into consideration, and thus would lead to poor performance when there is a large number of low quality users. To solve this problem, the idea of *truth discovery* [18] can be borrowed. Truth discovery methods aim to infer the true information and learn users' reliability degrees simultaneously. Researchers in this area focus on two fundamental questions: single truth discovery [5–7, 16, 17, 19, 23, 24, 26, 27, 29, 31, 36, 38, 41, 44] and multiple truth discovery [28, 32, 33, 47], in which single truth discovery assumes that there is only one correct value for each object, while multiple truth discovery allows multiple correct values. Since there may be more than one side-effects for a drug, drug side-effect discovery

KDD'17, August 13–17, 2017, Halifax, NS, Canada.

DOI: http://dx.doi.org/10.1145/3097983.3098129

[1] http://health.gov/hcq/pdfs/ade-action-plan-508c.pdf
[2] http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm

can be considered as a multiple truth discovery problem. However, identifying drug side-effects is quite different from traditional truth discovery in the following aspects.

Truth Discovery for Single Object with Multiple Claims. Traditional truth discovery methods require users providing single claim for each object and claiming on multiple objects. However, for drug side-effect discovery problem, we focus on one drug's side-effects, i.e., *single object*. Moreover, users report multiple side-effects (i.e., *multiple claims*) for the given drug. These lead to the fact that existing methods may learn unreasonable user quality and infer incorrect side-effects.

Inferring Truth from Multiple Platforms. Drug's side-effects can be collected or extracted from many platforms, but existing truth discovery methods do not consider platform information. We use the following example to illustrate the importance of considering multiple platforms' information. Table 1 shows an example of Thyroxine's side-effects extracted from FAERS and Healthboards[3]. Each entry denotes a claim, i.e., a potential side-effect provided by a user. *Dysphagia*, *Nausea* and *Dehydrated* are the true side-effects and the other three are incorrect ones. If only mining side-effects from single platform, such as FAERS, we can obtain at most two correct side-effects: *Dysphagia* and *Nausea*. However, using data from both FAERS and Healthboards, we can obtain all the three correct side-effects. Consequently, taking multiple platforms into account helps us to obtain more true side-effects.

Table 1: An Example of Thyroxine Dataset.

FAERS		Healthboards	
User ID	Side-Effect	User ID	Side-Effect
110696642	Dysphagia	2918	Migraine
108294651	Dysphagia	3171	Dysphagia
108294651	Nausea	3171	Nausea
108294651	Mood	3171	Anemia
108325471	Dysphagia	6871	Mood
108325471	Nausea	6871	Dehydrated
108325471	Migraine	27417	Dehydrated

Learning Quality for Different Platforms and Users. The data quality of different platforms contributed by users may be different. In Table 1, the overall quality of FAERS (5 correct claims) is better than Healthboards' (4 correct ones). However, only employing the quality of platforms to infer side-effects is insufficient. For example, the low quality of Healthboards will decrease the probability of *Dehydrated* being estimated as truth and lead to an incorrect result. Therefore, we should also consider the quality of users in platforms when discovering drug side-effects.

Two-Sided Quality of Platforms and Users. Most truth discovery approaches assign a single quality for each user according to the trustworthiness of claims (i.e., the number of correct claims), which is not enough for drug side-effect discovery. For example, based on the claims explicitly provided by users, named *positive claims*, *Dysphagia* and *Nausea* can be easily estimated as true information with FAERS data in Table 1. Correspondingly, User 108294651 and User 108325471 will be assigned high reliability degrees, which leads to the increase of the probabilities of *Mood* and *Migraine* estimated as truth. However, they are not true side-effects. To reduce the probabilities, we need to consider the side-effects not claimed by the users or platforms, i.e., the *negative claims*. Since user 110696642 and 108325471 do not report *Mood* in FEARS data, the probability of *Mood* being true side-effect may be decreased. Thus, modeling two-sided platform-level and user-level quality (i.e., considering both *positive* and *negative* claims simultaneously) is the key design for multiple truth discovery.

To solve the aforementioned challenges, in this paper, we propose a novel and effective unsupervised model Sifter for drug **si**de-e**f**fec**t** discov**er**y problem. It can simultaneously estimate the correct side-effects for given drugs and learn platform-level and user-level reliability degrees (i.e., information quality) based on the side-effects provided by users. To the best of our knowledge, Sifter is the first unsupervised model to leverage the quality of platforms and users for multiple truth discovery. By treating the truth as a latent random variable, the proposed model can naturally model both positive and negative claims, and learn platform-level and user-level quality in a principled way. We also propose an efficient inference algorithm based on collapsed Gibbs sampling to estimate correct side-effects. The experiments on five real-world datasets show that the proposed Sifter can significantly improve the performance of identifying drug side-effects compared with the state-of-the-art truth discovery approaches.

In the following sections, we formulate the problem formally in Section 2. In Section 3, we introduce the proposed model in details, and the inference processes are discussed in Section 4. The experimental results are shown on four real-world datasets in Section 5. Section 6 gives a brief overview on the related work in drug side-effect discovery and truth discovery. Finally, we conclude the paper in Section 7.

2 PROBLEM FORMULATION

In this section, we introduce some basic terminologies used in this paper and then define our problem formally.

Input

The inputs of the proposed model include the platform set, the user set in each platform, and the claims provided by users.

Definition 2.1. A platform $s \in \{1, \ldots, S\}$ is a database containing potential side-effects of drugs, where S is the number of platforms.

Definition 2.2. In each platform s, there are a set of users $\{u\}_1^{U_s}$ who provide potential side-effects for drugs, where U_s is the number of users in s.

FDA Adverse Event Reporting System (FAERS) can be seen as a platform, in which all the data are structured. In FAERS, each reporter or patient can be regarded as a user. Since we focus on ***single drug side-effect discovery***, side-effects are collected when patients only take the given drug (i.e., one drug). Online forum Healthboards can be seen as another platform where the data are unstructured. On Healthboards, each user can write posts, answer questions and communicate with others. Here we assume that users are independent when writing or reporting drugs' side-effects, and there are no common users among platforms.

Since forum users are not experts in the medical area, side-effects collected from these users are colloquial and informal. In order to

[3]http://www.healthboards.com

model these side-effects reasonably and fairly, we use MetaMap[4], a natural language processing tool for recognizing medical concepts in raw text, to convert all the potential side-effects collected from different platforms into Concept Unique Identifier (CUI) codes[5], such as $Nausea \rightarrow C0027497$. The benefit of using CUIs is that different expressions of side-effects can be mapped into the same CUI code. In this way, we can extract structured representations of side-effects collected from both FAERS and Healthboards.

In Section 1, we already introduced the benefit of considering both positive and negative claims when inferring true side-effects. There is a challenge that has to be addressed, i.e., the number of negative claims for each user may be large as users usually provide only a few side-effects for one drug. To decrease the effect caused by negative claims, we randomly sample negative claims with no replacement to ensure there are comparable numbers of positive and negative claims. Here we use observations to denote the values of claims defined as follows:

Definition 2.3. An observation o^s_{um} ($m \in \{1, \ldots, M\}$) provided by user u in platform s is a Boolean value True or False, where M is the total number of side-effects for a given drug. If user u reports the m-th side-effect (i.e., positive claim), then the corresponding observation o^s_{um} is 1; if this is a negative claim, $o^s_{um} = 0$; otherwise, o^s_{um} does not exist.

In Table 1, there are 6 side-effects, i.e., $M = 6$. User 6871 in platform Healthboards reports side-effects *Mood* and *Dehydrated*, but *Dysphagia*, *Nausea*, *Migraine* and *Anemia* are not reported. We randomly sample two potential side-effects, for example *Dysphagia* and *Anemia*, as negative claims. Correspondingly, the observations provided by User 6871 are shown in Table 2.

Table 2: Observations of User 6871.

Platform	User ID	Side-Effect	CUI	Observation
Healthboards	6871	Dysphagia	C0011168	False
Healthboards	6871	Mood	C0026516	True
Healthboards	6871	Anemia	C0002871	False
Healthboards	6871	Dehydrated	C0001721	True

Output

Given input raw data from multiple platforms, the proposed model aims to derive true side-effects for a given drug. We define truth indicators as follows:

Definition 2.4. Truth indicators $\{t_m\}_{m=1}^M$ are used to denote the correctness of side-effects, where t_m is a Boolean value, i.e., True or False. If the final output of the m-th side-effect t_m is True, i.e., $t_m = 1$, then it is considered to be a true side-effect of the given drug; otherwise, it is not a side-effect.

Besides inferring true side-effects for a given drug, we also want to automatically estimate the *quality* of different platforms and users in platforms. Platform-level quality information can be seen as a global indicator to measure how reliable each platform is for a certain drug, and user-level quality information can be seen as a local indicator to evaluate how credible each user is.

[4]https://metamap.nlm.nih.gov
[5]https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html

In order to model platform- and user-level quality, we first introduce the confusion matrix of platforms or users as shown in Table 3. *Precision* is defined as the probability of positive observations being correct, i.e., $\frac{TP_s}{TP_s+FP_s}$. *Sensitivity* or *Recall* is the probability of true observations being estimated as truth, i.e., $\frac{TP_s}{TP_s+FN_s}$. $Specificity$ is the probability of false observations not being estimated as truth, i.e., $\frac{TN_s}{TN_s+FP_s}$, and the *False Positive Rate* (*FPR*) is 1 - *Specificity*.

Table 3: Confusion Matrix.

	$t = True$	$t = False$
$o = True$	True Positives (TP_s)	False Positives (FP_s)
$o = False$	False Negatives (FN_s)	True Negatives (TN_s)

From Table 1, we can observe that the platform FAERS has a high $precision$ but low $recall$ compared with Healthboards. Obviously, we need to take both $precision$ and $recall$ into consideration when measuring platforms' quality. However, a drawback of modeling platform-level quality with $precision$ is that it ignores negative observations. In order to take both positive and negative observations into account, we use *False Positive Rate* and *Sensitivity* to characterize platform-level quality and user-level quality. *False Positive Rate* is associated with false positives and true negatives, and $Sensitivity$ is related to false negatives and true positives. With these two measures, we are able to characterize the complete spectrum of platform- and user-level quality.

Definition 2.5. *False Positive Rate* and *Sensitivity* are used to measure the reliability degrees of platforms and users. *False Positive Rate* $\{\phi^0_s\}_{s=1}^S$ and *Sensitivity* $\{\phi^1_s\}_{s=1}^S$ are assigned to each platform s. User-level quality $\{\theta^0_{su}\}_{s=1,u=1}^{S,U_s}$ (*False Positive Rate*) and $\{\theta^1_{su}\}_{s=1,u=1}^{S,U_s}$ (*Sensitivity*) are learned for each user u in each platform s.

Intuitively, whether an observation o^s_{um} is a true side-effect is determined by platform-level quality and user-level quality. In order to make the proposed model more general and simple, we define a reliability indicator as follows:

Definition 2.6. Reliability indicator y_m is used to denote whether the observation o^s_{um} is related to platform-level or user-level quality. If $y_m = 0$, then platform-level quality is used; otherwise, user-level quality is selected.

Based on these definitions, we can formally define our problem as follows: Given a platform set $\{s\}_1^S$, user set $\{u\}_1^{U_s}$ of each platform s, side-effect set $\{m\}_1^M$, and observations $\{o^s_{um}\}_{m=1}^M$ provided by users, our goal is to learn platform-level reliability degrees $\{\phi^0_s\}_{s=1}^S$ and $\{\phi^1_s\}_{s=1}^S$, user-level reliability $\{\theta^0_{su}\}_{s=1,u=1}^{S,U_s}$ and $\{\theta^1_{su}\}_{s=1,u=1}^{S,U_s}$, truth indicators $\{t_m\}_{m=1}^M$, and reliability indicators $\{y_m\}_{m=1}^M$ for drug side-effects.

3 SIFTER MODEL

In this section we formally introduce the proposed model, called Sifter, for discovering drug side-effects from various platforms. We first briefly summarize the proposed model Sifter, and then provide details about the proposed model.

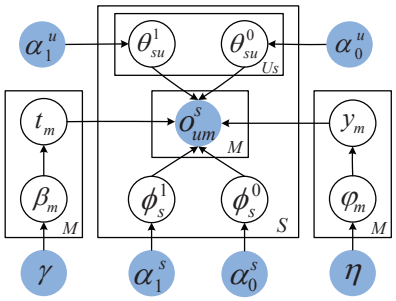

Figure 1: The Probabilistic Graphical Model of Sifter.

3.1 Model Overview

In contrast to existing methods in multiple truth discovery, we consider the differences among diverse platforms and learn platform-level reliability degrees. Within each platform, we learn user-level reliability degrees based on observations provided by users. Taking both platform-level and user-level reliability degrees into consideration, we estimate correct side-effects of drugs using observations reported by users from different platforms.

Figure 1 shows the proposed probabilistic graphical model for discovering drugs' side-effects. The inputs are S platforms, M side-effects of a given drug, $\{U_s\}_{s=1}^{S}$ users, and their corresponding observations $\{o_{um}^s\}_{s=1,u=1,m=1}^{S,U_s,M}$. The shaded circles represent hyper-parameters[6] except o_{um}^s. The outputs are truth indicators $\{t_m\}_{m=1}^{M}$, reliability indicators $\{y_m\}_{m=1}^{M}$, two-sided platform reliability degrees $\{\phi_s^0\}_{s=1}^{S}$ and $\{\phi_s^1\}_{s=1}^{S}$, and two-sided user quality $\{\theta_{su}^0\}_{s=1,u=1}^{S,U_s}$ and $\{\theta_{su}^1\}_{s=1,u=1}^{S,U_s}$. The remaining $\{\beta_m\}_{m=1}^{M}$ and $\{\varphi_m\}_{m=1}^{M}$ are the intermediate variables learned by the proposed model. The detailed generating process is introduced in the following subsections.

3.2 Truth & Reliability Indicator Generation

Since Sifter is an unsupervised model, each side-effect being true or false is unknown. We model the probability of each side-effect being true as a latent Boolean random variable. If the truth indicator is 1, then the side-effect is correct; otherwise, the drug cannot cause this side-effect. Moreover, the truth indicator is a switch to select *False Positive Rate* or *Sensitivity* used in the process of generating observations provided by users. In addition, the proposed model allows to set prior distributions on the truth probability. The truth indicator generating process is as follows.

We first generate prior probability β_m from a Beta distribution with parameter $\boldsymbol{\gamma} = (\gamma_1, \gamma_0)$, where γ_1 denotes the prior true pseudocount, and γ_0 is the prior false pseudocount for each side-effect m:

$$\beta_m \sim Beta(\gamma_1, \gamma_0).$$

Based on the prior distribution β_m, the truth indicator t_m (a Boolean variable) can be generated from a Bernoulli distribution:

$$t_m \sim Bernoulli(\beta_m).$$

Note that if there is no prior belief on side-effects, we can use a uniform priori.

[6] $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, $\boldsymbol{\alpha_0^u}$, $\boldsymbol{\alpha_1^u}$, $\boldsymbol{\alpha_0^s}$, and $\boldsymbol{\alpha_1^s}$ denote hyper-parameters of Beta distributions.

The reliability indicator y_m is used to select quality measure (platform-level or user-level) when generating observations. We model the probability of user-level quality being selected as a latent Boolean random variable. A Beta distribution with parameter $\boldsymbol{\eta} = (\eta_1, \eta_0)$ is used to generate a prior probability φ_m, where η_1 denotes the prior pseudocount of the side-effect generated with user-level quality, and γ_0 is the prior pseudocount for each side-effect m generated with platform-level quality.

$$\varphi_m \sim Beta(\eta_1, \eta_0).$$

The reliability indicator y_m is generated from a Bernoulli distribution according to the prior distribution φ_m:

$$y_m \sim Bernoulli(\varphi_m).$$

3.3 Platform-Level Quality Generation

Considering the differences among platforms, we generate platform-level quality for each platform. The quality of platforms affects side-effects' trustworthiness. If users in a platform s seldom provide erroneous side-effects, then platform s has a low *False Positive Rate*. Thus, the probability of a side-effect provided by a user in platform s being true is high. On the other hand, if platform s provides most of the true side-effects, it has a high *Sensitivity*. It leads to a larger probability for a side-effect being false if it is not posted by s.

We use two independent factors - *False Positive Rate* and *Sensitivity* - to characterize the quality of platforms, and two separate random variables are introduced to describe them. Moreover, we may have prior belief or assumptions with regard to each platform in practice. For example, users in the FAERS dataset typically provide correct side-effects, i.e., the *Sensitivity* of this platform should be high. On the contrary, users in online healthcare communities tend to provide noisy answers. Thus, they may have high *False Positive Rate* (i.e., low *Specificity*). In these cases, the model allows us to set such prior belief that characterizes platforms reasonably.

The generating process of platform-level quality is as follows. For each platform $s \in S$, we generate its *False Positive Rate* ϕ_s^0 from a Beta distribution with hyperparameter $\boldsymbol{\alpha_0^s} = (\alpha_{0,1}^s, \alpha_{0,0}^s)$, where $\alpha_{0,1}^s$ denotes the prior false positive pseudocount, and $\alpha_{0,0}^s$ is the prior true negative pseudocount for platform s:

$$\phi_s^0 \sim Beta(\alpha_{0,1}^s, \alpha_{0,0}^s).$$

Similar with the generation of ϕ_s^0, the *Sensitivity* of s, ϕ_s^1, is drawn from a Beta distribution with hyperparameter $\boldsymbol{\alpha_1^s} = (\alpha_{1,1}^s, \alpha_{1,0}^s)$, where $\alpha_{1,1}^s$ denotes the prior true positive pseudocount, and $\alpha_{1,0}^s$ is the prior false negative pseudocount for platform s:

$$\phi_s^1 \sim Beta(\alpha_{1,1}^s, \alpha_{1,0}^s).$$

3.4 User-Level Quality Generation

Similar with the generating process of platform-level quality, for each user u in platform s, we generate the user's *False Positive Rate* θ_{su}^0 from a Beta distribution with hyperparameter $\boldsymbol{\alpha_0^u} = (\alpha_{0,1}^u, \alpha_{0,0}^u)$ and the user's *Sensitivity* θ_{su}^1 from a Beta distribution with hyperparameter $\boldsymbol{\alpha_1^u} = (\alpha_{1,1}^u, \alpha_{1,0}^u)$ as follows:

$$\theta_{su}^0 \sim Beta(\alpha_{0,1}^u, \alpha_{0,0}^u),$$

and

$$\theta^1_{su} \sim Beta(\alpha^u_{1,1}, \alpha^u_{1,0}),$$

where $\alpha^u_{0,1}$ denotes the prior false positive pseudocount, $\alpha^u_{0,0}$ is the prior true negative pseudocount, $\alpha^u_{1,1}$ denotes the prior true positive pseudocount, and $\alpha^u_{1,0}$ is the prior false negative pseudocount.

3.5 Observation Generation

According to the above analysis, each observation o^s_{um} from user u in platform s is associated with a truth indicator t_m, a reliability indicator y_m, platform-level quality, and user-level quality.

For the observation o^s_{um}, truth indicator t_m indicates that *False Positive Rate* ($t_m = 0$) or *Sensitivity* ($t_m = 1$) is employed to generate observations. If $y_m = 1$, the observation is generated with user-level quality, and the generating process of observation o^s_{um} is as follows:

$$o^s_{um} \sim Bernoulli(\theta^{t_m}_{su}).$$

If $y_m = 0$, o^s_{um} is generated based on platform-level reliability degrees:

$$o^s_{um} \sim Bernoulli(\phi^{t_m}_{s}).$$

The generating process is quite different from that of existing truth discovery methods. First, we draw the claim o^s_{um} from a Bernoulli distribution with the parameters $\theta^{t_m}_{su}$ and $\phi^{t_m}_s$, which shows that different types of reliability degrees are considered when generating claims. However, traditional truth discovery methods only consider user-level quality measure. Moreover, we assign macroscopical reliability degrees to platforms, which is also different from existing methods. Existing methods do not distinguish the characteristic of platforms and treat all the platforms equally.

4 INFERENCE AND LEARNING

In this section, we present the joint likelihood function of the proposed model Sifter and discuss how to perform inference to estimate the truth of side-effects, and the quality of platforms and users from Sifter, given the observations.

4.1 Joint Likelihood Function

According to the generative process in Section 3, given the Sifter parameters, the probability of each observation o^s_{um} provided by user u in platform s is:

$$p(o^s_{um}|t_m, y_m, \theta^0_{su}, \theta^1_{su}, \phi^0_s, \phi^1_s) = y_m[p(o^s_{um}|\theta^0_{su})(1-t_m) + p(o^s_{um}|\theta^1_{su})t_m] + (1-y_m)[p(o^s_{um}|\phi^0_s)(1-t_m) + p(o^s_{um}|\phi^1_s)t_m].$$

Then the joint likelihood of all observations, latent variables, and unknown parameters given all the hyperparameters $\Omega = (\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha^s_0}, \boldsymbol{\alpha^s_1}, \boldsymbol{\alpha^u_0}, \boldsymbol{\alpha^u_1})$ is:

$$p(\boldsymbol{o}, \boldsymbol{t}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\phi^0}, \boldsymbol{\phi^1}, \boldsymbol{\theta^0}, \boldsymbol{\theta^1}|\Omega) = \prod_{s=1}^{S} p(\phi^0_s|\boldsymbol{\alpha^s_0})p(\phi^1_s|\boldsymbol{\alpha^s_1}) \prod_{u=1}^{U_s} p(\theta^0_{su}|\boldsymbol{\alpha^u_0})p(\theta^1_{su}|\boldsymbol{\alpha^u_1}) \prod_{m=1}^{M} p(o^s_{um}|t_m, y_m, \theta^0_{su}, \theta^1_{su}, \phi^0_s, \phi^1_s) p(t_m|\beta_m)p(\beta_m|\boldsymbol{\gamma})p(y_m|\varphi_m)p(\varphi_m|\boldsymbol{\eta}).$$

Obviously, it is intractable to perform exact inference on the posterior distribution of all the latent variables. Therefore, we employ collapsed Gibbs sampling algorithm to iteratively sample each variable from its full conditional distribution given all the other variables.

4.2 Latent Variable Inference

In the proposed model Sifter, there are two latent variables: truth indicator t_m and reliability indicator y_m. In the inference process, we use o to replace o^s_{um} for simplicity.

Truth Indicator Inference. Let $\boldsymbol{t_{-m}}$ be the truth of all side-effects except m. We iteratively sample for each side-effect given the current truth indicator of the other side-effects:

$$p(t_m = i|\boldsymbol{t_{-m}}, \boldsymbol{o}, \boldsymbol{y}) \propto \gamma_i \cdot \prod_{s=1}^{S}\prod_{u=1}^{U_s} \left[\frac{\alpha^s_{i,o} + n^{-m}_{s,u,i,o,y_m=0}}{\alpha^s_{i,1} + n^{-m}_{s,u,i,1,y_m=0} + \alpha^s_{i,0} + n^{-m}_{s,u,i,0,y_m=0}}\right]^{y_m=0} \left[\frac{\alpha^u_{i,o} + n^{-m}_{s,u,i,o,y_m=1}}{\alpha^u_{i,1} + n^{-m}_{s,u,i,1,y_m=1} + \alpha^u_{i,0} + n^{-m}_{s,u,i,0,y_m=1}}\right]^{y_m=1}, \tag{1}$$

where $n^{-m}_{s,u,i,j,k} = |\{m' \in \{1, \cdots, m-1, m+1, \cdots, M\}|\Delta\}|$ and $\Delta = \{s_{m'} = s_m, u_{m'} = u_m, t_{m'} = i, o_{m'} = j, y_{m'} = k\}$. Here, $s_{m'} = s_m$ and $u_{m'} = u_m$ denote the side-effects provided by the same user u in platform s. From Eq. (1), we can observe that the truth indicator is related to both platform and user information.

Reliability Indicator Inference. Similar with the truth indicator inference process, let $\boldsymbol{y_{-m}}$ be the reliability indicators of all side-effects except m. We can sample the current reliability indicator when given other indicators of side-effects:

$$p(y_m = k|\boldsymbol{y_{-m}}, \boldsymbol{t}, \boldsymbol{o}) \propto \eta_i \cdot \prod_{s=1}^{S}\prod_{u=1}^{U_s} \left[\frac{\alpha^u_{t_m,o} + n^{-m}_{s,u,t_m,o,k}}{\alpha^u_{t_m,1} + n^{-m}_{s,u,t_m,1,k} + \alpha^u_{t_m,0} + n^{-m}_{s,u,t_m,0,k}}\right]^{k} \left[\frac{\alpha^s_{t_m,o} + n^{-m}_{s,u,t_m,o,1-k}}{\alpha^s_{t_m,1} + n^{-m}_{s,u,t_m,1,i-k} + \alpha^s_{t_m,0} + n^{-m}_{s,u,t_m,0,1-k}}\right]^{1-k}, \tag{2}$$

where $n^{-m}_{s,u,t_m,j,k} = |\{m' \in \{1, \cdots, m-1, m+1, \cdots, M\}|\Lambda\}|$ and $\Lambda = \{s_{m'} = s_m, u_{m'} = u_m, t_{m'} = t_m, o_{m'} = j, y_{m'} = k\}$. Here, $s_{m'} = s_m$, $u_{m'} = u_m$ and $t_{m'} = t_m$ denote that the side-effects provided by user u in platform s have the same truth indicator with t_m. From Eq. (2), we can observe that when $k = 1$, the reliability indicator is inferred with user-level information; otherwise, it is sampled using platform information.

4.3 Parameter Estimation

In the proposed model Sifter, we utilize the conjugacy of exponential families when modeling the platform-level quality $\boldsymbol{\phi^0}$ and $\boldsymbol{\phi^1}$, user-level quality $\boldsymbol{\theta^0}$ and $\boldsymbol{\theta^1}$, so that they can be integrated out in the sampling process. We derive these parameters and make the following parameter estimations:

$$\theta^0_{su} \propto \frac{\sum_{m=1}^{M} n_{s,u,t_m=0,o=1,y_m=1} + \alpha^u_{0,1}}{\sum_{m=1}^{M}\sum_{j\in\{0,1\}} n_{s,u,t_m=0,o=j,y_m=1} + \alpha^u_{0,1} + \alpha^u_{0,0}}, \tag{3}$$

$$\theta^1_{su} \propto \frac{\sum_{m=1}^{M} n_{s,u,t_m=1,o=1,y_m=1} + \alpha^u_{1,1}}{\sum_{m=1}^{M}\sum_{j\in\{0,1\}} n_{s,u,t_m=1,o=j,y_m=1} + \alpha^u_{1,1} + \alpha^u_{1,0}}, \tag{4}$$

$$\phi_s^0 \propto \frac{\sum_{u=1}^{U_s}\sum_{m=1}^{M} n_{s,u,t_m=0,o=1,y_m=0} + \alpha_{0,1}^s}{\sum_{u=1}^{U_s}\sum_{m=1}^{M}\sum_{j\in\{0,1\}} n_{s,u,t_m=0,o=j,y_m=0} + \alpha_{0,1}^s + \alpha_{0,0}^s}, \quad (5)$$

$$\phi_s^1 \propto \frac{\sum_{u=1}^{U_s}\sum_{m=1}^{M} n_{s,u,t_m=1,o=1,y_m=0} + \alpha_{1,1}^s}{\sum_{u=1}^{U_s}\sum_{m=1}^{M}\sum_{j\in\{0,1\}} n_{s,u,t_m=1,o=j,y_m=0} + \alpha_{1,1}^s + \alpha_{1,0}^s}, \quad (6)$$

where $n_{s,u,t_m=i,o=j,y_m=k}$ denotes the number of side-effects provided by user u in platform s, and each side-effect satisfies that the truth indicator is i, the observation is j, and the reliability indicator is k.

4.4 Algorithm Flow

The model inference and parameter learning process are described in Algorithm 1. We first randomly assign the truth indicator t_m and the reliability indicator y_m for each side-effect, and then calculate the initial counts for each user. Then in each iteration, we re-sample each truth indicator (reliability indicator) from its distribution conditioned on all the other truth indicators (reliability indicator) and update the quality counts for each user accordingly. For the final prediction, discarding the first $I = 40$ samples (burn-in period), we use the results sampled from every $\lambda = 5$ iterations in the last several rounds to calculate the expectation of true side-effects (thinning). If the expectation of a potential side-effect equals to or is greater than 0.5, the estimated side-effect is true; otherwise, it is false. The time complexity of Algorithm 1 is $O(NMSU_s)$, which is linear in the number of observations, where $N = 100$ is the maximum of iterations.

Algorithm 1 Sifter Learning Algorithm.

while iter < N **do**
  for the m-th side-effect ($m = 1, 2, \cdots, M$)
    $p_{t_m} \leftarrow \gamma_{t_m}, p_{1-t_m} \leftarrow \gamma_{1-t_m}$;
    for the s-th platform ($s = 1, 2, \cdots, S$)
      for the u-th user ($u = 1, 2, \cdots, U_s$)
        Calculate p_{t_m} and p_{1-t_m} according to Eq. (1);
      end for
    end for
    if $random() < \frac{p_{1-t_m}}{p_{t_m}+p_{1-t_m}}$ **then**
      $t_m \leftarrow 1 - t_m$ and update counts;
    end if
  end for
  for the m-th side-effect ($m = 1, 2, \cdots, M$)
    $p_{y_m} \leftarrow \eta_{y_m}, p_{1-y_m} \leftarrow \eta_{1-y_m}$;
    for the s-th platform ($s = 1, 2, \cdots, S$)
      for the u-th user ($u = 1, 2, \cdots, U_s$)
        Calculate p_{y_m} and p_{1-y_m} according to Eq. (2);
      end for
    end for
    if $random() < \frac{p_{1-y_m}}{p_{y_m}+p_{1-y_m}}$ **then**
      $y_m \leftarrow 1 - y_m$ and update counts;
    end if
  end for
  if iter > l and iter % λ == 0 **then**
    $p(t_m = 1) \leftarrow p(t_m = 1) + \frac{t_m}{samplesize}$;
  end if
end while
for the s-th platform ($s = 1, 2, \cdots, S$) **do**
  Calculate ϕ_s^0 and ϕ_s^1 according to Eq. (5) and Eq. (6);
  for the u-th user ($u = 1, 2, \cdots, U_s$) **do**
    Calculate θ_{su}^0 and θ_{su}^1 according to Eq. (3) and Eq. (4);
  end for
end for

5 EXPERIMENTS

In this section, we describe a thorough evaluation of the proposed Sifter compared with state-of-the-art methods on real-world datesets.

5.1 Datasets

In our experiments, we select five widely used **drug families** (based on WebMD[7]), including *Thyroxine*, *Metformin*, *Omeprazole*, *Alprazolam* and *Ibuprofen* [25]. We extract the five drug families' data from two platforms: FAERS and Healthboards. The ground truth data are collected from SIDER[8]. Next, we describe these data sources.

The SIDER Ground Truth Data. We rely on the data from SIDER as the ground truth for drug side-effects, which contains information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts. The last release, SIDER 4, contains data on 1,430 drugs, 5,880 side-effects and 140,064 drug-side-effect pairs. We first collect side-effects for five drug families, then extract CUIs and their corresponding semantic types, and finally select CUIs with four semantic types[9] as ground truth data.

The FAERS Platform. FDA Adverse Event Reporting System (FAERS) is a database or platform that contains information on adverse event (i.e., side-effect) and medication error reports submitted to FDA. It is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products. Side-effects are coded to preferred terms in the Medical Dictionary for Regulatory Activities (MedDRA) disclaimer icon terminology. We convert side-effects from MedDRAs to CUIs using MetaMap and select CUIs with the four semantic types. The collected data are reported from September 2012 to September 2015.

The Healthboards Platform. Healthboards.com is one of the largest online health communities, with 850,000 members and over 4.5 million posts. We extract users' posts on the five drug families, then recognize side-effect terms, and finally covert these terms to CUIs using MetaMap and select CUIs included in the four semantic types.

Since all the datasets only contain positive claims, we randomly add negative claims for each user, which have the same quantities with positive claims. Table 4 lists the statistics of the five drug families, including the number of users, side-effects, observations and ground truth for each platform. From Table 4, we can observe that only the FAERS platform contains all the true side-effects for the drug family Metformin, which means that it is essential to account for different platforms when discovering drug side-effects.

[7] http://www.webmd.com
[8] http://sideeffects.embl.de
[9] Four semantic types include *Sign or Symptom*, *Disease or Syndrome*, *Mental or Behavioral Dysfunction*, and *Mental Process*.

Table 4: Data Statistics.

Dataset	FAERS Platform				Healthboards Platform				SIDER
	# Users	# Side-Effects	# Observations	# Truth	# Users	# Side-Effects	# Observations	# Truth	# Ground Truth
Thyroxine	399	210	3,734	37	78	258	5,524	34	42
Metformin	1,327	356	9,560	69	109	247	5,720	32	69
Omeprazole	1,460	553	19,014	137	198	476	14,040	79	139
Alprazolam	1,310	525	8,108	127	524	1,154	33,769	117	143
Ibuprofen	2,142	465	6,948	149	683	868	22,919	97	157

5.2 Experiment Setup

We compare the proposed Sifter model against several state-of-the-art algorithms briefly summarized as follows: Voting is a naive baseline, which regards a side-effect as truth if the proportion of the users that provide this side-effect (i.e., positive claims) exceeds a certain threshold. Ranking is a straightforward method, which selects side-effects with high frequency. In our experiments, we select side-effects ranked in the top 100 as the estimated truth. 3-Estimates [7] iteratively computes reliability degrees of users and trustworthiness of each observation. It also considers negative claims. MLE [32] is based on Expectation Maximization (EM) algorithm to quantify the reliability of users and the trustworthiness of their observations. MBM [33] is an integrated Bayesian approach to discover multiple truths, which clusters users and observations into different groups to improve truth-finding efficiency. LTM [47] is a probabilistic approach to discover multiple truth and take sensitivity and specificity as source quality, which is a special case of the proposed Sifter.

Note that since all the baselines do not distinguish observations' platforms, we merge observations provided by users in the FAERS platform and the Healthboards platform into one dataset and run them on this new dataset. Parameters for the baselines are set according to the best performance after running a series of experiments. For our method, we simply use a generic parameter setting for $\boldsymbol{\eta} = (10, 10)$ and $\boldsymbol{\gamma} = (10, 10)$. For both the FARES and Healthboards platform, we set $\boldsymbol{\alpha}_0^{\boldsymbol{u}} = \boldsymbol{\alpha}_0^{\boldsymbol{s}} = (100, 1000)$ and $\boldsymbol{\alpha}_1^{\boldsymbol{u}} = \boldsymbol{\alpha}_1^{\boldsymbol{s}} = (5000, 100)$.

We select three commonly used metrics to evaluate all the baselines and the proposed Sifter.

Precision, Recall and F1-score. *Precision* (*P*) measures the probability of the output side-effects being correct. *Recall* (*R*) measures the probability of real true side-effects being estimated as final outputs. *F1-score* (*F1*) computes the harmonic of precision and recall, i.e., $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$.

5.3 Performance Validation

Since we randomly sample negative claims and arbitrarily initialize the parameters in the proposed model, in order to show the robustness of Sifter, we run the algorithm 10 times and report the average of precision, recall and F1-score in Table 5. The experimental results show that the proposed Sifter can significantly improve the performance when identifying drug side-effects compared with baselines on the five drug families.

From Table 5, we can observe that the proposed Sifter is better than all the baseline methods in terms of F1-score. Voting-25 means that the side-effects is selected as the estimated truths if there are at least 25% users claimed them. On the three datasets (Thyroxine, Metformin and Ibuprofen), all the measures are 0, which means that none of side-effects is claimed by $\geq$ 25% users. Even though on the Alprazolam dataset, the precision of Voting-25 is 0.667, it only returns 6 side-effects in which four ones are correct. Since the number of real side-effects of Alprazolam is 143 in Table 4, the recall of Voting-25 is very low (0.028). From this observation, we can safely conclude that the five datasets are very noisy. Therefore, it is difficult to identify the correct drug side-effects on them.

Compared with Ranking-100, the proposed Sifter outperforms this simple baseline. However, the performance of Ranking-100 is better than that of other baselines on the precision. This also shows that there exits much noisy information in these datasets.

Some single truth discovery methods, such as TruthFinder [41], Investment and PooledInvestment [26], cannot achieve good performance due to the low quality of the five datasets. Therefore, we only select 3-Estimates as a baseline which takes both positive and negative claims into consideration. Compared with 3-Estimates, the precision of Sifter is higher, but the recall is lower on the Ibuprofen dataset. That is because 3-Estimates uses accuracy to measure users' quality, and some negative claims would be assigned higher trustworthiness than they should be.

Since MLE only considers the positive claims and ignores the negative ones, the performance of MLE is worse than Sifter's on the five datasets. MBM considers both positive and negative claims, and groups users based on the observations. LTM models user quality based on sensitivity and specificity. Since both MBM and LTM are sensitive on the noisy data, these two approaches return most of side-effects with high frequency as the estimated truth. Thus, the recalls of MBM and LTM are better than that of Sifter, but the precisions and F1-scores are much worse.

5.4 Model Validation

Since the proposed Sifter considers the differences among different platforms. We illustrate the benefit of distinguishing platforms when identifying drug side-effects by comparing with the methods that conduct drug side-effect discovery on different platforms separately. We first run all the baselines on each platform and obtain the possible correct side-effects for a given drug. Then, we combine all the estimated side-effects into a set as the final output.

Table 6 shows the results of model validation on the five drug families. We can see that the performance of Voting-25 improves significantly. This is because the number of users decreases when we run Voting-25 on each platform's data. For Ranking-100, the number of output side-effects improves, so the precisions and F1-scores decrease on the first three datasets but the recalls improve on

Table 5: Performance on the Five Drug Families.

Method	Thyroxine			Metformin			Omeprazole			Alprazolam			Ibuprofen		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Voting-25	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.667**	0.028	0.054	0.000	0.000	0.000
Ranking-100	0.250	0.595	0.352	0.350	0.507	0.414	0.530	0.381	0.444	0.340	0.238	0.280	0.350	0.223	0.272
3-Estimates	0.199	0.643	0.303	0.187	0.681	0.293	0.298	0.662	0.411	0.287	0.650	0.398	0.299	0.758	0.429
MLE	0.075	0.548	0.132	0.073	0.435	0.124	0.093	0.417	0.152	0.052	0.322	0.090	0.090	0.414	0.148
MBM	0.109	**1.000**	0.196	0.131	0.986	0.231	0.160	0.993	0.276	0.103	**1.000**	0.186	0.140	**1.000**	0.246
LTM	0.108	**1.000**	0.195	0.132	**1.000**	0.234	0.161	**1.000**	0.277	0.102	**1.000**	0.185	0.140	**1.000**	0.246
Sifter	**0.360**	0.738	**0.484**	**0.463**	0.693	**0.555**	**0.561**	0.727	**0.633**	0.427	0.728	**0.538**	**0.520**	0.680	**0.589**

Table 6: Results of Model Validation.

Method	Thyroxine			Metformin			Omeprazole			Alprazolam			Ibuprofen		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Voting-25	0.163	0.167	0.165	0.074	0.029	0.042	0.241	0.050	0.083	0.279	0.119	0.167	0.111	0.019	0.033
Ranking-100	0.199	0.810	0.319	0.240	0.609	0.344	0.387	0.468	0.424	0.352	0.427	0.386	0.382	0.420	0.340
3-Estimates	0.199	0.643	0.303	0.187	0.681	0.293	0.298	0.662	0.411	0.105	0.923	0.188	0.299	0.758	0.429
MLE	0.107	0.976	0.193	0.115	0.812	0.201	0.157	0.842	0.265	0.095	0.636	0.165	0.134	0.675	0.223
MBM	0.113	0.976	0.201	0.133	0.986	0.235	0.163	0.993	0.289	0.103	0.986	0.186	0.142	0.994	0.248
LTM	0.108	**1.000**	0.195	0.132	**1.000**	0.234	0.161	**1.000**	0.277	0.102	**1.000**	0.185	0.140	**1.000**	0.246
Sifter	**0.360**	0.738	**0.484**	**0.463**	0.693	**0.555**	**0.561**	0.727	**0.633**	**0.427**	0.728	**0.538**	**0.520**	0.680	**0.589**

the five datasets. Since the amount of noisy information decreases in model validation experiment, the performance of LTM improves significantly. However, the performance of MBM changes slightly, and the results of LTM do not change. Compared with all the baseline approaches, Sifter assigns different quality for platforms, which consequently performs better than baselines.

5.5 Estimated Truth Analysis

To validate the benefit of utilizing the information from multiple platforms in drug side-effect discovery task, we analyze the correctly estimated side-effects by the proposed Sifer and summarize the number of correct side-effects learned from the FAERS and Healthboards platform in Table 7. "Both" represents the correct side-effects from two platforms, "FAERS Only" denotes the number of correct side-effects only discovered from the FAERS platform, "Healthboards Only" is the number of correct side-effects only obtained from the Healthboards platform, and "Total" is the number of side-effects correctly estimated by the proposed Sifter. We can observe that the final correct side-effects come from different platforms, which is in accord with our motivation of the proposed model.

Table 7: Results of the Estimated Truth Analysis.

Dataset	Estimated Truth From			Total
	Both	FAERS Only	Healthboards Only	
Thyroxine	25	3	3	31
Metformin	26	21	0	47
Omeprazole	64	35	2	101
Alprazolam	82	12	10	104
Ibuprofen	67	30	8	105

5.6 Platform Quality Validation

From Table 4, we can observe that the ratio of $\frac{\#Truths}{\#Side-Effects}$ on the FAERS platform is greater than that on the Healthboards platform. It means that the overall quality of the FAERS platform is better than that of the Healthboards platform. To analyze the learned quality of each platform on the Ibuprofen dataset, we show the **learned** *Sensitivity* and *Specificity* of FAERS and Healthboards in Figure 2. We can observe that Sifter assigns FAERS higher *Sensitivity*, which corresponds to the fact that the data provided by the FAERS platform contains more correct side-effects (149 of 157) for the drug family Ibuprofen in Table 4. In Figure 2, the learned *Specificity* (1 - *False Positive Rate*) of FAERS is higher than that of Healthboards. It is confirmed by the fact that the FAERS platform just has 316 incorrect side-effects, but Healthboards contains 771 incorrect side-effects shown in Table 4, i.e., the data provided by the Healthboards platform contains a lot of noisy information. These observations show that the learned platform-level quality is reasonable, and also validate the intuition that modeling different platforms' quality with *Sensitivity* and *Specificity* is essential.

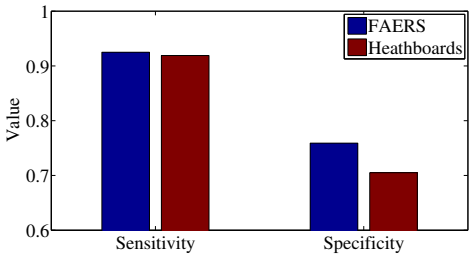

Figure 2: Platform Quality on the Ibuprofen Dataset.

6 RELATED WORK

In this section, we review some work related to drug side-effect discovery and truth discovery.

Drug Side-Effect Discovery. With the thriving growth of online social networks and forums, more and more health related data can be collected easily. Healthcare data mining has become a hot research topic [2, 3, 9, 10, 12, 20, 22, 30, 34, 35, 37, 42, 43, 45, 46]. Especially, drug side-effect discovery has become an active research area. Leman et. al. [14] use healthcare forum data to identify drug side-effects. Liu et. al. [21] extract a complete set of side effect expressions from patient-submitted drug reviews, and construct a hierarchical ontology of side effects to quantify associations between drugs and symptoms. Yang et. al. [39] employ association mining to identify side-effects. Yates et. al. [40] use the support and strength of co-occurrence of each drug-symptom interaction to train a classifier. Chee et. al. [4] use ensemble based classifier to classify drug side-effects. Bian et. al. [1] utilize SVM as classifier to classify whether drug-symptom interactions are side-effects. Katragadda et. al. [13] introduce a link classification method to detect drug side-effects on Twitter data. Mukherjee et. al. [25] propose a semi-supervised method, which uses linguistic features, user features and part of labels to learn user trustworthiness, statement credibility and language objectivity simultaneously. Some researchers also use large-scale web queries to identify adverse drug reactions [35, 43]. Since the aforementioned methods are supervised or semi-supervised approaches, they use drugs' features and labels as inputs. Different from these methods, we propose an unsupervised model to automatically identify side-effects.

Truth Discovery. There is extensive work in the area of truth discovery [15, 18, 49], including single truth discovery and multiple truth discovery. In this paper, we focus on the work for multiple truth discovery.

Single Truth Discovery. Yin et. al. [41] formally define the truth discovery problem and propose TruthFinder, a heuristic method, to compute the probability of each object being correct given the estimated user reliability degrees. Investment is proposed by Pasternack et. al. [26] in which sources "invest" their reliability uniformly on the observations they provide, and collect credits back from the confidence of those observations. In turn, the confidence of observations grows according to a non-linear function defined based on the sum of invested reliability from their providers. 3-Estimates [7] iteratively computes reliability degrees of users and trustworthiness of each observations, which also uses positive observations only. Li et. al. [17] propose an optimization framework, CRH, to model different data types jointly, and estimate source reliability and truth simultaneously. They also propose CATD [16] method to automatically estimate truth from conflicting data with long-tail phenomenon. Li et. al. [19] consider the temporal relations among both object truths and source reliability and propose an incremental truth discovery framework. Pasternack et. al. use a set of probabilistic model parameters to estimate the source credibility in [27]. Dong et. al. [6] focus on source selection problem in truth finding. Vydiswaran et. al. [31] propose models to estimate users' reliability and discover credible claims on unstructured data. Qi et. al. [29] propose a model to jointly learn group level source reliability and estimate true answers. Zheng et. al. [48] study how to leverage domain knowledge to accurately model a sourcefis quality. Ma et. al. [23] incorporate text information and propose a probabilistic graphical model to learn fine-grained source reliability and estimate the true answers.

Multiple Truth Discovery. There is extensive work in the area of truth discovery, and we mainly review the work for multiple truth discovery. Zhao et. al. [47] present a probabilistic graphical model LTM to resolve the problem of existence of multiple truths for a single entity in truth discovery tasks. Wang et. al. [32] propose a Maximum Likelihood Estimation (MLE) method, which deals with Boolean positive observations. Wang et. al. [33] propose an integrated Bayesian approach, named MBM, to the multiple truth discovery problem. PRECREC is proposed by Pochampally et. al. [28], which compute the trustworthiness using precision and recall of each source, but it needs a gold standard data to get source precision and recall.

All the above discussed methods cannot estimate platform-level and user-level reliability, and infer true information simultaneously. To the best of our knowledge, we are the first to build an unsupervised model to identify drug side-effects across multiple platforms.

7 CONCLUSIONS

Drug side-effect discovery is an important and practical issue in the world. Many existing work proposes supervised or semi-supervised methods to identify drug side-effects. However, the performance of these methods depends on the quality of the provided labels. If the known information's quality is low, the performance drops significantly. How to detect accurate side-effect in an unsupervised way is a promising research problem. Borrowing the idea of truth discovery, we can estimate true information in an unsupervised manner, but identifying drug side-effects is more challenging. Drug side-effect discovery focuses on single object with multiple claims, which is different from the problem setting of existing multiple truth discovery approaches. Since single data source or platform may not provide all the correct side-effects, it is important to collect data from multiple platforms. Moreover, the quality of data collected from different platforms should be different. To solve the aforementioned challenges, in this paper, we propose a probabilistic graphical model to identify the correct drug side-effects without any supervision. By modeling platform-level and user-level quality, the proposed model Sifter can characterize the quality of platforms and users accurately and estimate correct drug side-effect effectively. Experimental results on five real-world datasets show that the proposed Sifter can significantly improve the performance of identifying drug side-effects compared with the state-of-the-art truth discovery approaches.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work is supported in part by the US National Science Foundation under grants IIS-1319973, IIS-1553411, CNS-1566374, CNS-1652503, IIS-1218393 and IIS-1514204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Jiang Bian, Umit Topaloglu, and Fan Yu. 2012. Towards large-scale twitter mining for drug-related adverse events. In *Proc. of Workshop on Smart Health and Wellbeing*. ACM, 25–32.

[2] Karla L Caballero Barajas and Ram Akella. 2015. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proc. of KDD*. ACM, 69–78.

[3] Bokai Cao, Xiangnan Kong, Jingyuan Zhang, S Yu Philip, and Ann B Ragin. 2015. Mining brain networks using multiple side views for neurological disorder identification. In *Proc. of ICDM*. IEEE, 709–714.

[4] BW Chee, R Berlin, and B Schatz. 2011. Predicting adverse drug events from personal health messages.. In *Proc. of AMIA*, Vol. 2011. American Medical Informatics Association, 217–226.

[5] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Integrating conflicting data: The role of source dependence. *PVLDB* 2, 1 (2009), 550–561.

[6] Xin Luna Dong, Barna Saha, and Divesh Srivastava. 2012. Less is more: Selecting sources wisely for integration. *PVLDB* 6, 2 (2012), 37–48.

[7] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating information from disagreeing views. In *Proc. of WSDM*. ACM, 131–140.

[8] Kathleen M Giacomini, Ronald M Krauss, Dan M Roden, Michel Eichelbaum, Michael R Hayden, and Yusuke Nakamura. 2007. When good drugs go bad. *Nature* 446, 7139 (2007), 975–977.

[9] Aron Henriksson, Jing Zhao, Henrik Boström, and Hercules Dalianis. 2015. Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection. In *Proc. of BIBM*. IEEE, 343–350.

[10] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. 2014. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proc. of KDD*. ACM, 115–124.

[11] Yanqing Ji, Hao Ying, Peter Dews, Ayman Mansour, John Tran, Richard E Miller, and R Michael Massanari. 2011. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE ITB* 15, 3 (2011), 428–437.

[12] Isak Karlsson and Henrik Boström. 2016. Predicting adverse drug events using heterogeneous event sequences. In *Proc. of ICHI*. IEEE, 356–362.

[13] Satya Katragadda, Harika Karnati, Murali Pusala, Vijay Raghavan, and Ryan Benton. 2015. Detecting adverse drug effects using link classification on twitter data. In *Proc. of BIBM*. IEEE, 675–679.

[14] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. In *Proc. of Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, 117–125.

[15] Guoliang Li, Jiannan Wang, Yudian Zheng, and Michael J Franklin. 2016. Crowdsourced data management: A survey. *IEEE TKDE* 28, 9 (2016), 2296–2319.

[16] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A confidence-aware approach for truth discovery on long-tail data. *PVLDB* 8, 4 (2014), 425–436.

[17] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proc. of SIGMOD*. 1187–1198.

[18] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A Survey on Truth Discovery. *SIGKDD Explor. Newsl.* 17, 2 (2016), 1–16.

[19] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. On the discovery of evolving truth. In *Proc. of KDD*. ACM, 675–684.

[20] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. 2015. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proc. of KDD*. ACM, 705–714.

[21] Jingjing Liu, Alice Li, and Stephanie Seneff. 2011. Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. In *Proc. of IMMM*. 23–29.

[22] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *Proc. of KDD*. ACM.

[23] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proc. of KDD*. ACM, 745–754.

[24] Chuishi Meng, Wenjun Jiang, Yaliang Li, Jing Gao, Lu Su, Hu Ding, and Yun Cheng. 2015. Truth discovery on crowd sensing of correlated entities. In *Proc. of SenSys*. ACM, 169–182.

[25] Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu Niculescu Mizil. 2014. People on drugs: Credibility of user statements in health communities. In *Proc. of KDD*. ACM, 65–74.

[26] Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In *Proc. of Coling*. Association for Computational Linguistics, 877–885.

[27] Jeff Pasternack and Dan Roth. 2013. Latent credibility analysis. In *Proc. of WWW*. International World Wide Web Conferences Steering Committee, 1009–1020.

[28] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. 2014. Fusing data with correlations. In *Proc. of SIGMOD*. ACM, 433–444.

[29] Guo-Jun Qi, Charu C Aggarwal, Jiawei Han, and Thomas Huang. 2013. Mining collective intelligence in diverse groups. In *Proc. of WWW*. International World Wide Web Conferences Steering Committee, 1041–1052.

[30] Qiuling Suo, Fenglong Ma, Giovanni Canino, Jing Gao, Aidong Zhang, Pierangelo Veltri, and Agostino Gnasso. 2017. A Multi-task Framework for Monitoring Health Conditions via Attention-based Recurrent Neural Networks. In *Proc. of AMIA*.

[31] VG Vydiswaran, ChengXiang Zhai, and Dan Roth. 2011. Content-driven trust propagation framework. In *Proc. of KDD*. ACM, 974–982.

[32] Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. 2012. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proc. of IPSN*. ACM, 233–244.

[33] Xianzhi Wang, Quan Z Sheng, Xiu Susie Fang, Lina Yao, Xiaofei Xu, and Xue Li. 2015. An integrated Bayesian approach for effective multi-truth discovery. In *Proc. of CIKM*. ACM, 493–502.

[34] Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C Denny, Abel Kho, You Chen, Bradley A Malin, and Jimeng Sun. 2015. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proc. of KDD*. ACM, 1265–1274.

[35] Ryen W White, Nicholas P Tatonetti, Nigam H Shah, Russ B Altman, and Eric Horvitz. 2013. Web-scale pharmacovigilance: listening to signals from the crowd. *JAMIA* 20, 3 (2013), 404–408.

[36] Houping Xiao, Jing Gao, Qi Li, Fenglong Ma, Lu Su, Yunlong Feng, and Aidong Zhang. 2016. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *Proc. of KDD*. ACM, 1935–1944.

[37] Houping Xiao, Jing Gao, Long Vu, and Deepak S. Turaga. 2017. Learning Temporal State of Diabetes Patients via Combining Behavioral and Demographic Data. In *Proc. of KDD*. ACM.

[38] Houping Xiao, Jing Gao, Zhaoran Wang, Shiyu Wang, Lu Su, and Han Liu. 2016. A truth discovery approach with theoretical guarantee. In *Proc. of KDD*. ACM, 1925–1934.

[39] Christopher C Yang, Ling Jiang, Haodong Yang, and Xuning Tang. 2012. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In *Proc. of ACM SIGKDD Workshop on Health Informatics*.

[40] Andrew Yates and Nazli Goharian. 2013. ADRTrace: Detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In *Advances in Information Retrieval*. Springer, 816–819.

[41] Xiaoxin Yin, Jiawei Han, and Philip S Yu. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE TKDE* 20, 6 (2008), 796–808.

[42] Elad Yom-Tov, Diana Borsa, Andrew C Hayward, Rachel A McKendry, and Ingemar J Cox. 2015. Automatic identification of web-based risk markers for health events. *JMIR* 17, 1 (2015).

[43] Elad Yom-Tov and Evgeniy Gabrilovich. 2013. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *JMIR* 15, 6 (2013), e124.

[44] Hengtong Zhang, Qi Li, Fenglong Ma, Houping Xiao, Yaliang Li, Jing Gao, and Lu Su. 2016. Influence-aware truth discovery. In *Proc. of CIKM*. 851–860.

[45] Jingyuan Zhang, Bokai Cao, Sihong Xie, Chun-Ta Lu, Philip S. Yu, and Ann B. Ragin. 2016. Identifying Connectivity Patterns for Brain Diseases via Multi-side-view Guided Deep Architectures. In *Proc. of SDM*. SIAM, 36–44.

[46] Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. 2015. Label Propagation Prediction of Drug-Drug Interactions Based on Clinical Side Effects. *Scientific Reports* 5 (2015).

[47] Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. 2012. A Bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB* 5, 6 (2012), 550–561.

[48] Yudian Zheng, Guoliang Li, and Reynold Cheng. 2016. DOCS: a domain-aware crowdsourcing system using knowledge bases. *PVLDB* 10, 4 (2016), 361–372.

[49] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: is the problem solved? *PVLDB* 10, 5 (2017), 541–552.