An Automated Approach to Calculating the Daily Dose of Tacrolimus in Electronic Health Records

Hua Xu[1], PhD, Son Doan[1], PhD, Kelly A. Birdwell[2], MD, James D. Cowan[5], Andrew J. Vincz[2], David W. Haas[2,3], MD, Melissa A. Basford[4], MBA, Joshua C. Denny[1,2], MD, MS

[1]Department of Biomedical Informatics; [2]Department of Medicine; [3]Department of Microbiology and Immunology; [4]Vanderbilt Institute for Clinical and Translational Research, School of Medicine, Vanderbilt University, Nashville, TN, USA

Abstract*: Clinical research often requires extracting detailed drug information, such as medication names and dosages, from Electronic Health Records (EHR). Since medication information is often recorded as both structured and unstructured formats in the EHR, extracting all the relevant drug mentions and determining the daily dose of a medication for a selected patient at a given date can be a challenging and time-consuming task. In this paper, we present an automated approach using natural language processing to calculate daily doses of medications mentioned in clinical text, using tacrolimus as a test case. We evaluated this method using data sets from four different types of unstructured clinical data. Our results showed that the system achieved precisions of 0.90-1.00 and recalls of 0.81-1.00.*

INTRODUCTION

Clinical research projects, such as pharmacogenomics and drug-effect studies, often involve extraction of detailed drug information from electronic health records (EHRs). Medication information is typically recorded in hybrids of structured and unstructured formats in the EHR. Structured medication information can be obtained from systems such as provider order entry systems, pharmacy records, and drug administration databases. Unstructured medication mentions are often seen in clinical notes, clinical messaging, personal health records, and, sometimes, problem lists. For example, in an outpatient clinic visit note, physicians typically document a list of the patient’s current medications, and new medications may be mentioned in the “assessment and plan” section of the note. To obtain a complete medication profile of a patient, medication data need to be extracted from both the structured and unstructured sources of EHR. This is usually done via manual chart review by domain experts, which is both time and cost inefficient.

In this paper, we describe a preliminary study using natural language processing (NLP) to automatically extract and calculate daily dose of the drug tacrolimus, a common anti-rejection drug given to transplant patients. Tacrolimus dosages are tightly regulated based on plasma drug levels to prevent transplant rejection.

BACKGROUND

As a large amount of drug information is recorded in clinical narratives, automated methods are needed to extract structured drug information from clinical text. A number of studies have focused on extracting drug names from clinical notes using NLP technologies [1-3]. More recently, a few studies have investigated extraction of more detailed drug information from clinical text, including strength, route, and frequency. A recent study by Gold et al. [4] reported a regular expression based approach for extracting drug names and signature information. Jagannathan et al. assessed four commercial NLP engines for their ability to extract medication information (including drug names, strength, route, and frequency) and they reported a high F-measure of 93.2% on capturing drug names, but lower F-measures of 85.3%, 80.3%, and 48.3% on retrieving strength, route, and frequency, respectively [5]. At Vanderbilt University Medical Center (VUMC), we have developed a medication extraction system called MedEx [6]. MedEx extracts drug names as well as signature information about drug administration. An evaluation of MedEx using data sets from discharge summaries and clinical visit notes at VUMC showed that it could extract drug names, strength, route, and frequency with F-measures over 90%. The MedEx system consists of a semantic tagger and a context free grammar parser that parses textual sentences into structured forms based on pre-defined semantic patterns. Figure 1 shows the components of MedEx, as well as an example of input and output of MedEx.

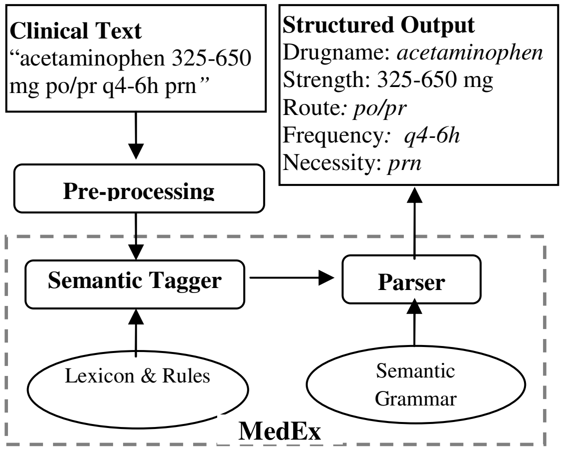

Figure 1. An overview of the MedEx system.

Recently, investigators in clinical research have used NLP-based medication extraction systems to automatically populate detailed drug information from EHR. For example, Peissig et al. [7] applied an NLP-based system [3] to construct atorvastatin dose-response relationships using data from the Marshfield Clinic Personalized Medicine Research Project (PMRP) database. Similar to that study, many other clinical researches also require detailed drug exposure information of patients, including accurate dosage information.

At VUMC, a de-identified copy of the EHR database, called Synthetic Derivative (SD), has been coupled with a large DNA biobank (with about 70,000 samples in November 2009), and has been used for many clinical research projects [8]. One of the ongoing projects is a pharmacogenomics study of the drug tacrolimus. Tacrolimus (TAC, Prograf®, FK506) is the most widely prescribed anti-rejection agent for solid organ and tissue transplants. Its use, however, is complicated by inter-individual pharmacokinetic variability that requires therapeutic drug monitoring to avoid adverse renal, neurologic, and metabolic events. The phenotype of primary interest is the *C/D value*, defined as the ratio of peripheral blood TAC concentration (ng/mL) to total daily TAC dosage (mg/day). Such information is useful for pharmacokinetic and pharmacogenomic studies.

The immediate goal of this study is to facilitate the daily dose extraction of a tacrolimus pharmacogenomics project by using the existing MedEx tool. However, MedEx (similar to many medication extraction systems) extracts dose-related information such as strength and frequency as textual strings, which are not normalized and cannot be directly used to calculate the daily dose. In this study, we extended MedEx by developing a method to normalize dose-related findings and to automatically calculate daily doses of medication mentions in clinical text. We then applied this method to the tacrolimus project and evaluated its performance on the task of determining daily doses of tacrolimus.

METHODS

Data Sets

As mentioned above, we used clinical data from the SD database, which is a de-identified copy of VUMC's EHR that contains detailed longitudinal health records of more than 1.7 million patients. Clinical notes in SD were de-identified using DE-ID®, a commercially-available software package from University of Pittsburgh Medical Center, combined with custom pre- and post-processing algorithms. For this study, 599 subjects were identified based on search criteria of 1) documented renal transplant; 2) multiple (>=3) measures of tacrolimus drug levels within a 30-day time span; 3) at least one tacrolimus-dose mention; 4) DNA sample availability (for future pharmacogenomic analysis). All the clinical records of those subjects were pulled from the SD and processed by MedEx to extract all the mentions of tacrolimus. Based on their sources according to the SD database, those clinical data were divided into four different types:

Clinical documentation (CD): This source contains various types of clinical notes, including admission notes, outpatient clinic visit notes, clinical communication messages, and more. Drug information is usually recorded as narrative sentences in these documents.

Discharge Summaries (DS): This source contains all the discharge or death summaries at VUMC, either dictated or typed by physicians. Drug information in DS can be either in a list or in a narrative sentence.

Problem List (PL): Problem lists contain free-text lists of key problems, medications, family and social history for patients, which are updated by physicians and staffs. Medications are typed into the system, each medication typically separated by newline characters, such as "PROGRAF 2MG PO bid."

WizOrder (WIZ): It is a database of physician orders at VUMC. In WIZ, the vast majority of drug information is in a native structured format. But some information such as frequency is stored as text strings in the database. Despite this internal structure, medications from Wiz are stored in the SD in a free text format not ready for direct daily dose calculation. In this study, we parsed the textual field in WIZ, which concatenates all pieces of drug information into one long string, such as "TACROLIMUS (FK506): 3. MG PO Q12H ALT".

All the sentences that mentioned tacrolimus were identified by MedEx and we obtained following numbers of sentences from each type of clinical notes: 31,746(CD), 2,335(DS), 3,960(PL), and 2,539(WIZ). For each data source, 200 sentences were randomly selected as a test data set, and the rest were used as a training data set. We developed the dose normalization/calculation program using the training set and evaluated its performance on the test set.

Extraction of Dose-related Findings

In a previous study, we used MedEx to extract structured drug findings including dose related information from discharge summaries and clinic visit notes [6]. In this study, we extended MedEx to other types of clinical data such as PL and WIZ. Some unexpected errors were noticed in training sets from those data sources. For example, the original version of MedEx did not extract the dose "3. MG" from the sentence "TACROLIMUS (FK506): 3. MG PO Q12H ALT", because it required another digit after the decimal point (e.g. "3.0 MG"). We modified MedEx to fix those errors.

For the tacrolimus project, we extracted four types of dose-related findings using MedEx: Strength, DoseAmount, Dose, and Frequency. "Strength" refers to the active dose contained in a single drug unit (e.g. a tablet); "DoseAmount" refers to the quantity of drug units for each drug intake (e.g., 2 tablets); "Dose" refers to the total dose needed for each drug intake; and "Frequency" refers to the how often one takes the drug. Based on their definitions, "Dose" is equivalent to the multiplication of "Strength" and "DoseAmount". A medication reference often contains either "Strength" + "DoseAmount" or "Dose" information. Table 1 shows two example sentences and their MedEx parse.

Sentences	1."prograf 1mg 5 tabs twice daily"	2."TACROLIMUS 7MG BID"
Strength	1mg	
DoseAmount	5 tabs	
Dose		7MG
Frequency	twice daily	BID

Table 1. Examples of dose-related findings.

Daily Dose Calculation

It is easy for a human to calculate daily doses using the above four types of dose-related findings. If a "Dose" is available, daily dose can be calculated from the "Dose" and the "Frequency" findings (e.g., example 2 in Table 1). If a "Dose" is not available, it can be calculated based on available "Strength" and "DoseAmount" information. For example 1 in Table 1, "Dose" of "5mg" can be derived based on the information of "Strength-1mg" and "DoseAmount-5 tabs". Then the "Daily Dose" can be obtained as "10mg" by multiplying by the frequency ("twice daily").

However, it is not straightforward for a computer to calculate daily doses based on above information. Those dose-related findings extracted by MedEx are still text strings, which have to be normalized to a numeric representation. We defined normalized representations for dose-related findings, as shown in Table 2. For Strength, DoseAmount, and Dose, they are all represented by a quantity part (Qty) and a unit part (UNIT). A simple regular expression was used to break those types of findings into the normalized Qty + UNIT form.

Finding	**Example**	**Normalization**
Strength	"1mg"	Qty:1, Unit:mg
DoseAmount	"5 tabs"	Qty:5, Unit:tablet
Dose	"7MG"	Qty:7, Unit:mg
Frequency	"twice daily" "q 12h"	Freq:2, Qty:1, Unit:day Freq:1, Qty:12, Unit:hour

Table 2. Examples of normalization of dose-related

Expressions of frequency information are diversified. Tacrolimus is usually given with a frequency of "2 times a day", which, we found, can be expressed in more than thirty different ways, including "b.i.d", "q 12h", "q. 12hours", "2 times a day", "2 times per day", "twice daily", "every 12 hours", etc. Frequency is usually represented as the number of times over a time period, as in TimeML[9]. In this application, we normalized frequency information into three parts: Freq (how many times), and Qty and Unit for the time period. For example, "q 12h" is normalized to "Freq:1, Qty:12, Unit:hour". With this representation, it is easy to convert a frequency from one Unit to another. For example, we can convert "Freq:1, Qty:12, Unit:hour" to its equivalent form of "Freq:2, Qty:1, Unit:day", based on the fact that 1 day has 24 hours.

MedEx recognizes frequency phrases through two ways: lexicon lookup and regular expression matching [6]. For frequency terms in the lexicon file (e.g. "b.i.d"), we manually added their normalized forms (e.g., "Freq:2, Qty:1, Unit:day") into the lexicon file. For a frequency term recognized through regular expressions (e.g., "every 12 hours"), first we break it into three parts (Freq, Qty, and Unit – e.g., "every", "12", "hours"). Then a set of rules were used to map each part to its target value. For example, "every" will be mapped to "Freq:1". After all dose-related findings are normalized to above representations; a function was built to calculate the daily dose of tacrolimus.

Evaluation

The performance of MedEx on extracting drug names has been evaluated in a previous study [6]. In this study, we focused on the evaluation of extracting dose-related findings and calculating daily doses of tacrolimus. Two hundred sentences containing tacrolimus were randomly selected from each of the four data sources (CD, DS, PL, and WIZ) as a test set for evaluation. If a sentence did not contain any dose-related information, it was excluded from the test set. The test set was processed by MedEx and the daily dose calculator. The output contained four dose-related findings: Strength, DoseAmount, Dose, and Frequency, and the calculated daily dose. A person, who is familiar with medication data, manually reviewed the sentences, as well as the system's output, to determine whether a finding was: 1) extracted and correct (TP); or 2) extracted but wrong (FP); or 3) should be extracted but missed (FN). When there was a question, a physician was consulted to make the final judgment. Precision (TP/TP+FP) and recall (TP/TP+FN) were calculated.

RESULTS

After removing the sentences that do not containing any dose-related information from the 200 randomly selected sentences (see Data Sets), we obtained 89, 109, 187, and 200 tacrolimus findings with dosing information for CD, DS, PL, and WIZ respectively, as the final test sets. In Table 3, precisions (Pre) and recalls (Rec) are reported for four types of extracted findings: Strength, DoseAmount, Dose, and Frequency, and the calculated finding of Daily Dose. For four types of dose-related finings, precisions were all above 0.90 for all four types of clinical data; recalls varied from 0.71 to 1.00, depending on the types of clinical data and the types of findings. For Daily Dose, precisions were high (over 0.90) for all four types of clinical data, but recalls were relative low (over 0.80) for CD and DS.

Clinical Notes Type		Strength	Dose Amount	Dose	Freq	Daily Dose
CD	Pre	0.96	0.96	0.93	0.95	0.94
	Rec	0.71	0.74	0.82	0.84	0.81
DS	Pre	0.92	0.92	0.90	0.93	0.90
	Rec	1.00	1.00	0.84	0.98	0.86
PL	Pre	1.00	0.95	0.97	0.98	0.95
	Rec	0.92	0.92	0.90	0.97	0.92
WIZ	Pre	N/A	N/A	1.00	1.00	1.00
	Rec	N/A	N/A	1.00	1.00	1.00

Table 3. Precisions and recalls of extracting dose-related findings and determining daily doses of tacrolimus using four different types of clinical text.

DISCUSSION

We describe an automated approach to calculating daily doses of medications mentioned in clinical text, and evaluated its performance over four types of clinical documentation. This evaluation showed that the extended MedEx system can extract dose-related findings and determine daily doses of tacrolimus mentioned in four different types of clinical data with precisions of 0.90-1.00 and recalls of 0.81-1.00. This demonstration project suggests that such a system may assist in clinical research studies that require structured drug dosage information.

The performance of MedEx on extracting dose-related findings of tacrolimus was similar to previous reported results [6] (with F-measure over 0.9) in the DS data set. As expected, both precisions and recalls were higher for more structured data such as WIZ and PL than for more narrative data such as CD and DS. As CD is a mixture of many different types of clinical notes, low recalls were obtained because of unexpected patterns. We analyzed the errors in the system's output with the purpose of further improvement. For example, some XML documents with drug information caused additional errors in CD. Some errors were caused by missing lexical terms, which are easily remedied. Another type of error resulted from sentences that contained multiple sets of dosing information. Figure 2 shows two examples of those sentences: the first one mentions a dose change, and the second one has two different doses for morning and afternoon. The current MedEx system outputted two records for those sentences: for example 1, it generated “FK-506 1mg b.i.d.” and “FK-506 0.5mg b.i.d.”; for example 2, it generated “FK506 4mg qam” and “FK506 3mg qpm”. It did not demarcate whether to pick one set of dosing information (example 1) or to combine two sets of dosing information (example 2). In the future, we will develop methods that can make the decision based on the context to handle those complicated cases.

Example 1: “FK-506 was redosed from 1 mg b.i.d. to 0.5 mg b.i.d.”
Example 2: “FK506 4mg po qam and 3 mg po qpm”

Figure 2. Examples of sentences with multiple sets of dosing information.

Despite the good performance of calculating the daily doses of tacrolimus from clinical text, it is only the first step toward automatic determination of daily doses for corresponding drug levels. As we noticed, there are often multiple, sometimes discrepant, mentions of tacrolimus dosages occurring on the same date, usually from different sources. Determining the correct daily dose from many drug mentions is another challenge. Possible solutions include prioritizing different types of clinical notes based on their reliability and building models to determine correct dose based on the information across a timeline.

This study has several limitations. First, it focused on daily dose extraction for only one drug. We are aware that some other drugs could have more complicated patterns about dosing information. For example, warfarin can have multiple different doses at different days within a week. However, the normalization model for dose-related information such as frequency is generalizable. We plan to apply it to more complicated drugs in future studies. Second, the gold standard in this study was generated by manual review of the system’s output, which may cause bias. However, the annotation task is relatively simple, such that we expect the quality of gold standard is high.

CONCLUSION

Capture and interpretation of detailed drug information is crucial to many type of EHR-based research, such as pharmacogenomics. Manually extracting and calculating drug information such as daily doses from clinical narratives is costly and time-consuming. We developed an approach to normalize dose-related findings that are extracted by an existing medication extraction system, thus enabling the automated calculation of daily doses of drugs. Evaluation for the drug tacrolimus showed that the system could calculate the daily dose of tacrolimus with high precision and acceptable recall from different types of clinical data.

ACKNOWLEDGEMENT

This study was supported by grants U01 HG004603 (NHGRI) and R01CA141307 (NCI). The datasets used were obtained from VUMC’s Synthetic Derivative which is supported by institutional funding and by the Vanderbilt CTSA grant 1UL1RR024975-01 from NCRR/NIH.

REFERENCES:

1. Chhieng D, Day T, et al. Use of natural language programming to extract medication from unstructured electronic medical records. AMIA. 2007, 908

2. Levin MA, Krol M, Doshi AM, Reich DL. Extraction and mapping of drug names from free text to a standardized nomenclature. AMIA. 2007, 438-42

3. Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. Pac.Symp.Biocomput. 2005;308-18

4. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. AMIA.2008, 237-41

5. Jagannathan V, Mullett CJ, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int.J.Med.Inform. 2008 Oct 4;

6. Xu H, Stenner S, Doan S, Johnson K, Waitman L, Denny JC, "MedEx – A Medication Information Exaction System for Clinical Narratives", JAMIA 2009 (accepted).

7. Peissig P, Sirohi E, Berg RL, Brown-Switzer C, Ghebranious N, McCarty CA, Wilke RA. Construction of Atorvastatin Dose-Response Relationships Using Data from a Large Population-Based DNA Biobank. Basic Clin Pharmacol Toxicol 2007; 100, 286-8.

8. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008 Sep;84(3):362-9.

9. Pustejovsky J, Castaño J, Ingria R, Saurí R, Gaizauskas R, Setzer A and Katz G. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. IWCS-5, Fifth International Workshop on Computational Semantics.