ARTICLE

DrugMetab: An Integrated Machine Learning and Lexicon Mapping Named Entity Recognition Method for Drug Metabolite

Heng-Yi Wu[1], Deshun Lu[2], Mustafa Hyder[3], Shijun Zhang[1], Sara K. Quinney[3], Zeruesenay Desta[3] and Lang Li[1,*]

Drug metabolites (DMs) are critical in pharmacology research areas, such as drug metabolism pathways and drug-drug interactions. However, there is no terminology dictionary containing comprehensive drug metabolite names, and there is no named entity recognition (NER) algorithm focusing on drug metabolite identification. In this article, we developed a novel NER system, DrugMetab, to identify DMs from the PubMed abstracts. DrugMetab utilizes the features characterized from the Part-of-Speech, drug index, and pre/suffix, and determines DMs within context. To evaluate the performance, a gold-standard corpus was manually constructed. In this task, DrugMetab with sequential minimal optimization (SMO) classifier achieves 0.89 precision, 0.77 recall, and 0.83 F-measure in the internal testing set; and 0.86 precision, 0.85 recall, and 0.86 F-measure in the external validation set. We further compared the performance between DrugMetab and whatizitChemical, which was designed for identifying small molecules or chemical entities. DrugMetab outperformed whatizitChemical, which had a lower recall rate of 0.65.

CPT Pharmacometrics Syst. Pharmacol. (2018) **7**, 709–717; doi:10.1002/psp4.12340; published online on 29 September 2018.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?
☑ There is no terminology dictionary containing comprehensive DM names, and there is no NER algorithm focusing on DM identification.

WHAT QUESTION DID THIS STUDY ADDRESS?
☑ First, four different DM presentation patterns were defined. Second, a gold-standard corpus was constructed to annotate DMs with the relationship between parent drugs and their metabolites within the abstracts. Third, a new DM, the NER algorithm, was developed to identify DMs in the scientific literature. Fourth, different from all the other NER systems, DrugMetab can not only annotate metabolite terms but also recognize metabolism reactions related to their parent drugs.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?
☑ It facilitates the exploration of relationships between drugs and their metabolites.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?
☑ It facilitates the exploration of drug-drug interaction in scientific literature.

A drug's pharmacokinetics (PKs) involves not only the parent compound, but also its metabolites.[1] For instances, codeine drugs have active metabolites (morphine) that possess more therapeutic activity against the targeted protein than its parent drug.[2] In addition, drug metabolites (DMs) play prominent roles in drug interactions. A notable example is itraconazole. Itraconazole itself is a potent cytochrome P450 (CYP)3A inhibitor, so are its metabolites, such as hydroxy-itraconazole.[3] Pharmacogenetics also has a major impact on the drug metabolism products. The tamoxifen active metabolite, endoxifen, is generated through the CYP2D6 enzyme. Among patients with breast cancer with CYP2D6 loss functional variants (e.g., *4, *5, and *10), the patients usually have very limited tamoxifen metabolite, endoxifen. Hence, these patients have much reduced endoxifen concentration such that the efficacy of tamoxifen treatment declined.[4] All these above examples demonstrate that DMs and their parent drugs are equally important in PK research.

Although there are several well-established dictionaries for drug and metabolome, the resource for DMs is still limited. The Human Metabolome Database (HMDB) reports data on >29,000 endogenous metabolites, but there are only

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, Ohio, USA; [2]Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, Indiana, USA; [3]Division of Clinical Pharmacology, School of Medicine, Indiana University, Indianapolis, Indiana, USA. *Lang Li (lang.li@osumc.edu)

Received 22 January 2018; accepted 25 June 2018; published online on 29 September 2018. doi:10.1002/psp4.12340

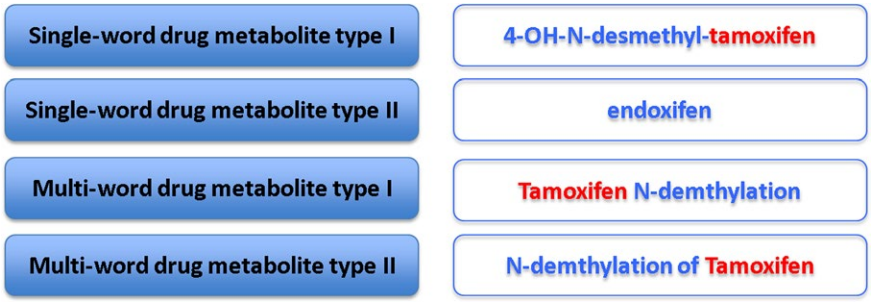


Figure 1 Patterns of drug metabolites. There are four patterns that show how the drug metabolites are presented. Tamoxifen metabolites are used as the primary example for the demonstration. Type I contains a substring of a drug name as well as a chemical prefix or suffix that represents its drug metabolism chemical reactions. Type II, however, does not contain either a substring of its parent drug or chemical reaction. It can be either an abbreviation or an unrelated name. The other two patterns are represented with the form of multiword entities containing either a preposition (type I) or conjunction for describing the drug metabolism chemical reaction (type II). ML, Machine Learning; POS, Part-Of-Speech.

2,485 drugs, and 948 DMs.[5] Similarly, DrugBank 4.0[6] and ChEBI[7] have 1,912 and 112 DMs, respectively, which are much less than the total number of generic drugs (8,184). Those numbers represent the gaps between drugs and their metabolites. For instance, two metabolites, desmethylflunitrazepam and 3-hydroxyflunitrazepam, can be found in PMID: 11259331. However, they are not available in either DrugBank or ChEBI. To solve this problem, biomedical literature is a good data resource to deliver high-quality information and text mining is the technology to transfer the information into a system. Among those works, named entity recognition (NER) is an initial and crucial processing step.

There are many NERs that annotate text with biomedical terminologies.[8–25] Some were designed to identify general terms, like proteins, DNA, RNA, cells, cell lines, etc.,[8–10,16,26] and some can annotate drugs, chemicals, or metabolome.[11–13,17] However, only a few using the dictionary lookup approach can annotate DMs. For instance, whatizitChemical is an NER system that can annotate DMs, if it selects dictionaries, like ChEBI or OSCAR3, containing DM names.[22]

To conquer the challenges above, first, four different DM presentation patterns were defined, and a gold-standard corpus was constructed. This annotated corpus facilitates the next step in DrugMetab development. Second, DrugMetab was proposed to identify DMs in biomedical literature.

METHODS

Define drug metabolite and reaction

The annotations for drug metabolism products (i.e., DMs), on the other hand, are not well investigated or integrated yet. To demonstrate the challenges in annotating DMs, we illustrated four patterns how DMs are presented in literature. Tamoxifen metabolites are used as the primary example for the demonstration in **Figure 1**. The first two categories (Single_Word_Drug_Metabolite type I and type II) are DM names in a single entity. Type I contains a substring of a drug name as well as a chemical prefix or suffix that represents its drug metabolism chemical reactions (e.g., 4-OH-N-desmethyltamoxifen in PMID: 15685451). Type II, however, does not contain either a substring of its parent drug or chemical reaction. It can be either an abbreviation or an unrelated name. For instance, endoxifen (PMID: 20400308) is the primary active metabolite of tamoxifen via CYP2D6 enzyme, which has an alternative name of 4-OH-N-desmethyltamoxifen. The other two patterns are represented with the form of multi-word entities containing either a preposition (type I) or conjunction for describing the DM chemical reaction (type II). The examples of Multi_Word_Drug_Metabolite type I and type II are tamoxifen N-demethylation in PMID: 24737844 and N-demethylation of tamoxifen in PMID: 8104124, respectively. Therefore, to characterize and extract DM names from biomedical literature, we will need two different but highly connected informatics tools: a DM annotation scheme and its NER algorithm. The DM annotation scheme shall characterize the substrings for the DM chemical reactions in annotating DMs and the relationship to its parent drug. Utilizing the annotations in the corpus, the NER can learn from annotations how DMs are presented in literature and facilitate the recognition in the unread text.

Corpus construction

Annotation material

DM corpus was constructed with 210 *in vitro* PK abstracts and 45 drug-drug interaction (DDI) abstracts. A detail of how to collect those documents can be found in ref. 27. To annotate DMs, a web-based tool, called *Brat* available in (http://brat.nlplab.org/index.html), was utilized.[28] Brat is an annotation tool, which is popular for annotations in many aspects, including term, relationship, event annotation, etc.

Annotation guideline

The guideline was generated to provide annotators a standard for annotation. It focused solely on the annotation for DM entities. The cognitive process is based on the context of an abstract. The annotated entity must be referred as if it was either a product or a reaction activity of a drug via the metabolism process. As we described in "Define drug metabolite and reaction," four entity types were proposed to annotate DMs.

Annotation process

The corpus construction is a manual process (**Figure 2**). Three annotators with different training backgrounds,

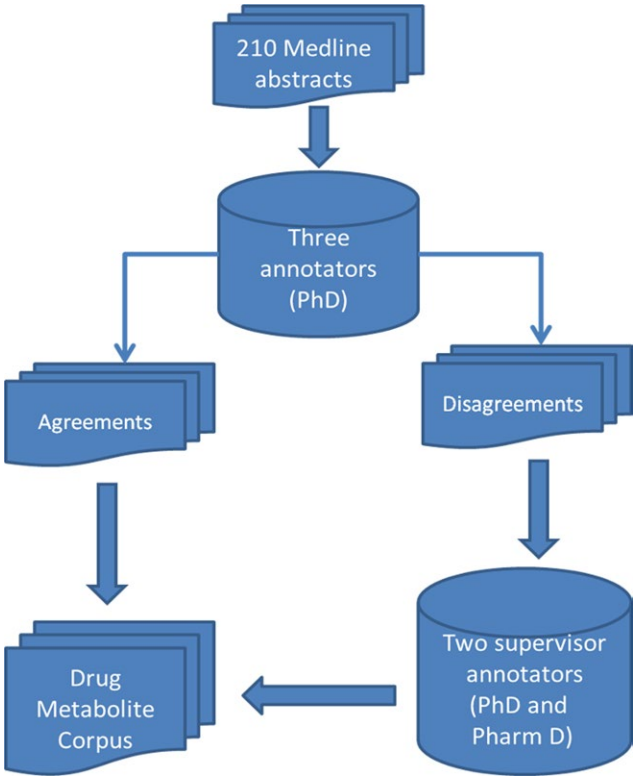


Figure 2 Drug metabolite annotation flow chart. The corpus construction is a manual process by three annotators. The result was created based on the consensus. The disagreed annotations among three annotators were judged by pharmacological research experts for the final decision.

including informatics, biochemistry, and pharmacology, conducted the annotation tasks independently. An agreed annotation was defined when its offset information is identical among three annotators. The offset information contains start-offset and end-offset. The start-offset is the index of the first character of the annotated span and the end-offset is the index of the last character of the annotated span in the text. If spans for a tag are overlapping but its start or end offset locations are different, it means three annotators agreed with this tag but disagreed with its offset location. In this case, it was solved among three annotators. If there exists two or less spans for a tag, in this case, the disagreed annotations are first discussed among three annotators for consensus. If the consensus is still not achieved, the disagreed annotations are further judged by pharmacological research experts (Professors in the Department of Pharmacology) for the final decision.

Gold-standard corpus

Once the gold-standard corpuses were created from the annotation procedure, then annotated text files were converted into GENIA format, invented by Tsujii Laboratory of University of Tokyo.[29] This corpus format was initially created to support the development and evaluation of information extraction and text mining system for the domain of molecular biology. Within corpus, the annotation for DMs was made with start-tag and end-tag and the names of their parent drugs were also embedded within the start-tag. In this way, the relationship between the parent drugs and their metabolites can be built. The data is available in the **Data S1**.

Annotation evaluation

To evaluate the consistency and quality of the annotation task, two types of measurements were calculated. First, pairwise percent agreement was used to measure the agreements between two annotators. Second, the results from three annotators are compared to the gold-standard corpus. Precision (P), recall (R), and F-measure (F) are adopted to assess the performance of an individual annotator.

Drug and drug metabolism reaction lexicon

Two lexica are built up, including the drug name lexicon and the DM reaction lexicon. The drug name lexicon is built upon the drug names in Drugbank 4.0[30] and the medical subject heading term.[31] In total, there are 70,712 unique drug names in the drug name lexicon, which is available in the **Data S2**.

The DM reaction lexicon are composed of 65 metabolites' prefix and suffix terms collected from the literature[32,33] and our previous work.[27] They are further evaluated by two domain experts. Within the lexicon, DM reactions are categorized into two groups: modification (phase I) and conjugation (phase II) reactions. The DM reaction lexicon is available in the **Data S3**.

DrugMetab: An integrated drug metabolite NER algorithm

DrugMetab has three phases, and the workflow is shown in **Figure 3**. In the first phase, drug names, their prefix/suffix, and their abbreviations are tagged and indexed in each abstract. In addition, Part-of-Speech (PoS) information was provided to illustrate the sentence structure grammatically. In the second phase, a searching window is created centering at a drug name entity. In the third phase, a machine-learning algorithm will be trained using the feature matrix created in the second phase. It predicts whether the candidate entities in the searching window are DMs or not.

Phase I: Part-of-Speech tagging

The Part-of-Speech Tagger in OpenNLP[34] was implemented for creating PoS features for entities. With the Penn Treebank tag set,[35] the English maxent PoS model in PoS Tagger read a tokenized sentence each time and echoed the sentence with PoS tags. In this step, some erroneous tags for drug names and reaction terms were manually modified.

Phase I: Lexicon-based tagging

A dictionary-based tagging is applied to identify whether an entity is a drug name, and whether a drug name and a metabolite's reaction term are the substring of that entity. Technically, drug names in lexicon are sorted based on the length of string in the hash table. Then, if an entity can be partially mapped against a drug name or a reaction term in the lexicon table, a drug name or pre/suffix name for that entity is annotated. However, some entities might be

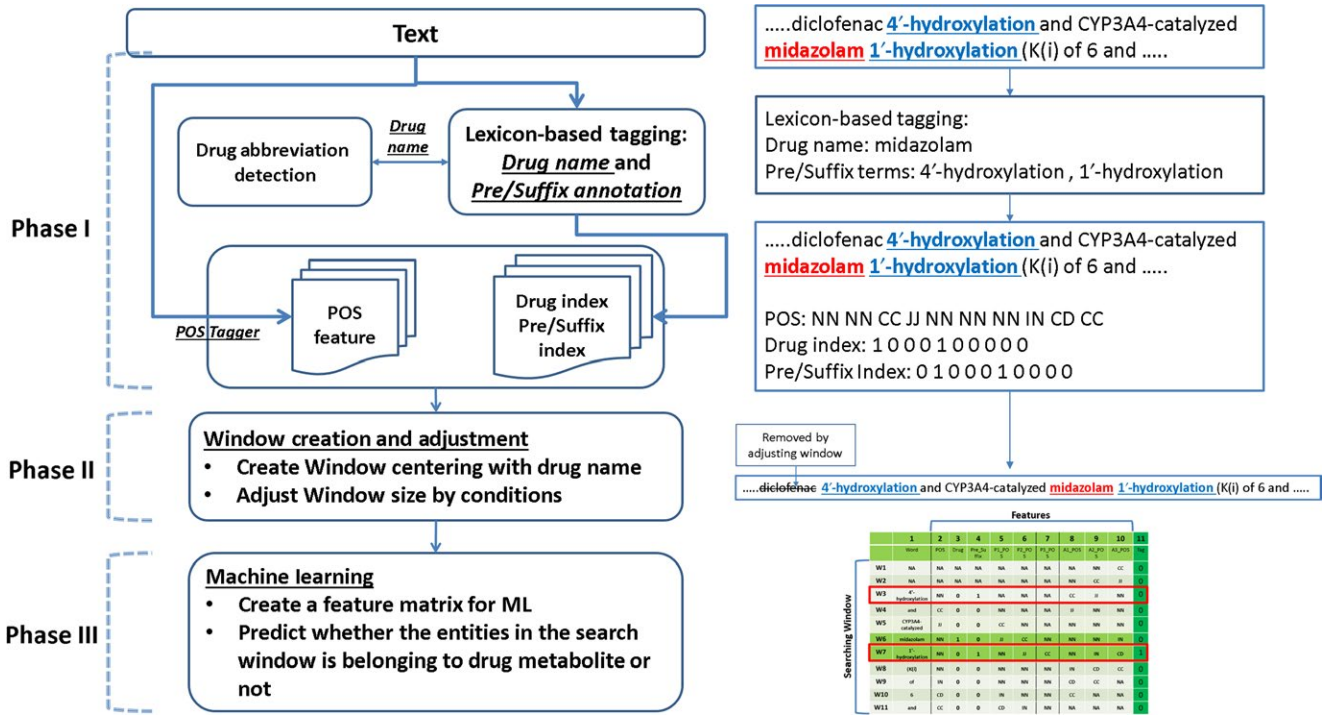


Figure 3 Workflow of drug metabolite annotation. DrugMetab has three phases in workflow. In the first phase, drug names, their prefix/suffix, and their abbreviations are tagged and indexed in each abstract. In the second phase, a searching window is created centering at a drug name entity. In the third phase, a machine learning algorithm will be trained using the feature matrix created in the second phase.

erroneously tagged because of some special brand names. For instance, "Control" is a brand name of chlordiazepoxide. To eliminate such false-positive results, these tags will be removed if the term was recognized as a verb with the PoS tagger.

Phase I: Detect drug abbreviations

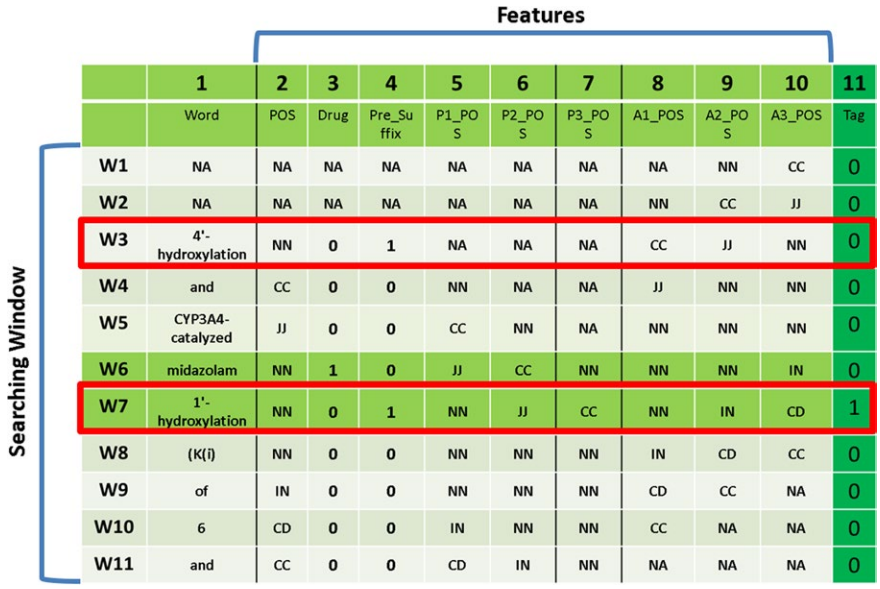
In PK studies, a drug abbreviation is usually annotated in a parenthesis after its full name is presented the first time in an abstract. This algorithm for detecting drug abbreviations first searches for the existence of parentheses after the tagged drug names in a range of five words. However, not all terms within parentheses are drug abbreviations. They might be an enzyme name (e.g., CYP3A4), drug dosage, or drug serum concentration in a PK experiment (e.g., 10 μM), PK parameters measured in a PK experiment (e.g., half-maximal inhibitory concentration), and statistical results (i.e., *P* value or confidence interval). These terms are then filtered using the technologies of regular expression. The regular expressions of these terms are well defined in our previously PK corpus.[27] Once the abbreviations of drug names were recognized, they were tagged as drug names.

Phase II: Construct window around a drug name

First, a window size of $2*n + 1$ (n is one-sided word span) is placed centering on the tagged drug name. To optimize DM identification, different window sizes were evaluated using our gold-standard corpus. Here, we have investigated the span size $n = 2, 3, 4, 5$, and 6. The best coverage (optimal recall rate ~100%) was obtained using the window of size 11 (span size $n = 5$). Second, the window is further trimmed according to the following rules: the window meets the start or end of a sentence; the window overlaps with another drug name; and the window meets the entity ending with a comma.

Phase III: Create the input feature matrix for the machine-learning algorithms

Three types of input features (PoS tags, drug indices, and metabolism reaction indices) from phase I were used to build a feature matrix for a searching window. **Figure 4** shows an example of a feature matrix created for the machine learning. In the rows of this matrix, there are 11 words (W1–W11) within a window. For those words assigned with "NA" (W1 and W2 in **Figure 4**), they are removed during the window size adjustment (see phase II: Construct window around a drug name). For the machine-learning prediction, 9 features (2nd to 10th columns) were created for each word. First, the *Part-of-Speech tag* (second column) represents the grammatical category for each word. *Drug index* (third column) means the availability of a drug name in each word. Midazolam (W6) in the center of the searching window is indexed as 1 because midazolam was recognized as a drug name. Metabolism reaction index (fourth column) represents the availability of a reaction term in each word. For two instances in **Figure 4**, both 4′-hydroxylation (W3) and 1′-hydroxylation (W7) are indexed with 1 because they contain a predefined reaction term (hydroxyl). PoS tags for the surrounding entities (±*N* words) of the current word were created to represent the syntactic environment



	1	2	3	4	5	6	7	8	9	10	11
	Word	POS	Drug	Pre_Suffix	P1_POS	P2_POS	P3_POS	A1_POS	A2_POS	A3_POS	Tag
W1	NA	NA	NA	NA	NA	NA	NA	NA	NN	CC	0
W2	NA	NA	NA	NA	NA	NA	NA	NN	CC	JJ	0
W3	4'-hydroxylation	NN	0	1	NA	NA	NA	CC	JJ	NN	0
W4	and	CC	0	0	NN	NA	NA	JJ	NN	NN	0
W5	CYP3A4-catalyzed	JJ	0	0	CC	NN	NA	NN	NN	NN	0
W6	midazolam	NN	1	0	JJ	CC	NN	NN	NN	IN	0
W7	1'-hydroxylation	NN	0	1	NN	JJ	CC	NN	IN	CD	1
W8	(K(i)	NN	0	0	NN	NN	NN	IN	CD	CC	0
W9	of	IN	0	0	NN	NN	NN	CD	CC	NA	0
W10	6	CD	0	0	IN	NN	NN	CC	NA	NA	0
W11	and	CC	0	0	CD	IN	NN	NA	NA	NA	0

Figure 4 Feature matrix of entities in a searching window for machine learning. An example of a feature matrix was created for machine learning. CC, Coordinating Conjunction; CD, Cardinal number; IN, Preposition; JJ, Adjective; NA, Not Available; NN, Noun; POS, Part-Of-Speech.

around the target word. In this experiment, $N = 3$ was determined based on the histogram of suffix or prefix terms in the training dataset, which covers 95% of drug and suffix/prefix terms combinations. Taking 1′-hydroxylation (W7) as the example in **Figure 4**, P1_POS, P2_POS, and P3_POS (5th to 7th columns) represent the PoS tags of one, two, and three words before 1′-hydroxylation (W7), respectively. On the other hand, A1_POS, A2_POS, and A3_POS (8th to 10th columns) represent the PoS tags of one, two, and three words after 1′-hydroxylation (W7), respectively. Final column (Tag) is used to identify whether the words (W3 or W7) containing reaction terms are part of metabolism reactions for a DM name; and in which "1" means yes, and "0" means no. It is the outcome variable that the machine-learning algorithms are either trained with or tested against.

Phase III: Machine-learning algorithms

The aim of this work is to predict whether the candidate entities (words with metabolism reaction terms) in the searching window are part of a DM or not. With the feature matrix generated from phase II, sequential minimal optimization (SMO),[36] J48,[37] and logistic model tree (LMT)[38] with the default parameter setting in Weka 3.8 was utilized to accomplish this task.[39] For the experimental setting, among 210 *in vitro* PK abstracts,[27] 168 abstracts were used to build the training model, 42 abstracts were used as internal validation, and 45 DDI abstracts were used for external validation.[40] Tenfold cross-validation was used in building up the training mode.

Phase III: Prediction performance evaluation

To evaluate DrugMetab's performance, the predicted DMs that matches both start and end positions in gold-standard corpus constitute true-positive results. The predicted DMs that do not match fully are false-positive results; and DM terms in the corpus that are not be predicted are false-negative results. Finally, the information-retrieval metrics: precision, recall, and F-measure are used for evaluation.

Comparison DrugMetab with whatizitChemical

To discover exiting NER systems for performance comparison, there is no one focusing on the annotation for DMs. One similar work done by Nobata *et al.*[17] proposed an NER tool to extract yeast metabolites using ChEBI and HMDB terms. This work compared their performance with whatizitChemical and demonstrated that whatizitChemical achieved lower precision and F-measure compared to their NER tool. Therefore, we recognize whatizitChemical can be a baseline for evaluation. WhatizitChemical is one of the modules in the Whatizit pipeline that analyzes text data based on TreeTagger.[22] By integrating both drug (WhatizitChebiDict) and chemical (whatizitOSCAR3) dictionaries, whatizitChemical can identify chemical and drugs names because whatizitChebiDict annotates DMs in the ChEBI. In this analysis, we compared the performance of DrugMetab with that of whatizitChemical.

Online materials

The gold-standard corpus for DMs is available in the **Data S1**. For the drug dictionary, metabolism reaction terms, and codes, they can be found in the **Data S2**, **S3**, and **S4**, respectively.

RESULTS

Performance of corpus construction

The measurement of inter-annotator agreement between two annotators was quantified using pairwise percent agreement (**Table 1**). The pairwise percent agreement suggests that a high level of agreement (87.6–89.8%) among three annotators was achieved. In addition, annotations

Table 1 Annotation performance evaluation

Evaluation type 1	Pairwise percent agreement (%)
Annotator 1–2	87.6
Annotator 1–3	88.8
Annotator 2–3	89.8

	Comparison between gold-standard corpus and the result of each annotator		
Evaluation type 2	**Precision** = $\frac{TP}{TP+FP}$	**Recall** = $\frac{TP}{TP+FN}$	**F-measure** = $\frac{2*TP}{2*TP+FP+FN}$
Annotator 1	0.989	0.913	0.950
Annotator 2	0.994	0.936	0.964
Annotator 3	0.986	0.97	0.978

FN, false negative; FP, false positive; TP, true positive.

are compared between annotators and the gold-standards using precision, recall, and F-measure in **Table 1**. The evaluation suggested that three annotators have comparable curating performance, where the F-values are 0.95, 0.964, and 0.978, respectively.

There are some disagreements due to the lack of clarity of the annotation guideline. For example, when an abbreviation was mentioned right behind its DM name, one annotator annotated both DM and its abbreviation as a tag, but another annotator only annotated DMs and ignored the abbreviation part. Consequently, two annotators consistently differed in the Single_Word_Drug_Metabolite type II. In this analysis, most disagreements between annotator 1 and annotator 2 occurred in this category. For instance, in PMID: 10859153, both annotator 1 and annotator 2 omitted "NORCIS," which is the abbreviation of norcisapride. In addition, many DMs written in the mixture form of drug abbreviation and a reaction term were missed (e.g., 3-hydroxyNVP (a metabolite of Nevirapine) in PMID: 10570031). Another frequent error is the unique drug metabolite names. For example, dihydroqinqhaosu in PMID: 10456689 is an active metabolite of artelinic acid. However, it is challenging to find clues to connect it to its parent drug.

Performance of the entity tagging

In the gold-standard corpus, there are 3,789 annotated drug entities. Only three drug entities (two unique drugs) were not tagged by our tagging algorithm because their drug names were not available in our dictionary. These two drug names are RPR-106541 and cholantene. There are 1,582 entities containing the metabolite reaction terms in our corpus, and only 7 were erroneously tagged by our tagging algorithm. This is because some drug names or their synonyms have our proposed pre/suffix terms. In tagging drug abbreviations, 138 true drug abbreviations are captured, and we have only 9 false positives. Overall, the tagging performance of drug entities and pre/suffix terms is very good. It minimizes the effect of error propagation due to the erroneous entity tagging.

Performance of DrugMetab on the internal validation *in vitro* PK abstracts

From 210 abstracts, 168 abstracts are selected for training, and the remaining 42 abstracts (internal validation) are used for testing. Using 10-fold cross-validation, three different algorithms in Weka pipeline[39] were trained: SMO, J48, and LMT. Their performances on the testing dataset were evaluated. The best precision (0.89), recall (0.77), and F-measure (0.83) are achieved by the SMO. The J48 tree has comparable precision (0.88) but much lower recall (0.62) and F-measure (0.73). The LMT obtained the worst performance with precision (0.85), recall (0.57), and F-measure (0.68).

Performance of DrugMetab on the external validation DDI abstracts

To further evaluate DrugMetab, 45 DDI abstracts containing DMs were used as an external validation data.[40] In this dataset, there are 233 drug metabolites, including 100 Single_Word_Drug_Metabolite type I, 95 Multiple_Word_Drug_Metabolite type I, and 38 Multiple_Word Drug_Metabolite type II. The overall performance of DrugMetab F-measure is 0.86. The precision and recall for Single_Word_Drug_Metabolite type I, Multi_Word_Drug_Metabolite type I, and Multi_Word_Drug_Metabolite type II are (92.6%/88%), (81.6%/84.2%), and (79.5%/81.6%; **Table 3**), respectively.

Compare with whatizitChemical

WhatizitChemical was designed to identify chemical entities, drugs, and protein names for EBIMed individually, but was not designed to identify DMs. For example, in PMID: 10460803, "dextromethorphan o-demethylation" is a metabolite of dextromethorphan. Using whatizitChemical, dextromethorphan and o-demethylation were tagged as a drug and a chemical, respectively. Thus, it is difficult to compare whatizitChemical to DrugMetab directly because they have different annotation criteria. Here, we assume whenever whatizitChemical correctly annotates both drug term and its metabolism reaction term, we treat it as a true-positive result. Otherwise, it is a false-negative result. To make a fair comparison, we only count the number of true-positive results and false-negative results of terms in gold-standard corpus, and calculate their recall rates. Overall, our result shows that DrugMetab has a recall of 0.77, whereas whatizitChemical has a recall of 0.65. In addition, **Table 2** compares their recall rates in each type of DM. Except for Single_Word_Drug_Metabolite type I, DrugMetab with SOM can outperform whatizitChemical in the rest of the categories.

Table 2 The comparison of DrugMetab using SMO algorithm with whatizitChemical on *in vitro* PK test data and reasons of errors in each type of drug metabolites

Recall = $\frac{TP}{TP+FN}$/Precision = $\frac{TP}{TP+FP}$	DrugMetab with SOM (recall/precision)	whatizitChemical (Recall = $\frac{TP}{TP+FN}$)	Reasons of errors
Single word drug metabolite type I	92.7%/91.3%	98.5%	• Drug name is not in dictionary • Error from machine learning
Single word drug metabolite type II	32.3%/87.5%	15.4%	• Unidentified drug abbreviations • Metabolite-like names • Unique drug metabolite names
Multiword drug metabolite type I	93.8%/96.3%	70.5%	• Unidentified drug abbreviations • Error from machine learning • Drug name is not in dictionaries
Multiword drug metabolite type II	77.3%/89.5%	65.1%	• Unidentified drug abbreviations • Error from machine learning • Drug name is not in dictionaries

FP, false positive; PK, pharmacokinetic; SMO, sequential minimal optimization; TP, true positive.

DISCUSSION

Error analysis

In the error analysis, a manual check was performed to investigate the causes of errors on internal validation dataset. In **Table 2**, we recognize five major reasons of errors for each type of DM, including unidentified drug abbreviations, misclassifications from the machine learning, metabolite-like names, unique DM names, and drug names are not in the dictionary. Unidentified drug abbreviations account for about 44% of errors. Misclassifications by the machine learning have the second highest error annotations (32%). False-positive results occurred when their PoS patterns are similar to that of true DMs. For example, "hydroxylation *in vitro* by nelfinavir" in PMID: 11159797 has a similar PoS pattern (NN_reaction + IN_by + NN_drug) to that of drug reaction type II (NN_reaction + IN_of + NN_drug). False-negative results occurred when using a long phrase to represent Multi_Word_Drug_Metabolite type II (e.g., "N-demethylation of rac-, (R)- and (S)-methadone" in PMID: 10233205 and "N-dealkylation of the antipsychotic drug perphenazine" in PMID: 11136295). The third reason is metabolite-like names, which accounts for 10% of errors. For instance, "dihydroergotamine" is recognized as the metabolite of a drug name ("ergotamine"), but it is a generic drug name. The fourth reason is a unique DM name, which accounts for 8% of errors. For example, UK-103 320 in PMID: 11298070 (the main metabolite of sildenafil) and cycloguanil in PMID: 9923577 (the metabolite of proguanil) are not identified because they are not named based on their parent drug and do not exist in our dictionary.

We further investigated DrugMetab performance in four different patterns of DMs. **Table 2** shows the recall and precision rates of the DrugMetab using SMO. The SMO performs the best in Multi_Word_Drug_Metabolite type I with R: 93.8% and P: 96.3%. Single_Word_Drug_Metabolite type II and Multi_Word_Drug_Metabolite type II have slightly worse performance with R: 92.7%/P: 91.3% and R: 77.27%/P: 89.47%, respectively. However, for Single_Word_Drug_Metabolite type II, a poor recall rate of 32.3% was obtained. Based on our observation, the best performance of predicting Multi_Word_Drug_Metabolite type I is its simpler structure. Both drug entity and reaction entity in this category are assigned to a grammatical category of noun (NN), and they are laid side by side (i.e., NN_drug + NN reaction). On the other hand, the unfavorable DrugMetab result of Single_Word_Drug_Metabolite type II is primarily due to the unidentifiable drug or metabolite names or their abbreviations. It can probably be solved by manual curation.

Performance of DrugMetab without single word DM type II

For Single_Word_Drug_Metabolite type II, the major reason of this poor prediction is that there is no standard naming clue. First, it is caused by the abbreviation of DM itself. For example, DQHS is the abbreviation of dihydroqinghaosu, which is the active metabolite of artelinic acid. The second error type is rare reaction terms. For example, cycloguanil is the metabolite of proguanil. Although they have the same six letters (guanil) within the name, cyclo and pro are hard to identify as a suffix or prefix string in our dictionary. The third error type is a unique name for the DM. For instance, UK-103 320, which is the metabolite of sildenafil, has no suffix/prefix string for representing its pathway. Thus, there is no clue to connect it to its parent drug.

Due to these challenges, we decided to rebuild a training model and study whether the performance can be improved without Single_Word_Drug_Metabolite type II. Comparing **Table 3** to **Table 2**, DrugMetab has improved precision and recall rates in all three DM types.

Practical usage

In our work,[40] DrugMetab were applied to improve DDI identification from biomedical literature. In reality, many DDI signals can be identified indirectly via DMs. For instance, endoxifen but not tamoxifen interacts with estrogen receptor alpha literally in a sentence.

In addition, a new tool can be innovative in improving the performance of DM NER in two aspects. First, many DMs are related with their parent drugs through chemical reactions via drug metabolism enzymes. DrugMetab can build the relationship between a parent drug and their metabolites, which is valuable in enriching some existing databases, such as Drugbank. In addition, utilizing such a relationship

Table 3 The recall and precision rates of the DrugMetab on the internal and external validation dataset without single-word drug metabolite type II

Drug metabolite types	DrugMetab with SMO (Recall = $\frac{TP}{TP+FN}$)/precision = $\frac{TP}{TP+FP}$)	
Validation dataset	*In vitro* PK abstracts (internal dataset)	DDI abstracts (external dataset)
Single-word drug metabolite type I	89.7%/98.4%	88.0%/92.6%
Multiword drug metabolite type I	95.5%/98.2%	84.2%/81.6%
Multiword drug metabolite type II	77.3%/97.1%	81.6%/79.5%
Overall performance	90.1%/98.1%	85.4%/86.1%

DDI, drug-drug interaction; FN, false negative; FP, false positive; PK, pharmacokinetic; SMO, sequential minimal optimization; TP, true positive.

can enable the normalization of DMs if they are representing in different ways.

Second, abbreviations are frequently used in the biomedical literature to cite drugs and metabolites. If these abbreviations can be integrated in the DM lexicons and the follow-up NER algorithm, it shall have a much better performance in recognizing not only drug names but also their metabolites. For instance, in HMDB, 4-Hydroxymidazolam has 4-OH-MDZ in synonyms. However, 4OH-tamoxifen does not have 4OH-TAM in synonyms. DrugMetab can also enrich abbreviation terminologies in existing databases, such as HMDB.

CONCLUSION

In this article, we propose a new DM NER tool, namely DrugMetab. We make major contributions in developing this innovative NER tool. First, four different DM presentation patterns are defined, and a gold-standard corpus is constructed. This annotated corpus facilitates the next step DrugMetab development. Second, DrugMetab can identify DMs and outperform whatizitChemical. Through our analysis, we discover that Single_Word_Drug_Metabolite type II is still challenging.

Supporting Information. Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website.(www.psp-journal.com).

Data S1. Gold-standard corpus
Data S2. Drug name dictionary
Data S3. Drug metabolism reaction lexicon
Data S4. Scripts for tools

Funding. No funding was received for this work.

Conflict of Interest. As an Associate Editor for *CPT: Pharmacometrics & Systems Pharmacology*, Lang Li was not involved in the review or decision process for this paper. The authors declared no competing interests for this work.

Author Contributions. H.Y.W. and L.L. wrote the manuscript. H.Y.W. and L.L. designed the research. H.Y.W. and L.L. performed the research. H.Y.W., D.L., M.H., S.Z., S.K.Q., and D.Z. analyzed the data.

1. Rowland, M., Tozer, T.N. & Rowland, M. *Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications* (Wolters Kluwer Health/Lippincott William & Wilkins, Philadelphia PA, 2011).
2. Obach, R.S. Pharmacologically active drug metabolites: impact on drug discovery and pharmacotherapy. *Pharmacol. Rev.* **65**, 578–640 (2013).
3. Isoherranen, N., Kunze, K.L., Allen, K.E., Nelson, W.L. & Thummel, K.E. Role of itraconazole metabolites in CYP3A4 inhibition. *Drug Metab. Dispos.* **32**, 1121–1131 (2004).
4. Stearns, V. *et al.* Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. *J. Natl. Cancer Inst.* **95**, 1758–1764 (2003).
5. Wishart, D.S. *et al.* HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res.* **41**, D801–D807 (2013).
6. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091–D1097 (2014).
7. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344–D350 (2008).
8. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**, 3191–3192 (2005).
9. Leaman, R. & Gonzalez, G. BANNER: an executable survey of advances in biomedical named entity recognition. *Pac. Symp. Biocomput.*, 652–663 (2008).
10. David, C., Sérgio, M. & José Luís, O. *Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools* (InTechOpen, London, UK, 2012).
11. Eltyeb, S. & Salim, N. Chemical named entities recognition: a review on approaches and applications. *J. Cheminform.* **6**, 17 (2014).
12. Rocktaschel, T., Weidlich, M. & Leser, U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* **28**, 1633–1640 (2012).
13. Usie, A., Alves, R., Solsona, F., Vazquez, M. & Valencia, A. CheNER: chemical named entity recognizer. *Bioinformatics* **30**, 1039–1040 (2014).
14. Segura-Bedmar, I., Martinez, P., Fau-Segura-Bedmar, M. & Segura-Bedmar, M. Drug name recognition and classification in biomedical texts: a case study. *Drug Discov. Today* **13**, 816–823 (2008).
15. McDonald, R. & Pereira, F. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinform.* **6**, S6 (2005).
16. Alias-i. LingPipe 4.1.0 <http://alias-i.com/lingpipe> (2008).
17. Nobata, C. *et al.* Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics* **7**, 94–101 (2011).
18. Zhou, G., Zhang, J., Su, J., Shen, D. & Tan, C. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics* **20**, 1178–1190 (2004).
19. Nadeau, D. & Sekine, S. A survey of named entity recognition and classification. *Lingvist. Investigat.* **30**, 3–26 (2007).
20. Neves, M. & Leser, U. A survey on annotation tools for the biomedical literature. *Brief. Bioinform.* **15**, 327–340 (2014).
21. Vazquez, M., Krallinger, M., Leitner, F. & Valencia, A. Text mining for drugs and chemical compounds: methods, tools and applications. *Mol. Inform.* **30**, 506–519 (2011).
22. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. & Jimeno, A. Text processing through Web services: calling Whatizit. *Bioinformatics* **24**, 296–298 (2008).
23. Fukuda, K., Tamura, A., Tsunoda, T. & Takagi, T. Toward information extraction: identifying protein names from biological papers. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, 707–718 (1998).
24. Krauthammer, M., Rzhetsky, A., Morozov, P. & Friedman, C. Using BLAST for identifying gene and protein names in journal articles. *Gene* **259**, 245–252 (2000).
25. Björne, J., Kaewphan, S. & Salakoski, T., eds. UTurku: Drug Named Entity Recognition and Drug-Drug Interaction Extraction Using SVM Classification and Domain Knowledge. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013); 2013: Association for Computational Linguistics.
26. Tsuruoka, Y. & Tsujii, J. Improving the performance of dictionary-based approaches in protein name recognition. *J. Biomed. Inform.* **37**, 461–470 (2004).
27. Wu, H.Y. *et al.* An integrated pharmacokinetics ontology and corpus for text mining. *BMC Bioinform.* **14**, 35 (2013).
28. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. & Tsujii, J., eds. BRAT: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics; Association for Computational Linguistics (2012).
29. Kim, J.-D., Ohta, T., Tateisi, Y. & Tsujii, J. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* **19**(suppl. 1), i180–i182 (2003).
30. Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **39**, D1035–D1041 (2011).
31. MeSH.< http://www.nlm.nih.gov/mesh/meshhome.html>
32. Knollmann, B.C. & Randa, H.-D. *Goodman and Gilman's the Pharmacological Basis of Therapeutics* (McGraw-Hill Professional, New York, NY, 2011).
33. Golan, D. *Principles of Pharmacology: The Pathophysiologic Basis of Drug Therapy* (Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA, 2012).

34. Albright, D. *et al.* Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J. Am. Med. Inform. Assoc.* **20**, 922–930 (2013).
35. Marcus, M.P., Marcinkiewicz, M.A. & Santorini, B. Building a large annotated corpus of English: the penn treebank. *Comput. Linguist.* **19**, 313–330 (1993).
36. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).
37. Quinlan, J.R. *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers Inc., Burlington, MA, 1993).
38. Landwehr, N., Hall, M. & Frank, E. Logistic model trees. *Mach. Learn.* **59**, 161–205 (2005).
39. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10–18 (2009).
40. Wu, H.-Y., Zhang, S., Desta, Z., Quinney, S. & Li, L., eds. Translational drug interaction evidence gap discovery using text mining. *Clinical Pharmacology & Therapeutics* (Wiley-Blackwell, Hoboken, NJ, 2017).

© 2018 The Authors *CPT: Pharmacometrics & Systems Pharmacology* published by Wiley Periodicals, Inc. on behalf of the American Society for Clinical Pharmacology and Therapeutics. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.