Drug side effect extraction from clinical narratives of psychiatry and psychology patients

Sunghwan Sohn,[1] Jean-Pierre A Kocher,[1] Christopher G Chute,[1] Guergana K Savova[2]

► Additional appendices are published online only. To view these files please visit the journal online (www.jamia.org/content/18/Suppl_1.toc).

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
[2]Children's Hospital Boston and Harvard Medical School, Boston, Massachusetts, USA

Correspondence to
Dr Sunghwan Sohn, Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA; sohn.sunghwan@mayo.edu

Received 3 May 2011
Accepted 18 August 2011
Published Online First
21 September 2011

ABSTRACT

Objective To extract physician-asserted drug side effects from electronic medical record clinical narratives.

Materials and methods Pattern matching rules were manually developed through examining keywords and expression patterns of side effects to discover an individual side effect and causative drug relationship. A combination of machine learning (C4.5) using side effect keyword features and pattern matching rules was used to extract sentences that contain side effect and causative drug pairs, enabling the system to discover most side effect occurrences. Our system was implemented as a module within the clinical Text Analysis and Knowledge Extraction System.

Results The system was tested in the domain of psychiatry and psychology. The rule-based system extracting side effects and causative drugs produced an F score of 0.80 (0.55 excluding allergy section). The hybrid system identifying side effect sentences had an F score of 0.75 (0.56 excluding allergy section) but covered more side effect and causative drug pairs than individual side effect extraction.

Discussion The rule-based system was able to identify most side effects expressed by clear indication words. More sophisticated semantic processing is required to handle complex side effect descriptions in the narrative. We demonstrated that our system can be trained to identify sentences with complex side effect descriptions that can be submitted to a human expert for further abstraction.

Conclusion Our system was able to extract most physician-asserted drug side effects. It can be used in either an automated mode for side effect extraction or semi-automated mode to identify side effect sentences that can significantly simplify abstraction by a human expert.

INTRODUCTION

Patients' drug histories and responses to drugs are critical information for future medical treatment. In particular, the detection of drug side effects is tightly linked to patient safety and pharmacovigilance. The early detection of side effects is critical to ensure patient population safety and has led to the implementation of post marketing drug safety surveillance (phase IV of clinical trials). From a preventive point of view, the tracking of a patient's drug intake along with the associated side effects advances further individualized medicine and the identification of genetic markers of drug metabolism and toxicity.[1]

Manual chart review is the traditional method for collecting side effects information. However, it is too time-consuming and effort-intensive to be practical for routine or particularly large-scale use. Thus, the detection of adverse drug events (ADEs) by informatics techniques has been widely studied, with varied definitions and methodologies.[2 3] Hospital voluntary reporting systems have not been entirely effective for ADE detection.[4–6] Kilbridge *et al*[7] constructed an expert system with a rule-based computer program to monitor pediatric patients, exploring the demographic, encounter, laboratory, and pharmacy data to identify possible ADEs. Honigman *et al*[8] used computerized data including diagnosis codes, allergy rules, event monitoring rules, and text searching to detect ADEs in outpatients. Visweswaran *et al*[9] investigated four naive Bayes models to identify whether discharge summaries are related to ADEs. Chen *et al*[10] used an association rule mining technique against tabular-format pregnancy data to derive diverse cases of multiple drug exposure that might induce side effects. Melton and Hripcsak[11] investigated the automated detection of adverse events through computer query on the coded data generated by natural language processing (NLP).[12]

Generally, ADE studies have placed an emphasis on computerized surveillance that monitors laboratory and pharmacy data to detect patient injury. Many cases of adverse event detection simply use numeric or coded data that are derived from various sources of the structured part of patient records.[11] However, a substantial amount of valuable information resides in unstructured clinical narratives that require the use of advanced techniques when extracting the information for clinical research. NLP techniques coupled with rule based and/or machine learning (ML) techniques have been shown to be effective at processing clinical free texts for text analysis and relationship extraction.[13 14] In this manuscript we study the application of these techniques to the task of the automated extraction of drug side effects from clinical narratives in the electronic medical records of Mayo Clinic.

We developed a rule-based method using NLP techniques for discovering the relationship between drugs prescribed to psychiatry and psychology patients and their physician-asserted side effects. Separately, we implemented a method that combines both rule based and ML techniques for the extraction of sentences that might possibly include side effects and drugs. This approach was used to discover as many side effect occurrences as possible in clinical notes.

DATA

We used 237 clinical notes from patients in the psychiatry and psychology department at Mayo Clinic. They were manually reviewed by a medical expert and annotated for the side effects and the causative drugs. There are 335 pairs of side effects and the drugs that caused them (including 37 pairs that spanned two sentences). One side effect is associated with one causative drug in each pair. The longest adjective or noun phrase is marked as the side effect. The system used two-thirds of randomly selected notes for the training set (260 side effect and drug pairs) and the remaining third of the notes for the test (75 side effect and drug pairs).

METHODS

Our system was implemented in a version of the clinical Text Analysis and Knowledge Extraction System (cTAKES).[15 16] cTAKES is an open-source, comprehensive NLP system developed specifically for the clinical domain. It is built within the Apache Unstructured Information Management Architecture framework[17] which provides an efficient way to add new components such as the side effect annotator described in this paper. Variations of cTAKES have been applied to a number of use cases.[18–22]

Figure 1 shows the annotation flow of our side effect pipeline. All components except for the Side Effect annotator are standard cTAKES components described in Savova *et al*.[15 16] The Tokenizer finds token boundaries, and the Sentence Detection annotator discovers end of sentences. The Normalization annotator provides one form for the various morphological variants of the same root. POS tagger assigns tags to each word, such as noun, verb, adjective, etc. The Chunker annotator combines tokens into phrases such as noun phrases or verb phrases. The Lookup Window annotator sets the spans, for example, noun phrase, which will be passed to the Named Entity (NE) Detection module to discover clinical concepts and map them to a set of dictionaries (SNOMED-CT,[23] MeSH,[24] ICD-9,[25] NCI Thesaurus,[26] RxNorm[27]) as well as to assign a UMLS[28] semantic type. The Drug Attributes module annotates attributes of the drug: dosage, frequency, duration, route, form, strength, status change. The Status and Negation annotators assign *confirmed*, *possible*, *probable*, *negated*, *not_negated* values to the entity attributes.

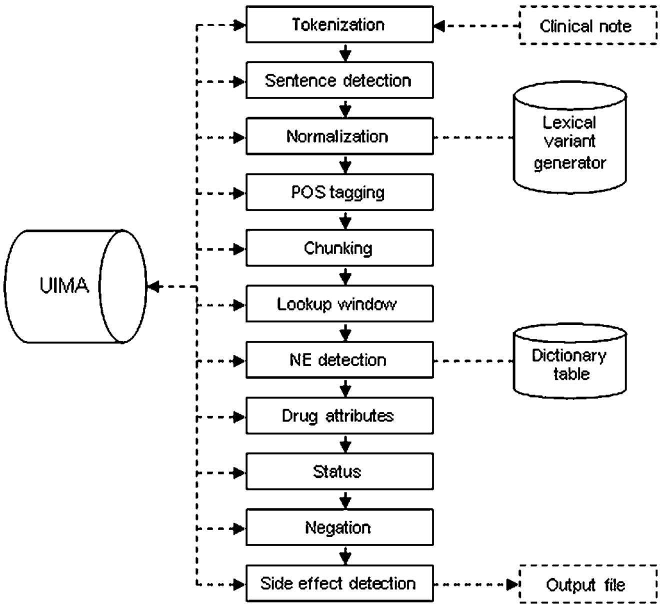


Figure 1 cTAKES (clinical Text Analysis and Knowledge Extraction System) UIMA (Unstructured Information Management Architecture) annotation flow of side effect pipeline.

Rule-based side effect extraction

We considered any NE of signs/symptoms and disorders/diseases type a potential side effect (PSE). If more than one overlapping NE was found within the same text string, then the longest (ie, the most specific) NE was selected as a PSE. Once we compiled all PSEs, a variety of rules were used to identify certain patterns that establish the UMLS semantic network *causes* relationship between the NE and drug occurrence. If a *causes* relationship was asserted, then the signs/symptoms or diseases/disorders mention became a side effect. The pattern matching rules were created manually using regular expression by examining actual patterns of the side effect expression in the training set.

For each PSE, we examined the surrounding text within a given text range (window). The window was defined as one or two sentences that contain both a PSE and a drug. If the PSE sentence (ie, the sentence that contains a PSE) contained a drug, then the window was set as this sentence. If not, the preceding sentence within the same paragraph was examined. If the preceding sentence contained a drug, then the window was set as both the PSE sentence and the preceding sentence. The pattern matching rules were then applied to identify the side effect and the causative drug pairs. The pseudo-code of side effect extraction rules is described in box 1.

Rules were prioritized based on their precision on the training data. The precision of each rule was determined by applying only a given rule on the training data and computing its precision. Specifically, the highest precision rule was applied first. If it did not find a match, then the next highest precision rule was applied. The process was iterated until either a match was found or all rules were exhausted. The detailed description for each rule follows. Of note, the rules described below are prioritized by their precision except for isInParenthesis and isInDictionary. These two rules were intended to be used as complementary to the core pattern matching rules.

In the following rule description, $^{+}$ means one or more entity, — means any string except PSE or drug, and *** means any string. Examples of each rule can be found in appendix A in the online supplementary materials.

1. isInAllergySection
   If PSE is within the Medication subsection under the allergy section, extract PSE and a corresponding drug if they occur in the same sentence.

Box 1 Pseudo code of side effect extraction

```
Identify all PSEs and drugs in a note
► For each non-negated PSE*
    – If PSE sentence contains drug, set window this sentence
    – Else if one previous sentence contains drug, set window
      PSE+previous sentence
    – Else skip this PSE
    – Apply a pattern matching rule in order of reliability
          If match found, extract PSE and drug
          Else apply next rule
    – If no match found, use drug side effect dictionary† to find
      match
```

*For detailed negation algorithm see Savova *et al*.[15]

†See item 12 in the list in the Rule-based side effect extraction section for details on isInDictionary.

PSE, potential side effect.

2. Drug$^+$—DueToWord—PSE$^+$
   Some side effects are expressed in a pattern of a drug mention followed by certain words (DueToWord) that implies a *causes* relationship to PSE. This pair is an indication of causative drug and side effect. This rule extracts those pairs.
   DueToWord={'due to', 'secondary to', 'because of', 'associate with', 'associates with', 'account for', 'accounts for'}
3. Drug$^+$***CauseVerb—PSE$^+$
   Side effects can be expressed in a pattern of a drug mention followed by certain types of verb forms (CauseVerb) that implies a *causes* or *produces* relationship to PSE. Then, this PSE is an indication of side effect.
   CauseVerb={'caused', 'causing', 'induced', 'inducing', 'resulted', 'resulting', 'yielded', 'yielding'}
4. Drug$^+$***made him/her PSE$^+$
   If a drug mention is followed by the phrase 'made him/her' and PSE, this PSE is an indication of side effect to this drug.
5. hasSideEffectWord
   Side effects can appear with direct side effect mentions (SideEffectWord). If PSE appears with SideEffectWord and this SideEffectWord is not negated, this rule extracts a corresponding PSE and drug.
   SideEffectWord={'side effect', 'side effects', 'reaction', 'reactions'}
6. PSE$^+$—DueToWord—Drug$^+$
   Some side effects are expressed in a pattern of PSE followed by certain words (DueToWord) that implies a *causes* relationship to a drug. This pair is an indication of causative drug and side effect.
   DueToWord={'due to', 'secondary to', 'because of', 'associated with', 'induced by', 'attributed to'}
7. hasSideEffectAsPSE
   In some cases, the side effect is not specifically expressed as signs/symptoms or diseases/disorders but simply expressed as a 'side effect' or 'reaction' textual string. If a given PSE is SideEffectWord defined in 5, is not mentioned together with certain words ('discussed,' 'concerned'), and no other PSEs appear along with it, then this rule extracts SideEffectWord itself as a side effect.
8. NoteVerb—PSE$^+$—with—Drug$^+$
   If certain verbs (NoteVerb) that imply the meaning of 'notification', 'development', or 'description' appear with PSE in a particular pattern, this PSE is an indication of side effect.
   NoteVerb={'noted', 'developed', 'reported', 'described'}
9. DiscontVerb—Drug$^+$—because/after—PSE$^+$
   Or Drug$^+$***DiscontVerb—because/after—PSE$^+$
   If a patient discontinued or decreased (DiscontVerb) a drug because of PSE or after experiencing PSE, this PSE is an indication of side effect to this drug.
   DiscontVerb={'discontinued', 'tapered', 'decreased', 'stopped'}
10. PSE$^+$—after taking/after starting—Drug$^+$
    If a patient has some PSE after taking a drug, this PSE might be a side effect of this drug.
11. isInParenthesis
    Some side effects appear within parentheses right after the drug is mentioned. If the word 'made' or 'got' is within parentheses along with PSE, we extract this PSE as a side effect.
12. isInDictionary
    Some side effects might appear with an indication word or phrase that is not covered by our pattern matching rules, or they might be expressed in an indirect way without a clear indication word or phrase, which makes it difficult for pattern matching rules to identify them. A dictionary of drug and side effect pairs could be utilized to catch those hard cases. Our psychiatry and psychology department has a list of 44 drugs that are commonly used in practice. We manually built a dictionary of those drugs and their side effects by compiling side effects from an online drug information resource (http://www.drugs.com/). However, it should be noted that this list is a small subset of all drugs in this study. Interestingly, some drugs have paradoxical reactions—that is, drug treatment has the opposite response which would usually be expected. For example, 'diazepam' is used for the management of anxiety disorder. However, 'anxiety' is also listed as an adverse reaction of this drug (source: MICROMEDEX). This situation can potentially cause false positive side effect extraction when using the dictionary lookup method. Therefore, we manually removed those cases from our dictionary to avoid potential false positive side effects.

If the drug and PSE within the window are in the side effect dictionary, the system extracts them. Note that this dictionary lookup is only applied if all previous pattern matching rules do not find any matches.

Side effect sentence extraction using machine learning and rules

This task is to extract sentences that contain a side effect and its causative drug (we call it simply the 'side effect sentence'). Note that the term *sentence* denotes the *window* in box 1. A human expert can further examine those sentences to validate the drug–side effects pairs. This semi-automatic procedure focuses on extracting a higher number of side effect occurrences to discover as many pairs as possible in clinical notes.

In the real world, not all side effect expressions fit into predefined patterns in our rules. Since ML techniques are commonly used for discovering hidden patterns, we used ML to refine the performance of automatic side effect sentence extraction. For this task, first we extracted all sentences that contain side effects and drugs found by our previous pattern matching rules. Second, expert judgments on the automatically extracted PSE sentences (defined in box 1) were used to train and test a C4.5 classifier from the Weka ML package,[29] independently from the first step. Finally, the union of the results from ML and rule based techniques was considered as our final set of side effect sentences. The allergy section was not considered for ML side effect sentence classification because it can be easily handled by a rule.

The features used in the C4.5 classifier consist of (1) the presence or absence of explicit side effect keywords (see SideEffect Word in table I of appendix B in the online supplementary materials) and (2) the locations of the other keywords. The keywords were manually selected. The location features are the pre-defined location keywords within the side effect sentence (see table II of appendix B in the online supplementary materials). Similar keywords are grouped together into a set of meta-keywords. Each meta-keyword becomes an element of a C4.5 feature vector—that is, the C4.5 classifier uses 21 dimensional feature vectors, in which each element represents the presence/absence of SideEffectWord and the location of the other meta-keywords. In this task, a large portion of data consists of non-side effect sentences (the training set has 190 side effect sentences and 1389 non-side effect sentences; the test set has 63 side effect sentences and 552 non-side effect sentences). This unbalanced data could result in a relatively lower accuracy on a minority class (ie, side effect sentences), although overall accuracy would not deteriorate significantly. To handle this problem, a number of approaches have been studied.[30–32] In this paper, we used the up-sampling technique to balance examples between the majority and minority classes.

Evaluation metrics

We used standard metrics to evaluate system performance for the discovery of side effects/drug pairs and sentences containing them:

$$precision = \frac{truePositives}{truePositives + falsePositives}$$

$$recall = \frac{truePositives}{truePositives + falseNegatives}$$

$$F - score = \frac{2*(precision*recall)}{(precision + recall)}$$

True positives represent cases in which the side effect and causative drug extracted by the system match the gold standard pair. Both exact and partial matches were used for evaluation. Exact matches are those cases where the offsets of the system outputs are exactly the same as those of the gold standard. Partial matches are the cases where there is an overlapped offset between the system output and the gold standard. For example, the side effects 'marked insomnia' in the gold standard and 'insomnia' from the system are considered a partial match but not an exact match. Exact matches are a subset of partial matches. In this paper, we emphasized partial match results because they still capture a main concept when the side effect is associated with a descriptive word.

RESULTS

Figure 2 shows a simple example of a drug side effect extracted by the system along with its attributes.

Rule-based side effect extraction

Table 1 shows evaluation results as obtained on the test set. Results are differentiated by the inclusion of the side effects found in the allergy section. The system produced a partial match F score of 0.799 when the allergy section was included and 0.554 when the allergy section was excluded. We attribute this to the fact that side effects other than those mentioned in the allergy section are much more difficult to identify. As expected, the exact match F score was lower than the partial match F score. In general, recall is higher than precision.

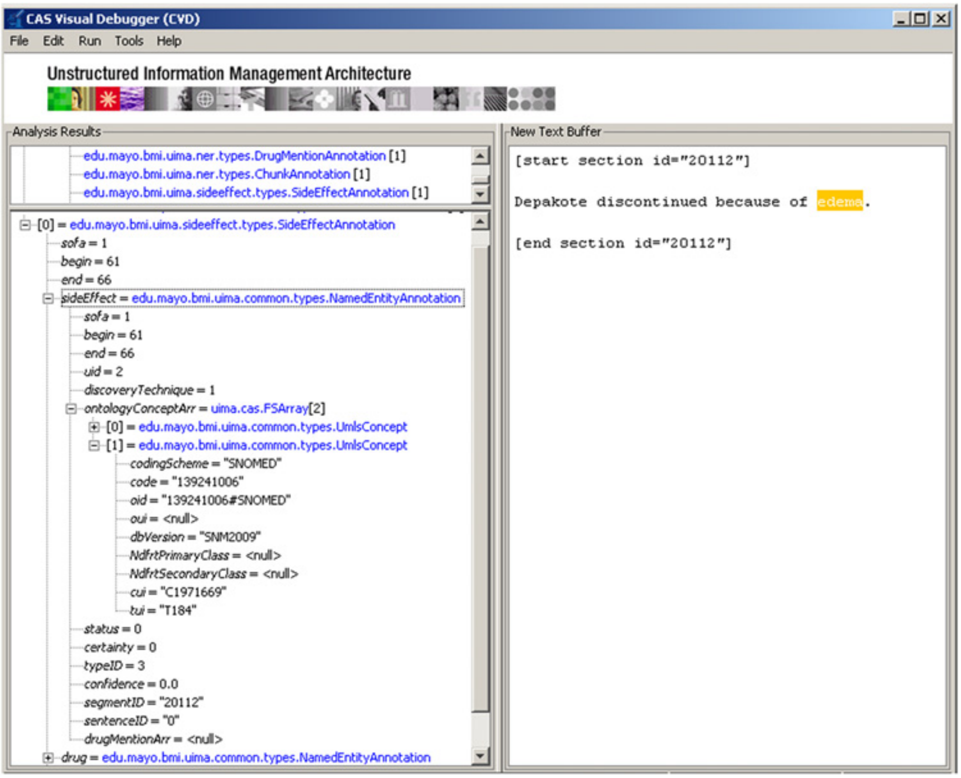
Some side effects in clinical narratives were described indirectly without clear indication words (see examples of isInDictionary, appendix A in the online supplementary materials). Context understanding is necessary to properly catch those side effects. Because of that, pattern matching rules fail to recognize these hard cases, which results in false negatives.

Side effect sentence extraction using machine learning and rules

Our baseline simply extracts all sentences containing PSE and drug pairs. Table 2 shows the baseline results. As expected, recall was high but precision was very low because we merely extracted side effect sentence candidates without any filtering process. Table 3 shows the evaluation results of our methods on the test set which produced much higher F scores although recalls were lower than those of the baseline. In table 3, the 'Rules' column contains the results of using only a rule-based method, and the column marked 'Rules+C4.5' is the hybrid method combining rules and C4.5. Both recall and SE_recall in the hybrid approach were higher than those of the rule-based method at the expense of precision. Of note, our focus was on achieving a high recall of side effect occurrences with tolerable precision. SE_recall results (0.893 and 0.861) were higher than the partial match recall results presented in table 1 (0.827 and 0.639).

DISCUSSION

Some side effect mentions in the gold standard include modifiers, for example, in '*marked* insomnia' there is a severity


CAS Visual Debugger (CVD)
File
Edit
Run
Tools
Help
Unstructured Information Management Architecture
Analysis Results
New Text Buffer

```
edu.mayo.bmi.uima.ner.types.DrugMentionAnnotation [1]
edu.mayo.bmi.uima.ner.types.ChunkAnnotation [1]
edu.mayo.bmi.uima.sideeffect.types.SideEffectAnnotation [1]

[0] = edu.mayo.bmi.uima.sideeffect.types.SideEffectAnnotation
  sofa = 1
  begin = 61
  end = 66
  sideEffect = edu.mayo.bmi.uima.common.types.NamedEntityAnnotation
    sofa = 1
    begin = 61
    end = 66
    uid = 2
    discoveryTechnique = 1
    ontologyConceptArr = uima.cas.FSArray[2]
      [0] = edu.mayo.bmi.uima.common.types.UmlsConcept
      [1] = edu.mayo.bmi.uima.common.types.UmlsConcept
        codingScheme = "SNOMED"
        code = "139241006"
        oid = "139241006#SNOMED"
        oui = <null>
        dbVersion = "SNM2009"
        NdfrtPrimaryClass = <null>
        NdfrtSecondaryClass = <null>
        cui = "C1971669"
        tui = "T184"
    status = 0
    certainty = 0
    typeID = 3
    confidence = 0.0
    segmentID = "20112"
    sentenceID = "0"
    drugMentionArr = <null>
  drug = edu.mayo.bmi.uima.common.types.NamedEntityAnnotation
```

```
[start section id="20112"]

Depakote discontinued because of edema.

[end section id="20112"]
```

Figure 2 Drug side effect annotation visualized through the UIMA's (Unstructured Information Management Architecture) CAS Visual Debugger. The right window shows a clinical narrative snippet processed to populate annotations as they appear in the left window. The bottom left window shows the list of side effects and causative drugs identified by the system and their attributes.

Table 1 Evaluation of drug side effect extraction on the test set

	Including allergy section		Excluding allergy section	
	Exact	Partial	Exact	Partial
Precision	0.464	0.774	0.390	0.489
Recall	0.520	0.827	0.444	0.639
F score	0.491	0.799	0.416	0.554

modifier ‘marked’ describing the main clinical concept of ‘insomnia.’ The current system does not discover associated severity modifiers, which led to a relatively lower exact match recall as compared to the partial match recall.

A dictionary lookup was helpful to find indirect side effect mentions. However, it searches for matches without examining particular patterns, which resulted in false positives. For example, ‘She has noted some upset stomach with Cymbalta. If the *upset stomach* continues, she can discuss with her primary-care provider whether to go back on *Prozac*.’ *Upset stomach* was falsely extracted as Prozac’s side effect. Cases like that require deeper semantic processing of the text such as flagging hypothetical constructions.

Many false positive cases occurred in sentences containing multiple drugs and PSEs (17 out of 19 false positives). Some false positives occurred when a two-sentence window was used (7 cases: most overlapped with the previous 17 cases). For example, ‘He discontinued *Antabuse* before drinking alcohol. In July 1995 he did attempt to attend treatment at hospital, and was dismissed due to *depression*.’ *Depression* was falsely extracted as the side effect of Antabuse. This again draws attention to the need for more sophisticated semantic processing techniques aiming at abstracting the meaning of the text through semantic role labeling as well as temporal reasoning[34] to distinguish different event occurrences.

There are several main sources for the false negatives cases. The first one is that the drug and/or side effect was not recognized as an NE by the system (7 out of 13 false negatives) such as these side effects: *could not sleep well*, *knocked him for a loop*. The second one stems from indirect expression of side effects (2 cases). The third reason lies in incorrect negation (2 cases). For example, in the sentence, ‘She tolerated the switch well and did *not* complain of any side-effects other than *fatigue*,’ the side effect indication ‘side-effects’ was negated and caused it to miss the actual side effect ‘*fatigue*.’ The last error source is due to co-referenced drugs. For example, ‘He took *Paxil* from November 2005 to February 2006. He reported *sluggish* with *this medication* and stopped *it*,’ where *this medication* and *it* refer to Paxil. Our system does not have a co-reference resolution module, thus the side effect ‘*sluggish*’ was not extracted.

In this paper, side effect sentence classification focused on extracting as many sentences as possible that may contain side effects and their causative drugs despite some loss of precision. Therefore, the system could cover more side effect mentions and enable a human expert to extract them with less effort than by simply reviewing the clinical notes in full. An ML approach using C4.5 leveraged the performance of side effect sentence classification. Our system was able to extract most side effect sentences. The main error sources were: (1) side effect is a descriptive phrase such as *make his compulsions worse*, *sensation of kidneys hurting*, and (2) incorrect segmentation of sentences that contain side effect and drugs.

Table 2 Evaluation of the baseline side effect sentence extraction on the test set

	Including allergy section	Excluding allergy section
Precision	0.102	0.043
Recall	0.984	0.926
F score	0.186	0.083
SE_recall*	0.960	0.917
90% CI of SE_recall†	0.901 to 0.987	0.804 to 0.971

*Number of side effect and drug pairs in retrieved side effect sentences/total number of side effect and drug pairs.
†CI was obtained by using a modified Wald method.[33]
SE, side effect.

Table 3 Evaluation of side effect sentence extraction on the test set

	Including allergy section		Excluding allergy section	
	Rules	Rules+C4.5	Rules	Rules+C4.5
Precision	0.862	0.640	0.750	0.423
Recall	0.875	0.891	0.778	0.815
F score	0.868	0.745	0.764	0.557
SE_recall*	0.867	0.893	0.806	0.861
90% CI of SE_recall†	0.788 to 0.920	0.820 to 0.940	0.675 to 0.893	0.738 to 0.934

*Number of side effect and drug pairs in retrieved side effect sentences/total number of side effect and drug pairs.
†CI was obtained by using a modified Wald method.[33]
SE, side effect.

Some potential ways to improve the system in the future have been observed. A flexible window that utilizes syntactic or semantic information might increase the accuracy of side effect extraction. Applying co-reference resolution could increase recall of side effects. To extract more informative side effects, the system needs to extract the associated modifiers and qualifiers with those named entities. Deeper semantic processing of the text is needed to distinguish between a broad set of relationships such as the UMLS semantic network relationships of *causes*, *indicates*, *treats*, and *manages*. Another crucial feature is temporal relation discovery, in particular the *after* relationship between administered medication and signs/symptoms and diseases/disorders. We believe that using temporal information could increase the precision of side effect extraction.

LIMITATIONS

The side effect extraction presented in this study was specifically applied to the domain of psychiatry. Our approach needs to be broadly tested against a variety of domains. For other domains which use a different style of side effect descriptions in their clinical notes, it might be necessary to adjust pattern matching rules to properly extract side effects. Our system extracts clearly defined side effects and causative drugs with limited indication words and patterns. Thus, a complicated or indirect description of side effect might not be identified by the system. A relatively a small corpus was used; a larger corpus would be beneficial to make the system more robust.

CONCLUSION

We investigated techniques to extract drug side effects from the clinical text building on an existing comprehensive clinical NLP system, cTAKES. Pattern matching rules were able to extract clearly expressed side effects. However, more sophisticated semantic processing is required to handle complex side effect descriptions. Side effect sentence extraction relaxed the requirement for finding side effects and discovered more occurrences of them at the expense of precision. This latter approach is feasible for a semi-automated paradigm which requires human validation for the extraction of individual side effects and the causing drugs. The Side Effects module will be released as an open source cTAKES (http://www.ohnlp.org) component under the Apache licensing mechanism. Its pattern matching rules are easily customizable.

Acknowledgments We would like to thank Josiah Allen for annotating the gold standard, Cindy Bandel and Donna Ihrke for building the side effect dictionary, and James Masanz for setting up the Knowtator framework for gold standard annotation. We also thank anonymous reviewers for their valuable comments.

Funding This work was funded by an AT&T foundation grant (PI Kocher) and SHARP 4 grant 90TR000201 (PI Chute).

Competing interest None.

Ethics approval The Mayo Clinic Institutional Review Board provided ethics approval.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Evans WE,** McLeod HL. Pharmacogenomics—drug disposition, drug targets, and side effects. *N Engl J Med* 2003;**348**:538—49.
2. **Bates DW,** Evans RS, Murff H, *et al*. Detecting adverse events using information technology. *J Am Med Inform Assoc* 2003;**10**:115—28.
3. **Wang X,** Hripcsak G, Markatou M, *et al*. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* 2009;**16**:328—37.
4. **Cullen DJ,** Sweitzer BJ, Bates DW, *et al*. Preventable adverse drug events in hospitalized patients: a comparative study of intensive care and general care units. *Crit Care Med* 1997;**25**:1289—97.
5. **Field TS,** Gurwitz JH, Harrold LR, *et al*. Strategies for detecting adverse drug events among older persons in the ambulatory setting. *J Am Med Inform Assoc* 2004;**11**:492—8.
6. **Jha AK,** Kuperman GJ, Teich JM, *et al*. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *J Am Med Inform Assoc* 1998;**5**:305—14.
7. **Kilbridge PM,** Noirot LA, Reichley RM, *et al*. Computerized surveillance for adverse drug events in a pediatric hospital. *J Am Med Inform Assoc* 2009;**16**:607—12.
8. **Honigman B,** Lee J, Rothschild J, *et al*. Using computerized data to identify adverse drug events in outpatients. *J Am Med Inform Assoc* 2001;**8**:254—66.
9. **Visweswaran S,** Hanbury P, Saul M, *et al*. Detecting adverse drug events in discharge summaries using variations on the simple Bayes model. *AMIA Annu Symp Proc* 2003:689—93.
10. **Chen Y,** Pedersen LH, Chu WW, *et al*. Drug exposure side effects from mining pregnancy data. *ACM SIGKDD Explorations Newsletter* 2007;**9**:22—9.
11. **Melton GB,** Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc* 2005;**12**:448—57.
12. **Hripcsak G,** Bakken S, Stetson PD, *et al*. Mining complex clinical data for patient safety research: a framework for event discovery. *J Biomed Inform* 2003;**36**:120—30.
13. **Demner-Fushman D,** Chapman W, McDonald C. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009;**42**:760—72.
14. **Uzuner O,** Zhang X, Sibanda T. Machine learning and et al to assertion classification. *J Am Med Inform Assoc* 2009;**16**:109—15.
15. **Savova G,** Masanz J, Ogren P, *et al*. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507—13.
16. **Savova GK,** Kipper-Schuler K, Buntrock JD, *et al*. *UIMA-Based Clinical Information Extraction System. LREC 2008: Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*. Marrakech, Morocco: The 6th International Conference on Language Resources and Evaluation, 2008.
17. **UIMA.** http://uima-framework.sourceforge.net/.
18. **Kullo IJ,** Fan J, Pathak J, *et al*. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* 2010;**17**:568—74.
19. **D'Avolio LW,** Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010;**17**:375—82.
20. **Nielsen RD,** Masanz J, Ogren P, *et al*. An architecture for complex clinical question answering. *Proceedings of the 1st ACM International Health Informatics Symposium*. Arlington, VA: ACM New York, NY, 2010:395—9.
21. **Sohn S,** Savova GK. *Mayo Clinic Smoking Status Classification System: Extensions and Improvements, 2009 Improvements*. San Francisco, CA: AMIA Annual Symposium, 2009:619—23.
22. **Savova GK,** Olson J, Murphy SP, *et al*. *Pharmacogenomic Study of Tamoxifen and Aromatase Inhibitors in Women with Breast Cancer: Drug Extraction Through Natural Language Processing*. Rochester, MN: PGRN meeting, 2009.
23. **UIMA.** http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html.
24. **MeSH.** http://www.nlm.nih.gov/mesh/.
25. **ICD-9.** http://www.cdc.gov/nchs/icd/icd9.htm.
26. **NCI Thesaurus.** http://ncit.nci.nih.gov/.
27. **RxNORM.** http://www.nlm.nih.gov/research/umls/rxnorm/.
28. **UMLS.** http://www.nlm.nih.gov/research/umls/.
29. **Hall M,** Frank E, Holmes G, *et al*. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 2009;**11**:10—18.
30. **Japkowicz N,** Stephen S. The class imbalance problem: A systematic study. *Intell Data Anal* 2002;**6**:429—50.
31. **Maloof M.** Learning when data sets are imbalanced and when costs are unequal and unknown. *Proc of the ICML-2003Workshop:Learning with Imbalanced Data Sets II*. Washington, DC: AAAI Press, 2003:73—80.
32. **Sohn S,** Kim W, Comeau DC, *et al*. Optimal Training Sets for Bayesian Prediction of MeSH® Assignment. *J Am Med Inform Assoc* 2008;**15**:546—53.
33. **Agresti A,** Coull B. Approximate is better than "Exact" for interval estimation of binomial proportions. *Am Stat* 1998;**52**:119—26.
34. **Savova GK,** Bethard S, Styler W, *et al*. Towards temporal relation discovery from the clinical narrative. *AMIA Annu Symp Proc* 2009;**2009**:568—72.