MedXN: an open source medication extraction and normalization tool for clinical text

Sunghwan Sohn,1 Cheryl Clark,2 Scott R Halgrim,3 Sean P Murphy,1 Christopher G Chute,1 Hongfang Liu1

ABSTRACT

Objective We developed the Medication Extraction and Normalization (MedXN) system to extract comprehensive medication information and normalize it to the most appropriate RxNorm concept unique identifier (RxCUI) as specifically as possible.

Methods Medication descriptions in clinical notes were decomposed into medication name and attributes, which were separately extracted using RxNorm dictionary lookup and regular expression. Then, each medication name and its attributes were combined together according to RxNorm convention to find the most appropriate RxNorm representation. To do this, we employed serialized hierarchical steps implemented in Apache’s Unstructured Information Management Architecture. We also performed synonym expansion, removed false medications, and employed inference rules to improve the medication extraction and normalization performance.

Results An evaluation on test data of 397 medication mentions showed F-measures of 0.975 for medication name and over 0.90 for most attributes. The RxCUI assignment produced F-measures of 0.932 for medication name and 0.864 for full medication information. Most false negative RxCUI assignments in full medication information are due to human assumption of missing attributes and medication names in the gold standard.

Conclusions The MedXN system (http://sourceforge.net/projects/ohnlp/files/MedXN/) was able to extract comprehensive medication information with high accuracy and demonstrated good normalization capability to RxCUI as long as explicit evidence existed. More sophisticated inference rules might result in further improvements to specific RxCUI assignments for incomplete medication descriptions.

INTRODUCTION

Patients’ medication history stored in electronic medical records (EMRs) provides critical information for medical treatment, patient safety, and secondary use of EMRs. Although some medication information can be extracted from the structured data, a substantial amount of the medication information resides in clinical narratives, which can be extracted only with advanced informatics techniques. In EMRs, medication information is described using different vocabularies and patterns, causing medication errors within or across institutions, hindering effective medication management and jeopardizing patient safety.

Medication errors are the most common patient safety error.1 The majority of medication errors are due to inaccurate reconciliation during the admission, transfer, and discharge of patients.2 One fifth of these errors are believed to cause adverse effects to the patient.3 To improve patient safety, it is crucial to accurately reconcile patient medication history across the continuum of care (Joint Commission’s National Patient Safety Goals, http://www.jointcommission.org). Medication reconciliation is defined as ‘the process of identifying the most accurate and complete list of the patient’s current medications, including name, dosage, frequency, and route, by comparing the medical records’.5

The accurate exchange of medication information requires drug name standards that can be mapped to drug name variations from different sources.6 RxNorm7 addresses this issue by providing a normalized drug name (RxNorm name) that links to drug name variants from many different sources and therefore allows drug management systems to reconcile information. Today, RxNorm is becoming part of Meaningful Use to support the expanding functionality of health record technology. A system that extracts medication information from clinical notes and maps it to its corresponding RxNorm representation will facilitate medication reconciliation and enable better medication management.

This paper describes the Medication Extraction and Normalization (MedXN pronounced [med-eks-n]) system that we developed to extract comprehensive medication information from clinical notes and normalize it to the most appropriate RxNorm concept unique identifier (RxCUI). MedXN focuses on medication normalization by mapping the comprehensive medication description to the best matching RxCUI. The MedXN system was implemented using Apache’s Unstructured Information Management Architecture (UIMA),9 which provides an efficient way to manage the system architecture.

BACKGROUND

Clinical text normalization

The translation of raw data extracted from clinical text to a normalized form plays a vital role in their use in searches within the EMRs, as well as secondary use of EMRs. After data are normalized, they can be shared with other researchers to promote collaboration and enhance clinical discovery. Medical text normalization has been carried out by using the Unified Medical Language System (UMLS),10 MeSH terms,11 and SNOMED-CT.12 For medication exchange, RxNorm7 provides normalized drug names linking to drug variants. In the biology domain, BioCreative is a community-wide effort for gene normalization.13,14 The BioCreative II.5 challenge posed the task of gene normalization to identify genes that play an interactor role in...
protein–protein interaction and map them to unique IDs. Many researchers used dictionary matching to identify proteins and their synonyms,\textsuperscript{15–18} and machine learning approaches\textsuperscript{15 17} or context information\textsuperscript{16 18} to assign the normalized gene mentions.

Natural language processing in clinical applications

Over the past decade EMRs have grown rapidly and a large amount of clinical data is stored in free-text format. Natural language processing (NLP) techniques can convert unstructured text to a structured format, and multiple clinical NLP systems have been developed, such as MedLEE,\textsuperscript{19} cTAKES,\textsuperscript{20} MedTagger,\textsuperscript{21} YTEX,\textsuperscript{22} MTERMS,\textsuperscript{23} and HITEX.\textsuperscript{24} Clinical NLP has been successfully employed in various clinical applications including pathology information extraction,\textsuperscript{25} patient medical status extraction,\textsuperscript{26 27} sentiment analysis,\textsuperscript{28} decision support,\textsuperscript{29 30} genome-wide association studies,\textsuperscript{31 32} and diagnosis code assignment.\textsuperscript{33 34} Our system uses basic NLP techniques (eg, tokenization, sentence segmentation) from cTAKES\textsuperscript{20} and the dictionary lookup algorithm from MedTagger\textsuperscript{21} (http://sourceforge.net/projects/ohnlp/files/MedTagger/) to process clinical free text.

Medication extraction systems

Early medication extraction systems focus on extracting the medication name itself\textsuperscript{35 36} and mapping it to a standardized nomenclature.\textsuperscript{37} In 2009, the third i2b2 workshop on NLP challenge focused on extracting comprehensive medication information, including medication name, dosage, mode, frequency, and duration from clinical discharge summaries.\textsuperscript{38} Many performing teams used rule-based systems.\textsuperscript{39–45} They used drug vocabularies available from public sources including the UMLS\textsuperscript{10} and the web to identify drug names.\textsuperscript{38} Regular expression pattern matching was also employed to parse medication attributes. Some researchers used a hybrid system, although they still employed drug vocabularies and coupled them with various machine learning approaches.\textsuperscript{38} Patrick and Li\textsuperscript{46} used conditional random field (CRF) to extract medications and attributes, support vector machines to link them, and pattern matching rules to aid the detection of fields. Meystre et al\textsuperscript{47} used knowledge engineering with existing software, OpenNLP (http://opennlp.apache.org/) and MetaMap.\textsuperscript{48} Li et al\textsuperscript{49} applied CRF for tagging medication name and attributes and AdaBoost for associating medication to attributes.

Our system emphasizes normalizing medication information (ie, it maps comprehensive medication information to a specific RxCUI), as well as extracting comprehensive information from clinical notes. Since medication descriptions in clinical notes often vary across institutions and even within note types,\textsuperscript{50} an easy interface for system adjustment would be desirable. MedXN externalizes pattern-matching rules and allows end users to easily customize them according to their needs.

RxNorm

RxNorm is a terminology for clinical drugs produced by the US National Library of Medicine (NLM). An RxNorm clinical drug consists of the ingredient, strength, and dose form and conceptually unique medication descriptions are assigned by a concept unique identifier (RxCUI). RxNorm provides normalized medication names and links them to various medication data frequently used in pharmacy management such as First Databank,\textsuperscript{51} Micromedex,\textsuperscript{52} MediSpan,\textsuperscript{53} Gold Standard,\textsuperscript{54} Multum,\textsuperscript{55} and NDF-RT.\textsuperscript{56} RxNorm enables various systems using different medication terminologies to share and exchange data.

MATERIALS

Medication information

Clinicians describe medication information (ie, a medication name and/or its attributes) in a variety of patterns. Table 1 shows the medication annotations used in our system. ‘normRxCUI’ is the RxCUI of comprehensive medication information, typically clinical drugs, in RxNorm (ie, medication +strength+form). Not all medication descriptions in clinical notes are assigned to normRxCUI because many of them occur as medication name alone. The medication description patterns vary in clinical notes and also vary across different institutions. However, dominant patterns do exist that account for most of the medication descriptions, so therefore, it follows that a system developed on a specific data set would work reasonably well on other data with some modification.\textsuperscript{10}

Data

We used three different data sets in our study: development, validation, and test sets. The development set was composed of 159 clinical notes randomly selected from Mayo Clinic and contained 639 manually annotated medication mentions and their attributes. This set was used to examine a variety of medication attribute patterns and develop the initial MedXN system. For medication reconciliation and clinical study, current medications that are currently being taken by patients are the most important. The ‘Current Medication’ section in clinical notes contains those medications and, in fact, covers the majority of medication mentions. This section contains medication descriptions arranged like a grocery list (ie, one medication description per line), while the other sections generally contain medication descriptions in narrative (ie, unstructured).\textsuperscript{50} Since our development set did not have many Current Medication sections (only 28% of notes contain a Current Medication section), we randomly selected an additional 51 clinical notes from Mayo Clinic that included the Current Medication section and annotated complete medication information, as well as the RxCUIs of medication name alone and specific medication description. Two medical experts annotated the notes and the inter-annotator agreement is described in appendix C in the online supplement. These notes were divided into validation (25 notes with 386 medications) and test sets (26 notes with 397 medications). The validation set was used to evaluate the initial

<table>
<thead>
<tr>
<th>Table 1 Medication annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>Medication</td>
</tr>
<tr>
<td>Dosage</td>
</tr>
<tr>
<td>Strength</td>
</tr>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td>Route</td>
</tr>
<tr>
<td>Duration</td>
</tr>
<tr>
<td>Form</td>
</tr>
<tr>
<td>DrugRxCUI</td>
</tr>
<tr>
<td>NormRxCUI</td>
</tr>
</tbody>
</table>
MedXN system and further refine the system to improve performance. The test set was used to test the final MedXN system.

**METHODS**

To achieve effective medication information extraction and normalization, MedXN employs a strategy of decomposition (ie, it extracts medication and attributes separately from text) and composition (ie, it associates medication and the corresponding attributes together). After this step, the MedXN system converts this medication information to the RxNorm standard and maps it to the corresponding RxCUI. Figure 1 shows the high-level procedures of MedXN with examples. Each step is explained below (refer to appendix A in the online supplement for additional descriptions and examples).

**Step 1. Medication extraction**
This step extracts the medication name. A dictionary lookup, based on the AhoCorasic algorithm57 (a fast-string matching algorithm), was used. A medication dictionary consists of ingredient (IN), brand name (BN), precise ingredient (PIN), and multiple ingredient (MIN) from RxNorm. We also included any other medication variations that have the same RxCUI as the above medications in order to include various medication synonyms.

**Step 2. Medication attributes extraction**
This step extracts medication attributes (ie, dosage, strength, frequency, form, and route) using regular expressions. We examined patterns of attribute descriptions and manually implemented regular expressions to parse them. Those regular expressions were placed into an externalized file.

**Step 3. Medication and attributes association**
This step links a medication name with corresponding attributes, if they appear within a given text window. A text window is defined as the text from the beginning of a given medication to the smaller offset of either the beginning of the next medication or the end of the next two sentences. However, there are a few exceptions: (1) if a brand name is followed by an ingredient within parentheses or square brackets or vice versa (eg, ‘Loratadine [CLARITIN] 10-mg oral tablet’), these two medications are treated as one to allow the system to catch all following attributes; (2) if time- or volume-related information precedes the medication with specific dose forms (eg, ‘24 HR Indur 30 MG Extended Release Tablet’ or ‘0.2 ML Somatuline 300 MG/ML Prefilled Syringe’), these are considered as part of the attributes of a given medication; (3) for patterns such as ‘IV+medication’ and ‘of+medication,’ the beginning offset of the window expands to 20 characters before the medication; and (4) Mayo Clinic’s ‘Current Medication’ section often contains medications followed by the subtitles ‘Instructions’ and ‘Indication’ in the next line. Otherwise, medication information is in the same line. Thus, we only expand the text window span to the next lines if we see these two subtitles (this is specific to Mayo Clinic).

**Step 4. Conversion to RxNorm standard**
After medication names and their attributes were linked together, we converted them to the RxNorm standard order (ie, ‘medication+strength+dose form’). If this resulted in both IN and BN, we moved BN to the end (IN+strength+dose form +[BN]).

We also performed additional steps to follow RxNorm conventions, including: (1) following RxNorm attribute formats (remove the hyphen between strength number and unit, add a space between them, etc); (2) using expanded abbreviations; (3) inference of nearby context in order to obtain a specific dose form (eg, Tylenol 325 mg tablet by mouth→Tylenol 325 mg oral tablet); and (4) using the RxNorm preferred medication name.

**Step 5. Conversion to RxCUI representation**
Although most medication descriptions can be captured in the previous step, it is not able to catch some medication variations, especially RxNorm synonyms (SY). For example, ‘white petrolatum 100% topical gel’ is an RxNorm synonym but its representation in step 4, ‘petrolatum, white 100% topical gel’ (IN+strength+dose form), has no match in RxNorm. For this reason, we further generalized string-based medication descriptions to RxCUI representations by converting medication and dose form to RxCUI. Then, we matched it with the RxCUI-represented dictionary. For example,

Medication text: White petrolatum 100% topical gel
Step 4: petrolatum, white <in>100%<strength>topical gel <dose form>
Step 5: 82074<in>100%<strength>346286<dose form>
The string in step 4 does not match either the RxNorm name or synonym. However, the RxCUI representation in step 5 can find the correct specific RxCUI representation of it because both ‘petrolatum, white’ and ‘white petrolatum’ share the same RxCUI.

**Step 6. Normalization to specific RxCUI**
In this step, we matched the RxCUI represented form from the previous step with the dictionary terms that were compiled from RxNorm SCDC, SCDF, SCD, SBDC, SBDF, SB, and SY. Here, medication names and dose forms were represented by RxCUI as explained previously. An AhoCorasic string-matching algorithm was also used. If there was more than one match, we selected the most specific one (ie, the longest match). In the following examples, we select the last one as the specific RxCUI (ie, normRxCUI).

- ‘Sulfasalazine [AZULFIDINE] 500-mg tab 2 tabs by mouth two times a day’
  - SCDC: Sulfasalazine 500 mg (RxCUI=316751)

Figure 1 MedXN algorithm flow for medication extraction and normalization. RxCUI, RxNorm concept unique identifier.
SCD: Sulfasalazine 500 mg oral tablet (RxCUI=198232)

SBD: Sulfasalazine 500 mg oral tablet [AZULFIDINE] (RxCUI=208437)

In addition to the above steps, we also employed extra processes to further refine medication mention extraction. These processes are described below.

**Synonym and abbreviation expansion**

Although RxNorm IN, BN, PIN, and MIN and their variations cover most medication names, they do not cover some medication synonyms and abbreviations (eg, ‘ASA 81 mg oral tablet’ and ‘0.3 mL EPO 10000 unit/mL prefilled syringe’ where ASA and EPO are abbreviations of Aspirin and Epoetin Alfa, respectively). Thus, we compiled the RxNorm medication mentions that do not start with those types mentioned above and manually examined whether they were synonyms or abbreviations. As a result, we generated 302 synonyms/abbreviations and added them to our medication dictionary.

**Removing false positive medications**

Some medication names are often confused with general words, which can cause false positive medication extraction. For example, ‘Today,’ ‘Level,’ and ‘Tomorrow’ are medication names in RxNorm, but also often occur as non-medication names in clinical notes. To avoid this issue, we manually compiled potential false positive medications based on our observations. We applied the following criteria to these terms and did not treat them as medication (1) if they started with a lowercase character, or (2) if there were no corresponding attributes associated with them, such as dosage, strength, frequency, etc.

**System evaluation**

We evaluated medication, individual attribute, and RxCUI assignment using precision, recall, and F-measure defined by:

\[
\text{precision} = \frac{\text{truePositives}}{\text{truePositives} + \text{falsePositives}}
\]

\[
\text{recall} = \frac{\text{truePositives}}{\text{truePositives} + \text{falseNegatives}}
\]

\[
F\text{- measure} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}
\]

Both exact and partial matches were used for evaluation. Exact matches are defined as those cases where the offsets of the system outputs are exact matches to those of the gold standard. Partial matches are the cases where the offsets between the system output and the gold standard are overlapped.

We compared the performance of RxCUI assignment in MedXN to the National Center for Biomedical Computing (NCBO) annotator and MedEx. NCBO provides web services that use ontologies and terminologies to annotate data, and its tools are widely used by biomedical investigators in clinical and translational research. The NCBO annotator has the capability to annotate biomedical free text using certain ontologies selected by users and assigns the corresponding concept identification. In our case, we used RxNorm to compare its performance with MedXN. However, the NCBO annotator generates all possible matches from RxNorm, which includes dose form (eg, tablet, capsule, etc) in RxNorm, and does not have an option to generate drugRxCUI and normRxCUI separately. To fairly compare the NCBO annotator’s performance to MedXN, we removed dose forms and any medication annotation completely subsumed by another in the NCBO output. We merged drugRxCUI and normRxCUI in MedXN by applying criteria whereby, if normRxCUI was found by the system, it was chosen; otherwise, drugRxCUI was selected. In other words, we used RxCUIs as specific as possible found by the system in this comparison. MedEx is a publicly available medication information extraction system that performed the second best in the i2b2 medication information extraction challenge. Originally, MedEx was developed using discharge summaries to extract medication name and attributes. A recently released version also includes RxCUI assignment functionality. In this comparison, we used Medex_UIMA_1.1 (https://code.google.com/p/medex-uima/downloads/list).

In addition, we compared the performance of medication and/or attributes extraction in MedXN to MedEx and MetaMap (2012 release). MetaMap is a widely available program developed by NLM to map biomedical text to UMLS concepts including medication.

**RESULTS**

The MedXN system was built under Apache UIMA and the annotation process is defined in the serialized annotation flow, which allows for the efficient addition of new components. Figure 2 shows a simple example of medication extraction and RxCUI assignment in the MedXN system through the UIMA graphical view. The left window shows annotation types, such as medication name, attributes, and the corresponding RxCUI. The right window contains a simple example of clinical text. Here, the given full medication information was mapped to the RxNorm name ‘Aspirin 81 MG Oral Tablet’ with RxCUI=243670.

An initial MedXN system was developed using the development set (159 clinical notes from Mayo Clinic) and evaluated on the validation set (results shown in table I of appendix B in the online supplement). Then, we further refined the MedXN system based on error analysis on the validation set and performed the final evaluation on the test set (results shown in table 2 and figure 3). After refinement, the system performance was noticeably improved in the test set compared to the validation set.

An exact match F-measure of frequency in the validation set was relatively low. This was because both regular mentions and Latin abbreviations were annotated together in the gold standard if they occurred consecutively (eg, ‘twice a day prn’ or ‘BID as needed’), but the system produced separated annotation (eg, ‘twice a day,’ ‘prn,’ ‘BID,’ or ‘as needed’). This problem was fixed in the refinement stage. Duration had a relatively low number of cases (13 in the validation and 17 in the test sets) and its descriptions varied in clinical notes. As a result, its performance was not high compared to other attributes.

The normRxCUI assignment process assigns the most appropriate RxCUI to the full medication information. However, not all medication mentions are applicable for normRxCUI assignment, as it requires not only medication name but also extra attributes, such as strength and form. The validation set contains 83 normRxCUI cases and the test set contains 126 normRxCUI cases in the gold standard. MedXN produced 70 and 106 normRxCUI assignments in the validation and test sets, respectively.

Many false negative cases in normRxCUI (approximately 45% in the test set) are due to inference by human annotators (ie, missing attributes or string variations are inferred by human annotators). If we ignore them, the F-measure would be 0.912 in the partial match.
MedXN used inference rules to assume dose form (refer to step 4 in the ‘Methods’ section). This approach improved the performance of normRxCUI assignment (0.280 and 0.188 increases in exact and partial F-measures, respectively).

The performance of medication information extraction in MedXN is also compared to MedEx (for both medication and attributes) and MetaMap (for medication) in table 3. The definitions of medication and/or attributes are slightly different among systems and therefore a partial match was used for evaluation.

MedXN consistently produced a higher precision, recall, and F-measure than the other systems except the recall of medication in MetaMap. For medication extraction, MetaMap had the highest coverage (ie, recall), but its precision was lower than that of MedXN and MedEx. We used semantic types of clinical drug, pharmacologic substance, antibiotic, and vitamin with scores greater than or equal to 610, which produced the best performance, for medication extraction in MetaMap. We performed the significance test (bootstrapping and Student t test with p=0.01) on F-measures of medication and individual attribute and it showed that MedXN’s performance of medication and its attribute extraction was significantly different from that of the other systems (see table II in appendix B of the online supplement for details).

The performance comparison of RxCUI assignment in MedXN to the NCBO annotator and MedEx is shown in both table 4 and figure 4. MedXN produced a higher F-measure for RxCUI assignment than both the NCBO annotator and MedEx, and the differences were more significant in precision than in recall.

DISCUSSION

MedXN focuses on extracting comprehensive medication information and normalizing it to the most appropriate RxCUI. Often medication descriptions in clinical notes do not exactly follow the patterns in RxNorm names, which requires flexible matching in normalization to achieve high performance in RxCUI matching. MedXN’s systematic approach, which decomposes the process into multiple simple processes and combines them together in the final step, effectively handles various medication description patterns and maps them to the appropriate RxCUI.

MedXN produced high accuracy in drugRxCUI assignment (ie, extracting medication names and mapping them to RxCUI). Medication name extraction produced false positives because some names in our medication dictionary might also represent...
other than medication (e.g., ‘Adenosine sestamibi scan in January of 2003’ where ‘Adenosine’ is a medication name in RxNorm but is part of the procedure representation in the clinical note). Though MedXN employs a set of rules to filter out general words that can be confused with actual medications, some cases require better context information in order to be disambiguated from the true medication as in the above example. False negatives of medication names were largely due to lack of human-like inference in the system—that is, medications were not exactly the same string as in the actual medication names but human annotators assumed the appropriate one (e.g., ‘B-12’ was annotated as ‘B-12 vitamin’ by human annotators) and also annotated misspelled cases (e.g., ‘Advair Discus’ was identified as ‘Advair Diskus’ by human annotators but not by the system).

MedXN was able to find the most specific medication information beyond medication name and map it to the corresponding RxCUI (i.e., normRxCUI). False positives for normRxCUIs were due to incomplete matching. For example, the system mapped ‘Aspirin 81 mg enteric coated 1 tablet by mouth’ to ‘Aspirin 81 MG Oral Tablet’ (RxCUI=243670) instead of the true case of ‘Aspirin 81 MG Enteric Coated Tablet’ (RxCUI=308416). This is because the system extracted ‘tablet’ as a form instead of ‘enteric coated tablet’ due to the complete form being interrupted by ‘1.’ Another example of a false positive is that the system mapped ‘Lisinopril/hydrochlorothiazide 10/12.5 mg one tablet by mouth’ to ‘Hydrochlorothiazide Oral Tablet’ (RxCUI=370643) instead of the true case of ‘Lisinopril/hydrochlorothiazide 12.5 MG/Lisinopril 10 MG Oral Tablet’ (RxCUI=197785). In normRxCI assignment, a major cause of false negatives was human annotators’ inference for missing attributes or medication names. For example, human annotators mapped ‘Claritin, 10 mg p.o.’ to ‘Claritin 10 MG Oral Tablet’ with RxCUI=206805 by assuming the missing form ‘oral tablet.’ Similarly, human annotators mapped ‘Prednisone 10 mg taking 2 tabs twice a day’ to ‘Prednisone 10 MG Oral Tablet’ with RxCUI=198145 by assuming the missing route ‘oral.’ For some cases, human annotators used brand name and ingredient interchangeably to obtain the normRxCI. For example, ‘Pravastatin 20 mg one tablet by mouth’ was mapped to ‘Pravachol 20 MG Oral Tablet’ with RxCUI=904469. Here ‘Pravachol’ is a brand name and ‘Pravastatin’ is its ingredient.

As described in step 5 in the ‘Methods’ section, MedXN used RxCUI representation to improve normRxCI assignment. However, this approach did not improve the performance noticeably in our test set. Depending on the data set, step 5 could be optional.

MedXN performed better than both the NCBO annotator and MedEx in RxCI assignment. The NCBO annotator does not filter out general words (e.g., Today, Level, Air, etc) or potential laboratory test items (e.g., Cholesterol, Glucose, etc) that can be confused with medications, and therefore produces many false positive medications. The NCBO annotator did not catch medication synonyms that have the same RxCI as the RxNorm name but a different string. For example, ‘HCZ’ is a synonym of the RxNorm name ‘Hydrochlorothiazide’ but was missed in NCBO annotation. The NCBO annotator failed to assign a specific RxCI if the medication description had a different pattern from the RxNorm name, even though it had the same components. For example, from the medication description ‘Sildenafil [VIAGRA] 25 mg tablet 1 tablet by mouth,’ the NCBO annotator generated two individual medications ‘Sildenafil’ and ‘Viagra’ and mapped them to each individual RxCI. Comparatively, MedXN was able to map such descriptions to a correct RxNorm name ‘Sildenafil 25 MG Oral Tablet [Viagra]’ with the correct RxCI.

MedEx allowed relaxed RxCI assignment in some cases. For example, in the case of ‘Pepcid one po’, MedEx assigned it to ‘Pepcid Oral Product’ (RxCUI=1185206), but both MedXN and the NCBO annotator produced ‘Pepcid’ (RxCUI=196458). In another example, ‘Levaquin 250 mg’ was assigned by MedEx to ‘Levofloxacin 250 MG [Levaquin]’ (RxCUI=572148), but both MedXN and the NCBO annotator produced ‘Levaquin’ (RxCUI=217992). MedXN did not allow this kind of partial match because a complete medication description like ‘Levofoxacin 250 MG [Levaquin]’ actually occurs in clinical notes. There is also ‘Levofloxacin 250 MG’ in RxNorm with a

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Comparison of medication information extraction on the test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>MedXN P</td>
</tr>
<tr>
<td>Medication</td>
<td>0.982</td>
</tr>
<tr>
<td>Dosage</td>
<td>0.910</td>
</tr>
<tr>
<td>Strength</td>
<td>0.979</td>
</tr>
<tr>
<td>Form</td>
<td>0.944</td>
</tr>
<tr>
<td>Route</td>
<td>0.990</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.954</td>
</tr>
<tr>
<td>Duration</td>
<td>0.714</td>
</tr>
</tbody>
</table>

Cells with ‘–’ denote unsupported functionality for a given system. In MedEx ‘form’ is extracted as ‘dosage’ if they occur together. F, F-measure; P, precision; R, recall.

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Comparison of RxCI assignment on the test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>MedXN Precision</td>
</tr>
<tr>
<td>Medicine</td>
<td>0.877</td>
</tr>
</tbody>
</table>

Overlapped drug span and RxCI match, and false negative does not include cases that also occur in false positive.
NCBO, National Center for Biomedical Computing.

Figure 4 MedXN versus National Center for Biomedical Computing (NCBO) annotator versus MedEx performance of RxCI assignment on the test set.
different RxCUI. These results contributed to relatively lower precision in MedEx’s RxCUI assignment.

Other MedEx errors in RxCUI assignment are similar to those of the NCBO annotator. MedEx seemed to filter out potentially incorrect medications but still produced some false positive medications (eg, ‘Colon,’ ‘triglycerides,’ and ‘dry skin’). Also, for the example ‘Sildenafil [VIAGRA] 25 mg tablet 1 tablet by mouth,’ MedEx generated two individual medications ‘Sildenafil’ and ‘Viagra’ and mapped them to each individual RxCUI. Overall, MedEx produced a lower F-measure in RxCUI assignment than MedXN and the NCBO annotator. However, it should be noted that MedEx was primarily developed for medication information extraction, not for RxCUI assignment.

MedXN also produced higher F-measures of medication and attributes extraction than MedEx and MetaMap. Although MetaMap produced the highest recall for medication extraction, its precision was much lower than the others. This might be due to the lack of a fine-tuning mechanism in MetaMap for medication information extraction.

Medication descriptions in clinical notes vary depending on institutions and even on note types or sections. The current MedXN system was developed on Mayo Clinic clinical notes, particularly on the Current Medication section. It might be necessary for end users to customize the system if they have different patterns of medication description in their clinical notes. MedXN’s externalized regular expression patterns for attribute extraction allow end users to easily customize them according to their needs.

In the future, we plan to improve the inference capability for matching full medication information through: (1) compiling RxNorm medication mentions that have a certain unique attribute, therefore safely assuming missing attributes in the system, even if they do not appear in clinical notes; and (2) examining medication name and frequent co-occurrence attribute statistics within a large set of Mayo Clinic clinical notes, in order to assume missing attributes with certain confidence.

CONCLUSION
In this paper, we presented an open source medication extraction tool, MedXN (http://sourceforge.net/projects/ohnlp/files/MedXN/), which aims to extract comprehensive medication information with high accuracy from clinical notes and normalize the information to the most specific RxCUI with reasonably good performance. The inference rules (ie, ‘inference of nearby context’ in step 4 in the ‘Methods’ section) resulted in a significant improvement in RxCUI assignment; however, additional rules are necessary to mimic the behavior of human assumption in order to further improve mapping capability. The UIMA framework, with its series of simple annotation steps, allows flexibility in full medication information matching and efficiency in the system architecture. In clinical information extraction, a system developed in a certain domain often requires adaptation in order to perform well in different domains. We believe that MedXN’s externalized resources (ie, medication dictionaries and rules for attribute patterns) create a flexible mechanism for the adaptation process.

Acknowledgements We would like to thank Donna Ihrike and Liwei Wang for annotating medication data and James Masanz for setting up the Knowtator framework for gold standard annotation.

Contributors SS and HL conceived and designed the experiments and developed the system. All authors participated in system analysis. SS wrote the first draft of the manuscript. All authors contributed to the writing of the manuscript. All authors reviewed and approved the final manuscript.

Funding This study was supported by the Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002) administered by the Office of the National Coordinator for Health Information Technology, National Institute of Health grants R01LM009959A1, R01GM102282A1, and National Science Foundation grant ABI-0945523. The contents of the manuscript are solely the responsibility of the authors.

Competing interests None.

Ethics approval The Mayo Clinic Institution Review Board approved this study.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES
8 RxNORM. http://www.nlm.nih.gov/research/umls/rxnorm/
9 UIMA. http://incubator.apache.org/uima/
10 UMLS. http://www.nlm.nih.gov/research/umls/
11 MeSH. http://www.nlm.nih.gov/me/sh/
15 Chen Y. Normalizing Interactor Protein and Extracting Interaction Protein Pairs using Support Vector Machines. BioCreative II.5 Workshop 2009; Madrid; Spain; 2009: 29.


Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010;17:524–7.


