Automated Off-label Drug Use Detection from User Generated Content

Mengnan Zhao
College of Computing and Informatics
Drexel University
Philadelphia, Pennsylvania
mz438@drexel.edu

Christopher C. Yang
College of Computing and Informatics
Drexel University
Philadelphia, Pennsylvania
chris.yang@drexel.edu

ABSTRACT

Off-label drug use refers to using marketed drugs for indications that are not listed in their FDA labeling information. Such uses are very common and sometimes inevitable in clinical practice. To some extent, off-label drug uses provide a pathway for clinical innovation, however, they could cause serious adverse effects due to lacking scientific research and tests. Since identifying the off-label uses can provide a clue to the stakeholders including healthcare providers, patients, and medication manufacturers to further the investigation on drug efficacy and safety, it raises the demand for a systematic way to detect off-label uses. Given data contributed by health consumers in online health communities (OHCs), we developed an automated approach to detect off-label drug uses based on heterogeneous network mining. We constructed a heterogeneous healthcare network with medical entities (e.g. disease, drug, adverse drug reaction) mined from the text corpus, which involved 50 diseases, 1,297 drugs, and 185 ADRs, and determined 13 meta paths between the drugs and diseases. We developed three metrics to represent the meta-path-based topological features. With the network features, we trained the binary classifiers built on Random Forest algorithm to recognize the known drug-disease associations. The best classification model that used *lift* to measure path weights obtained F1-score of 0.87, based on which, we identified 1,009 candidates of off-label drug uses and examined their potential by searching evidence from PubMed and FAERS.

KEYWORDS

Off-label drug use; Online health community; Heterogeneous network; Meta path; Classification

ACM-BCB'17, August 20-23, 2017, Boston, MA, USA.

ACM ISBN 978-1-4503-4722-8/17/08...$15.00.
DOI: http://dx.doi.org/10.1145/3107411.3107475

1. INTRODUCTION

Off-label drug use refers to using marketed drugs for indications that are not listed in their FDA labeling information. Although FDA currently controls which medications can be brought to the market, they are not able to supervise and control if the medications are prescribed in accordance with the approved labels. Off-label drug uses are very common in clinical practice. Up to one-fifth prescriptions are off-label, and the number is even higher in certain diseases (e.g. psychiatric diseases), certain conditions (e.g. intensive care), and special subpopulations (e.g. pediatric and pregnant patients) [1]. Although physicians prescribe off-label drugs for the benefit of patients, and off-label drug uses deliver effective treatment and provide a pathway for clinical innovation in some cases, they could cause serious outcomes due to lacking scientific research and tests. For example, morphine has never received an FDA approval for pain treatment in pediatric patients, but it is extensively used in hospitalized practice. However, the misuse or overdose of such narcotic can cause addiction or death, especially in children.

As off-label drug use is inevitable to some degree, most healthcare related participants have the interest in gaining information about off-label uses in a timely manner. Healthcare providers and patients are concerned about the observation information of off-label drug uses in practice; pharmaceutical companies are required to provide the postmarketing surveillance reports of drug uses; biomedical and clinical researchers are interested in the novel, especially those successful, off-label drug uses to assess the potential benefits and risks [2]. Identifying the off-label drug uses could present a clue to the stakeholders to further the investigation on drug efficacy and safety.

Despite these needs and the wide practice of off-label drug uses, there is no regulatory agency that monitors such uses. FDA Adverse Event Reporting System (FAERS) [3] contains reports on the adverse event and medication error of approved drugs and therapeutic biologic products to support the FDA's postmarketing safety surveillance program, while FAERS does not designate the off-label uses particularly. Survey is another approach of detecting off-label uses from physicians but limited by many conditions such as the number of respondents, the quality and truthfulness of answers, and the cost. Therefore, the above factors raise the demand for a systematic way to detect off-label drug uses.

In this study, we focus on developing an automated approach to detect off-label drug uses from the health consumer contributed

data. A variety of resources has been used to identify drug-disease associations, mainly including biomedical database, medical literature, EHR, and clinical note, and involving medical entities such as protein, compound, gene, pathway, and adverse drug reaction (ADR). While these data sources are useful in detecting off-label uses, they are mainly contributed by healthcare professionals and researchers. None of them capture the information offered by health consumers or patients. Meanwhile, the development of Web 2.0 not only breeds the various online social media sites like Facebook and Twitter, but also fosters online health communities (OHCs) such as MedHelp, PatientsLikeMe, and DailyStrength. OHCs have been growing in popularity across the world and provide a convenient way to exchange health information. It has been claimed that 80% of adults in US and 66% of adults in Europe seek online health advice [4]. 72% of Internet users said they searched online for health information in 2011 [5]. Taking MedHelp for instance, it empowers over 12 million people each month to seek and offer healthcare information on the site, which provides huge volumes of timely and valuable health-related information. With data coming from the OHC, we detected the most common three medical entities involved in user-generated contents: drugs, diseases, and adverse drug reactions (ADRs, and constructed a heterogeneous healthcare network. On the heterogeneous network, we determined the meta paths between drugs and diseases and extracted the meta-path-based topological features that are used for training a binary classification model to recognize the known drug-disease pairs and to predict the potential off-label drug-disease relationships.

2. RELATED WORK

A number of studies employed research methods such as survey and narrative interview to identify off-label uses. Conroy et al. [6] designed a prospective study to explore the use of unlicensed and off-label drugs in pediatric patients in European countries and found 39% of them involved off-label uses. Leslie & Rosenheck [7] investigated the off-label uses of antipsychotic drugs among patients who are enrolled in Medicaid by operating the retrospective analysis of administrative data, and found 57.6% patients received antipsychotics for off-label indications.

Recently, with the availability of online biomedical resources such as medical literature, pharmaceutical databases, electronic medical records (EMRs), and clinical text, a number of studies utilized automated methods to discover novel drug-indication/disease relationships to identify the off-label pairs. Mesgarpour et al. [8] focused on generating highly sensitive search strategies to detect off-label related documents in Excerpta Medica Database (EMBASE)-a major bibliographic database in biomedicine. Jung et al. [9] developed a predictive model to detect novel off-label uses from the clinical text. By utilizing NCBO Annotator, they tagged the words in the corpus of clinical notes first, and extracted the empirical relationships between drugs and indications from the population level rather than the textual level. In total, they calculated nine measures and used them as features to train a SVM classifier to predict novel off-label drug-indication pairs. Then they added extra 16 domain knowledge features from two pharmaceutical databases (Medi-Span and DrugBank) to develop a new SVM predictive model, and compared the performance of different feature sets [10]. In result, they discovered 6142 novel off-label pairs and validated 403 of them on MEDLINE literature and FAERS reports.

To unveil the drug-disease relationships that are critical for identifying off-label drug uses, previous studies have developed approaches based on natural language processing [11] and text mining [12]. Gottlieb et al. [13] calculated and ranked the similarity between potential drug-disease pairs with those in the gold standard set that they created by referring to several pharmaceutical databases, based on the assumption that similar drugs are indicated for similar diseases.

Additionally, some studies implemented heterogeneous network-based methods to discover novel drug-disease associations. Huang et al. [14] proposed a network propagation model to infer drug-disease associations, based on the integrated networks of three homogeneous networks and two heterogeneous, with the weights of edges assigned by knowledge from biomedical repositories. Chen et al. [15] also constructed a heterogeneous network of drug, disease, and protein, and predicted drug-disease associations by using a random walk based algorithm. Yu et al. [16] created a tripartite heterogeneous network of three types of nodes: diseases, drugs and protein complexes, with the weights of edges computed on a symmetrical conditional probability model. Moreover, heterogeneous network-based approaches have also been used to discover relationships between some other medical entities, such as drug-drug interactions [17], drug-targets [18], and multi drug-pathways [19]. For instance, Lee et al. [20] built a large heterogeneous network to discover drug-drug interactions (DDIs) by involving drugs, proteins, genes, pathways, side effects, targets and their interactions into the network.

3. OFF-LABEL DRUG USES DETECTION FROM ONLINE HEALTH COMMUNITIES

The previous studies render a promising way to unveil drug-disease relationships by heterogeneous network-based methods, based on which, we proposed a systematic method to identify off-label drug uses by using health consumer contributed data from OHCs, as illustrated in Figure 1. Firstly, we collected data from the OHC website to create the text corpus, and extracted the entities of diseases, drugs, and ADRs with lexicon based approach. Secondly, we constructed a heterogeneous healthcare network that contained three types of nodes (drug, disease, ADR) and six types of associations (drug-drug, disease-disease, ADR-ADR, drug-disease, drug-ADR, disease-ADR), determined the meta-paths, and utilized path count methods to measure the weights of associations between two nodes. Thirdly, we used the extracted weights as the features to train an effective classifier that distinguished known drug-disease pairs with those unknown, and then identified the off-label drug uses from the unknown pairs.

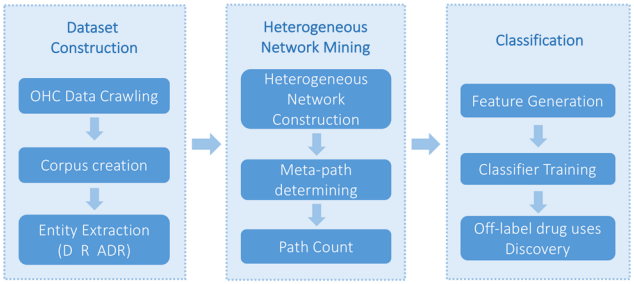

Figure 1: Flow of detecting off-label drug uses from OHC data

3.1 Dataset

In the current age, a large number of health consumers exchange health information on online health communities (OHCs). MedHelp (www.medhelp.org), as a pioneer in OHCs, has 176 health communities and empowers over 12 million visitors each month to discuss health related topics on the site. We collected the user generated data from MedHelp by implementing an automatic web crawler, with which, we retrieved all the posts and comments in the most popular 50 disease communities and obtained 70,960 threads (70,960 posts + 319,993 comments).

Within the extracted corpus, we identified the medical entities (i.e. drug, disease, ADR) by utilizing a lexicon-based approach, which has shown its advantage to OHC data in previous studies [21]. Unlike the pharmaceutical databases that include multiple medical entities (e.g. protein, compound, and gene), OHC data are mostly contributed by consumers without medical background talking about their diseases, the drugs they take, and the reactions or adverse reactions they experience. Therefore, we only employ the above three medical entities here.

For diseases and drugs, we utilized all their alternate names suggested in PharmGKB and UMLS and their abbreviations (e.g. OCD for Obsessive-Compulsive Disorder) to locate them, and tagged them with their UMLS-id. For example, the terms used to identify Parkinson included "parkinson" "parkinson disease" and "parkinson's disease", then all of those words were tagged with "C0030567".

However, the detection of ADR signals is more complicated than drugs and diseases in OHCs, because consumers use quite diverse and various expressions to describe the concepts and their adverse reactions. Thus, the standard medical lexicons managed by professionals are not applicable for analyzing user generated data. To address this problem, we resorted to Consumer Health Vocabulary (CHV) Wiki to create our ADR lexicon [21]. CHV Wiki connects the everyday expressions of healthcare-related topics with the professional expressions to bridge the communication gap between consumers and healthcare professionals [22]. For ADRs, it provides the preferred name of an ADR in UMLS and a list of its corresponding consumers' expressions. For example, "anorexia" is a professional expression of ADR, CHV Wiki extends it to "appetite lost" "appetite loss" "appetite lack" "no appetite" and several other common expressions of health consumers. Here we used all the expressions suggested by CHV Wiki to identify ADR signals in the corpus and tagged them with their UMLS-id. In result, we tagged 50 diseases, 1,297 drugs, and 185 ADRs on the corpus.

3.2 Heterogeneous network mining

3.2.1 Heterogeneous network construction

A heterogeneous network is defined as a graph consisting of nodes connected by links with at least two types of nodes or two types of links [23]. Most real-world networks are actually heterogeneous networks rather than homogeneous networks where the nodes and links are treated as the same type. Analysis based on homogeneous network may miss important semantic and schema-level information, while heterogeneous networks can deliver more essential, accurate and complete features, thus unveiling the underlying knowledge and patterns. Let $N = \{n_1, n_2, ..., n_k\}$ be a set of nodes and $L = \{l_1, l_2, ..., l_m\}$ be a set of links, then $G = (N, L)$ denotes the graph. In the graph G, each node $n_i \in N$ belongs to a particular type from γ; each link $l_i \in L$ belongs to a particular type from τ, and $|\gamma| > 1$ or $|\tau| > 1$, and can be directional or non-directional. Then $M_G = (\gamma, \tau)$ denotes the node types γ and link types τ in graph G.

By involving the medical entities we identified from the MedHelp corpus, we constructed a heterogeneous network that contains three types of nodes (disease(D), drug(R), ADR) and six types of links (drug-drug, disease-disease, ADR-ADR, drug-disease, drug-ADR, disease-ADR), as shown in Figure 2. That is, $N = \{d_1, \dots, d_k, r_1, \dots, r_m, adr_1, \dots, adr_n\}$, $\gamma = \{D, R, ADR\}$, and $\tau = \{L_{D-D}, L_{R-R}, L_{ADR-ADR}, L_{D-R}, L_{D-ADR}, L_{R-ADR}\}$. In this heterogeneous network, the interactions between two nodes, $w(n_i, n_j)$, were built on their co-occurrence in the same thread. Instead of inferring if a sentence contains use-to-treat relationship, we measured co-mention relationships at a population level by using co-occurrence-based indicators.

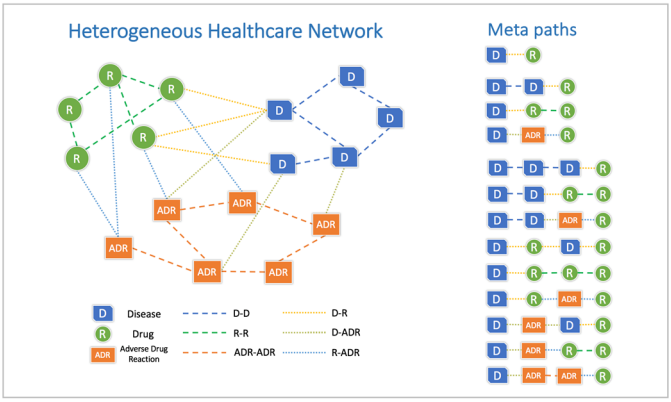

Figure 2: Heterogeneous healthcare network structure and meta-paths

3.2.2 Meta Path and Path Count

A meta path is a path defined on the network schema in the form of $A_1 \xrightarrow{L_1} A_2 \xrightarrow{L_2} \dots \xrightarrow{L_l} A_{l+1}$, which composes the relations between nodes in the heterogeneous network. Meta path-based approaches could describe the meta structure of the network and of the paths that derived from the network. To discover all the

possible and reliable associations between diseases and drugs, we defined the topology between them using meta paths, and by limiting path length within three, we found 13 meta paths, as shown in Table 1:

Table 1: Meta paths

Length	Meta Path	Path structure
1	*D-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_j$
2	*D-D-R*	$d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-R}}{\longleftrightarrow} r_j$
2	*D-R-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{R-R}}{\longleftrightarrow} r_j$
2	*D-ADR-R*	$d_i \overset{L_{D-ADR}}{\longleftrightarrow} adr_k \overset{L_{R-ADR}}{\longleftrightarrow} r_j$
3	*D-D-D-R*	$d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-D}}{\longleftrightarrow} d_m \overset{L_{D-R}}{\longleftrightarrow} r_j$
3	*D-D-R-R*	$d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-R}}{\longleftrightarrow} r_m \overset{L_{R-R}}{\longleftrightarrow} r_j$
3	*D-D-ADR-R*	$d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-ADR}}{\longleftrightarrow} adr_m \overset{L_{R-ADR}}{\longleftrightarrow} r_j$
3	*D-R-D-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{D-R}}{\longleftrightarrow} d_m \overset{L_{D-R}}{\longleftrightarrow} r_j$
3	*D-R-R-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{R-R}}{\longleftrightarrow} r_m \overset{L_{R-R}}{\longleftrightarrow} r_j$
3	*D-R-ADR-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{R-ADR}}{\longleftrightarrow} adr_m \overset{L_{R-ADR}}{\longleftrightarrow} r_j$
3	*D-ADR-D-R*	$d_i \overset{L_{D-ADR}}{\longleftrightarrow} adr_k \overset{L_{D-ADR}}{\longleftrightarrow} d_m \overset{L_{D-R}}{\longleftrightarrow} r_j$
3	*D-ADR-R-R*	$d_i \overset{L_{D-ADR}}{\longleftrightarrow} adr_k \overset{L_{R-ADR}}{\longleftrightarrow} r_m \overset{L_{R-R}}{\longleftrightarrow} r_j$
3	*D-ADR-ADR-R*	$d_i \overset{L_{D-ADR}}{\longleftrightarrow} adr_k \overset{L_{ADR-ADR}}{\longleftrightarrow} adr_m \overset{L_{R-ADR}}{\longleftrightarrow} r_j$

Given the topological features determined by meta paths, the associations between *D* and *R* can be measured by the popular indicator – Path Count (PC), which counts the number of path instances between two objects under a given meta path, denoted as PC_P, where *P* is the given meta path. Here we developed two metrics to reveal the associations between two objects d_i and r_j ($d_i \in D, r_j \in R$): given a meta path *P*, (1) Meta-Path-Indicator (MPI) indicates whether there exists a path $p \in P$ that associates d_i with r_j; (2) Path-Count-Indicator (PCI) counts the number of *p* that associates d_i with r_j.

$$MPI_P(d_i, r_j) = \begin{cases} 1, (\exists\, p \in P \Rightarrow d_i \overset{p}{\leftrightarrow} r_j = 1) \\ 0 \end{cases}$$

$$PCI_P(d_i, r_j) = \sum_{p \in P} 1, (d_i \overset{p}{\leftrightarrow} r_j = 1)$$

For example, $MPI_{D-R-R}(d_i, r_j) = 1$ denotes there is at least a path in the form of $d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{R-R}}{\longleftrightarrow} r_j$ that connects d_i with r_j.

3.2.3 Association rule mining - lift

In the above equations for MPI and PCI, the weights of links were not taken into consideration. In order to incorporate more network information, here we utilized association rule mining to measure the weights of associations and embedded the results into the computation of Path Count to develop a new metric.

In association rule mining, let $I = \{I_1, I_2, ..., I_m\}$ be a set of items and let $T = \{T_1, T_2, ..., T_n\}$ be a set of transactions, where each transaction is a subset of items such that $T_i \subseteq I$. An itemset that contains *k* items is a *k*-itemset; the occurrence frequency of an itemset is the number of transactions that contain the itemset. The association rule is an implication in the form of $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$, which is deemed as an itemset. In this study, *I* denotes the whole set that contains diseases, drugs, and ADRs, or *N* in the heterogeneous network definition; *T* denotes the dataset of all threads and each thread represents a transaction; there are both 1-itemset ({*D*}, {*R*}, {*ADR*}) and 2-itemset ({*D, D*}, {*R, R*}, {*ADR, ADR*}, {*D, R*}, {*D, ADR*}, {*R, ADR*}) involved in our calculation. The goal is to mine and evaluate the associations presented in 2-itemset.

Support is a common indicator used in association rule mining, defined as the percentage of transactions that contain 1-itemset or 2-itemset, for instance:

$$support\ (n_i) = \frac{count(n_i)}{total\ threads}$$

$$support\ (n_i, n_j) = \frac{count(n_i \cup n_j)}{total\ threads}$$

in which, $count(n_i)$ is the number of threads that contain target n_i; $count(n_i \cup n_j)$ is the number of threads that contain both n_i and n_j; *total threads* denotes the total number of threads. Nevertheless, for the 2-itemset, *support* is appropriate only when the co-occurrence frequency of the items is high. However, when consumers mention a drug, they might discuss different aspects of drugs, so that threads that are related to ADR only occupy a small portion in all the threads. To address this problem, another indicator *lift* is often used. *Lift* is a measure based on probability and reflects the division of the actual probability and theoretical probability. For instance, when measuring the strength of rule $R \Rightarrow ADR$, lift not only takes account of $support(R \cup ADR)$ but also the the correlation between 1-itemset *R* and 1-itemset *ADR*, by calculating the ratio of the proportion of threads containing both *R* and *ADR* above those expected if *R* and *ADR* are independent of each other. For a direct link $A_1 \leftrightarrow A_2$, the equation for calculating $support\ (n_i, n_j)$ $(n_i \in A_1, n_j \in A_2)$ is:

$$lift(n_i, n_j) = \frac{support(n_i, n_j)}{support(n_i) \times support(n_j)}$$

Given a meta path *P* in the form of $A_1 \overset{L_1}{\rightarrow} A_2 \overset{L_2}{\rightarrow} \dots \overset{L_l}{\rightarrow} A_{l+1}$, Path-Count-Lift (PCL) $PCL_P(n_1, n_l)$ $(n_1 \in A_1, \dots, n_l \in A_l)$ is calculated by:

$$PCL_P(n_1, n_l) = \sum_{p \in P} lift(n_1, n_2) \times lift(n_2, n_3) \times \dots \times lift(n_{l-1}, n_l), (n_1 \overset{p}{\leftrightarrow} n_l = 1)$$

Taking the meta path *D-D-R* for instance:

$$PCL_{D-D-R}(d_i, r_j) = \sum_{p \in P} lift(d_i, d_k) \times lift(d_k, r_j), (d_i \overset{p}{\leftrightarrow} r_j = 1)$$

in which, $d_i \overset{p}{\leftrightarrow} r_j = 1$ denotes there is a meth path p that associates d_i and r_j via $d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-R}}{\longleftrightarrow} r_j$.

3.3 Classification

In order to recognize associations between drugs and diseases from all the possible pairs, we developed a supervised learning model by training a binary classifier. The outputs of the classifier were 'positive' or 'negative', indicating whether the predicted drug-disease pair could be a known pair or not. The inputs of the classifier came from the meta-path-based topological features. For each drug-disease pair, we constructed a set of features that were used to predict whether this pair is possibly to be a known association. Under the length of three, we defined 13 meta paths between drugs and diseases; with *MPI* or *PCI* or *PCL* on each meta path representing a feature, it created a set of 13 features. As the supervised learning model required both positive and negative labels, we built the gold standard dataset of known drug-disease pairs by referring to PharmGKB and DrugBank

together, which suggested 2,087 known usages, and generated 28,000 negative instances randomly out of the unobserved associations between 50 diseases and 1,297 drugs.

In this work, we used a machine learning algorithm Random Forest as our classification method, which operates by constructing a multitude of decision trees that grow from bootstrap the training samples and outputs the labels based on the majority votes of the individual trees. In addition, considering there are much more negative instances than the positive in the dataset, we performed undersampling to avoid the imbalanced classification problem. Specifically, we divided the whole dataset into training (65%) and test (35%) sets firstly, then randomly split the negative pairs in the training set into 10 chunks. Each chunk and the positive training instances formed a sub-training set, where the classifier was trained and then tested on the hold-out test set.

3.4 Detection of off-label drug uses

Off-label drug uses are actually quite common in clinical practice and can be widely entrenched in certain clinical conditions. The document of medication providers such as EHRs and clinical notes provide the opportunity to detect off-label drug uses in an automated and scalable way. Meanwhile, from the perspective of medication receivers, they often talk about their drug uses and reactions on OHCs, which provides another resource to detect off-label uses. If the features extracted from user generated data enables us to recognize the known drug usages effectively, the other drug-disease pairs that present similar features with the known pairs are possibly to be off-label practices, in other words, the negative pairs that are falsely classified as "positive" are quite potential to be the off-label drug-disease usages. Therefore, the goal of this step is to identify the false positive (FP) predictions.

4. EXPERIMENTAL RESULTS

4.1 Classification results & evaluation

Evaluation of the classifier performance could indicate the effectiveness and accuracy of using network-based features to reveal drug-disease associations, furtherly, the effectiveness of identifying off-label drug-disease associations via this way [10]. In order to investigate the performance of different meta-path-based features, we did three experiments by using *MPI*, *PCI*, and *PCL* respectively. We trained the Random Forest classifiers on each sub-training set by performing 10-fold cross validation and evaluated the performance on the hold-out test set. Thus, we obtained 10 test results and took the average to represent the overall performance. Here, the classification performances were evaluated using Precision, Recall, and F1-score.

We trained and tested the Random Forest classifier with the *scikit-learn* tool in Python package [24]. The evaluation results of the binary classification were shown in Table 2. It is obvious that using MPI to represent the meta-path features obtained the worst results in all the three metrics, with F1-score equaling 0.639 in the test set. When incorporating the information about the number of meta paths between two objects, that is, using PCI to describe the meta-path features, the F1-score was improved by 18% compared with using MPI and achieved 0.754. Furtherly, after embodying the weights of links in the heterogeneous network and using PCL for features, the F1-score was improved by 36% compared with MPI and 15% with PCI. Besides, we obtained a very high Recall of 0.969, which means that among all the known drug-disease usages we classified 96.9% of them correctly. The Precision of 0.795 indicated that 79.5% of the predicted "positive" drug-disease pairs were indeed known usages, meanwhile, the other 21.5% were falsely classified as "positive" from the randomly generated negative pairs and might be the off-label pairs we were seeking. In summary, meta-path-based topological features have enabled us to develop successful supervised classification models for recognizing known drug-disease associations.

Table 2: Evaluation results of classification models built on different features

Feature	Dataset	Random Forest		
		Precision	Recall	F1 score
MPI	Training	0.670	0.713	0.692
	Test	0.618	0.663	0.639
PCI	Training	0.817	0.791	0.804
	Test	0.763	0.745	0.754
PCL	Training	0.897	0.971	0.932
	Test	0.795	0.969	0.870

As using PCL for features outperformed the other two indicators based on the evaluation results, we then applied the best trained classification model to the whole balanced dataset built by oversampling the minority class, to find all the possible off-label predictions. In result, we found 1,009 false-positive instances that might be potential candidates of off-label drug-disease associations, shown in Table 3.

Table 3: classification results of using PCL features

		Predicted	
		P	N
Actual	P	14480	300
	N	1009	26991

4.2 Validate results on PubMed and FAERS data

As the identified off-label usages have not been included in any pharmaceutical databases, the popular validation way is to find co-mention support from medical literature and reports [10]. The detected 1,009 potential candidates of off-label drug uses were checked for positive evidence in PubMed literature and FAERS reports. PubMed is a free resource developed and maintained by the National Center for Biotechnology Information (NCBI) and covers the titles and abstracts of more than 26 million biomedical publications, which can be accessed using the Entrez Programming Utilies (E-utilities). FAERS is the most important spontaneous reporting system as well as the primary data source for the study and identification of ADRs in United States. In PubMed, we found 407 of the novel predictions that have at least ten articles that both the drug and the disease were mentioned in

the abstract; in FAERS, we only found 10 of the novel predictions that have at least ten reports co-annotating with both the drug and the disease. The reason why there was so less evidence in FAERS might be that: first, FAERS is designed for reporting adverse drug effects rather than off-label uses, therefore, if there is no ADR involved, people have no intension to report their situations to the system; second, people report ADRs spontaneously and voluntarily, which leads to a surprisingly low reporting rate because of the nature of passiveness, with a median of 6%; third, it usually takes FDA a long time to complete the whole process of collecting reports, investigating cases and releasing alerts, which limits the manner of timely for information.

5. CONCLUSIONS

Off-label drug uses are very common and sometimes inevitable in clinical practice. Although the stakeholders including healthcare providers, patients, researchers, and medication manufacturers have the interest to gain information about off-label drug uses comprehensively and timely, there is no regulatory agency supervising such usages yet, which raises the demand for a systematic way to detect and manage off-label drug uses. The documents of medication providers such as EHRs and clinical notes provide the resource of detecting off-label uses and have been exploited in some previous studies, meanwhile, the large volumes of data coming from medication receivers also render the opportunity to detect off-label uses in a systematic and scalable way. In this study, we developed an automated way to detect off-label uses based on heterogeneous network mining. With data collected from a popular OHC - MedHelp, we extracted the medical entities of diseases, drugs, and ADRs with lexicon-based approaches and constructed a heterogeneous healthcare network that contained three types of nodes and six types of edges. On the heterogeneous network, we determined 13 meta paths between drugs and diseases and developed three metrics to describe the meta-path-based topological features: Meta-Path-Indicator (MPI), Path-Count-Indicator (PCI), and Path-Count-Lift (PCL). Then we utilized these features as inputs for the Random Forest classifier to recognize the known drug-disease associations from all the possible pairs. Using MPI, PCI, and PCL for the features respectively, we conducted three classification experiments, and the classification model built on PCL achieved the best performance with F1-score reaching 0.87. The results showed that meta-path-based topological features could enable us to develop well-performed supervised classification models to recognize known drug-disease associations, furtherly, the other drug-disease pairs that present similar features with those known pairs are possibly to be the off-label practices, or the false-positive predictions are quite potential to be off-label drug-disease usages. Based on this hypothesis, we identified 1,009 candidates of off-label drug uses and examined them for positive support from PubMed and FAERS. In result, 407 of them were found evidence from at least ten articles in PubMed and 10 of them were found evidence from at least ten reports in FAERS.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under the Grant IIS-1650531 and DIBBs-1443019.

REFERENCES

[1]. Wittich, C. M., Burkle, C. M., & Lanier, W. L. (2012, October). Ten common questions (and their answers) about off-label drug use. In Mayo Clinic Proceedings (Vol. 87, No. 10, pp. 982-990). Elsevier.

[2]. Kao, J. (2016). White Paper: Pharmaceutical Regulation and Off-Label Uses.

[3]. Questions and Answers on FDA's Adverse Event Reporting System (FAERS). https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugeffects/

[4]. Barry, M. M., Domegan, C., Higgins, O., & Sixsmith, J. (2011). A literature review on health information seeking behaviour on the web: a health consumer and health professional perspective.

[5]. White, R. W., Tatonetti, N. P., Shah, N. H., Altman, R. B., & Horvitz, E. (2013). Web-scale pharmacovigilance: listening to signals from the crowd. Journal of the American Medical Informatics Association, 20(3), 404-408.

[6]. Conroy, S., Choonara, I., Impicciatore, P., Mohn, A., Arnell, H., Rane, A., ... & Rocchi, F. (2000). Survey of unlicensed and off label drug use in paediatric wards in European countries. Bmj, 320(7227), 79-82.

[7]. Leslie, D. L., & Rosenheck, R. (2012). Off-label use of antipsychotic medications in Medicaid. The American journal of managed care, 18(3), e109-17.

[8]. Mesgarpour, B., Müller, M., & Herkner, H. (2012). Search strategies to identify reports on "off-label" drug use in EMBASE. BMC medical research methodology, 12(1), 190.

[9]. Jung, K., LePendu, P., & Shah, N. (2013). Automated detection of systematic off-label drug use in free text of electronic medical records. AMIA Summits on Translational Science Proceedings, 2013, 94.

[10]. Jung, K., LePendu, P., Chen, W. S., Iyer, S. V., Readhead, B., Dudley, J. T., & Shah, N. H. (2014). Automated detection of off-label drug use. *PloS one*, *9*(2), e89324.

[11]. Fung, K. W., Jao, C. S., & Demner-Fushman, D. (2013). Extracting drug indication information from structured product labels using natural language processing. Journal of the American Medical Informatics Association, 20(3), 482-488.

[12]. Xu, R., & Wang, Q. (2013). Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC bioinformatics, 14(1), 181.

[13]. Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular systems biology, 7(1), 496.

[14]. Huang, Y. F., Yeh, H. Y., & Soo, V. W. (2013). Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. BMC medical genomics, 6(3), S4.

[15]. Chen, X., Liu, M. X., & Yan, G. Y. (2012). Drug–target interaction prediction by random walk on the heterogeneous network. Molecular BioSystems, 8(7), 1970-1978.

[16]. Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., & Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. BMC medical genomics, 8(2), S2.

[17]. Yang, H., & Yang, C. C. (2016, October). Discovering Drug-Drug Interactions and Associated Adverse Drug Reactions with Triad Prediction in Heterogeneous Healthcare Networks. In Healthcare Informatics (ICHI), 2016 IEEE International Conference on (pp. 244-254). IEEE.

[18]. Yan, X. Y., Zhang, S. W., & Zhang, S. Y. (2016). Prediction of drug–target interaction by label propagation with mutual interaction information derived from heterogeneous network. Molecular BioSystems, 12(2), 520-531.

[19]. Soliman, T. H. A. (2014). Mining Multi Drug-Pathways via A Probabilistic Heterogeneous Network Multi-label Classifier. Bonfring International Journal of Research in Communication Engineering, 4(2), 10.

[20]. Lee, K., Lee, S., Jeon, M., Choi, J., & Kang, J. (2012, October). Drug-drug interaction analysis using heterogeneous biological information network. In Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on (pp. 1-5). IEEE.

[21]. Jiang, L., & Yang, C. C. (2015, March). Expanding Consumer Health Vocabularies by Learning Consumer Health Expressions from Online Health Social Media. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 314-320). Springer International Publishing.

[22]. Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, *13*(1), 24-29.

[23]. Han, J., Sun, Y., Yan, X., & Yu, P. S. (2010, July). Mining heterogeneous information networks. In Tutorial at the 2010 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'10), Washington, DC.

[24]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.