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.

A drug’s pharmacokinetics (PKs) involves not only the parent compound, but also its metabolites. For instances, codeine drugs have active metabolites (morphine) that possess more therapeutic activity against the targeted protein than its parent drug. 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 hydroxyitraconazole. 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. 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...
There are many NERs that annotate text with biomedical terminologies. Some were designed to identify general terms, like proteins, DNA, RNA, cells, cell lines, etc., and some can annotate drugs, chemicals, or metabolome. 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.

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. 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).

The guidelines were 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. Three annotators with different training backgrounds,
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. 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 and the medical subject heading term. 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 and our previous work. 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 was implemented for creating PoS features for entities. With the Penn Treebank tag set, 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
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 $2n + 1$ (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 $n = 2, 3, 4, 5, \text{and } 6$. 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.
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
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: 10460803, “dextromethorphan o-demethylation” is a metabolite of dextromethorphan. Using whatizitChemical, 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 0.926/88%, (81.6%/84.2%), and (79.5%/81.6%; Table 3), respectively.

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. 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 EBI Med 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.
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: 77.2%/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 artelinc 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,4 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 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

| Type of Drug Metabolite | Recall = TP/ (TP + FN) | Precision = TP/ (TP + FP) | Reasons of Errors |
|-------------------------|------------------------|---------------------------|-------------------|
| Single word drug metabolite type I | 92.7%/91.3% | 98.5% | - Drug name is not in dictionary |
| Single word drug metabolite type II | 32.3%/87.5% | 15.4% | - Error from machine learning |
| Multiword drug metabolite type I | 93.8%/96.3% | 70.5% | - Unidentified drug abbreviations |
| Multiword drug metabolite type II | 77.3%/89.5% | 65.1% | - Unique drug metabolite names |

FP, false positive; PK, pharmacokinetic; SMO, sequential minimal optimization; TP, true positive.
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 = TP/TP+FN)/precision = 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-Hydroxytamoxifin 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 whatzitChemical. 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).
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.