MedEx: a medication information extraction system for clinical narratives

Hua Xu,[1] Shane P Stenner,[1,2] Son Doan,[1] Kevin B Johnson,[1,3] Lemuel R Waitman,[1] Joshua C Denny[1,2]

[1]Department of Biomedical Informatics, Vanderbilt University, School of Medicine, Nashville, Tennessee, USA [2]Department of Medicine, Vanderbilt University, School of Medicine, Nashville, Tennessee, USA [3]Department of Pediatrics, Vanderbilt University, School of Medicine, Nashville, Tennessee, USA

Correspondence to
Dr Hua Xu, Department of Biomedical Informatics, Vanderbilt University, 2209 Garland Ave, 412 EBL, Nashville, TN 37232, USA; hua.xu@vanderbilt.edu

Received 9 August 2009
Accepted 21 October 2009

ABSTRACT

Medication information is one of the most important types of clinical data in electronic medical records. It is critical for healthcare safety and quality, as well as for clinical research that uses electronic medical record data. However, medication data are often recorded in clinical notes as free-text. As such, they are not accessible to other computerized applications that rely on coded data. We describe a new natural language processing system (MedEx), which extracts medication information from clinical notes. MedEx was initially developed using discharge summaries. An evaluation using a data set of 50 discharge summaries showed it performed well on identifying not only drug names (F-measure 93.2%), but also signature information, such as strength, route, and frequency, with F-measures of 94.5%, 93.9%, and 96.0% respectively. We then applied MedEx unchanged to outpatient clinic visit notes. It performed similarly with F-measures over 90% on a set of 25 clinic visit notes.

INTRODUCTION

In electronic medical records (EMRs), medication data are often recorded in narrative clinical notes. For example, hospital discharge summaries usually contain some instructions on medications after discharge (eg, 'Will start Orapred x 5 days and increase Pulmicort to 0.5 mg inh BID'), and outpatient clinic visits often document medication changes. Even in electronic prescribing tools, free text inputs such as medication signatures are often allowed[1] (eg, 'prednisone 10 mg tablets—take 4 tablets for 4 days, then 2 tablets for 4 days'). The free-text medication data is inaccessible to other computerized applications that rely on coded data in daily healthcare settings (eg, electronic medication reconciliation systems), as well as to clinical research that uses structured medication data, such as EMR-based post-marketing surveillance.

Medication errors often increase when a patient is transferred from one care setting to another.[2 3] Medication reconciliation is a formal process for creating a most complete and accurate list of a patient's medications, for the purpose of supporting correct medication orders. In 2005, the Joint Commission listed 'medication reconciliation across the care continuum' as a National Patient Safety Goal in 2005.[4] Given the increasing use of EMRs, automated medication reconciliation methods[5 6] have received great attention. One particular challenge involves the heterogeneity of clinical data, which consists of both coded and narrative medication data.

EMR-based clinical research often requires detailed medication information as well. Ongoing EMR-based pharmacogenomic studies at Vanderbilt University Medical Center require accurate identification of drug exposure and drug responses of patients based on available EMR data. In a heterogeneous EMR such as ours with data inputs from free-text clinical documentation, electronic prescription writers, and cross-disciplinary problem lists (mostly in free text), a complete understanding of a patient's medication status requires extraction of medication information from all sources including clinical narratives.

In clinical notes, medication data are often expressed with medication names and signature information about drug administration, such as dose, route, frequency, and duration. In this study, all the items except for 'drug name' in table 1 are referred as medication signatures. All these types of information are necessary to create an accurate medication profile for a patient. In this paper, we describe an automated medication extraction system (MedEx), which can accurately extract medication names and signatures from clinical narratives.

BACKGROUND

Over the last 2 decades, there have been many efforts to apply natural language processing (NLP) technologies to clinical text. The Linguistic String Project[7 8] developed one of the earliest clinical NLP systems that used very comprehensive semantic and syntactic knowledge. Friedman and her colleagues[9] developed a clinical NLP system called MedLEE (Medical Language Extraction and Encoding System), which was originally designed for decision support applications in the domain of radiology reports of the chest,[10 11] and was extended to other domains, such as mammography[12] and discharge summaries[13] later. SymText[14 15] is a NLP system developed at the University of Utah, which has been used for various applications such as encoding chief complaints into ICD-9 codes[15] and extracting pneumonia-related finings from chest radiograph reports.[16 17] KnowledgeMap[18] is a NLP system developed at Vanderbilt University and it has been used to extract medical concepts from clinical and education documents.[19] Other research groups have also developed various NLP systems[20–24] for processing clinical text and have shown good performance in different sub-domains of medicine.

Several studies have worked on extracting drug names from clinical notes using NLP. In 1996, Evans *et al*[25] showed that drug and dosage phrases in

Table 1 Semantic categories of medication findings

Semantic categories	Examples
DrugName	'Lisinopril', 'Famotidine'
Strength	'50 mg', '500/50'
Route	'by mouth', 'intravenous'
Frequency	'twice daily', 'every 2 days'
Form	'tablet', 'ointment'
Dose Amount	'take one tablet'
IntakeTime	'cc', 'at 10 am'
Duration	'for 10 days'
Dispense Amount	'dispensed #30'
Refill	'refills: 2'
Necessity	'prn', as needed

discharge summaries can be identified by the CLARIT system with an accuracy of 80%. Chhieng *et al*[26] reported a precision of 83% when using a string matching method to identify drug names in clinical records. Levin and colleagues[27] developed an effective rule-based system to extract drug names from anesthesia records and map to RxNorm[28] concepts with 92.2% sensitivity and 95.7% specificity. Sirohi and Peissig[29] studied the effect of lexicon sources on drug extraction.

Current medication extraction systems extract drug names with high accuracy, but they perform less well on identifying drug signature information. Furthermore little has been done on extracting contextual level information such as status about medications (eg, 'start', 'increase', or 'discontinue'). A recent study by Gold *et al*[30] reported a regular expression based approach for extracting drug names and signature information such as dose, route, frequency. Evaluation on a data set of 26 discharge summaries showed that drug names were identified with a precision of 94.1% and a recall of 82.5%, but other signature information such as dose and frequency had much lower precisions. There are several commercial systems that extract drug names and signature information from clinical notes, such as LifeCode, FreePharma, and Coderyte. Jagannathan *et al*[31] 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.

DESIGN OBJECTIVES

Sentences containing medications can be complicated for parsing. One sentence can have multiple medications, and one medication can contain multiple sets of signatures. For example, it can contain structures such as 'Midrin 2 po initial then 1 po q6 h prn #5'. Understanding the contextual level information of medications, such as status (eg, 'start' vs 'stop') and temporal information ('last year' vs 'now'), is even more challenging. Tasks such as to determine whether a drug reference is to an allergy or therapy often require information beyond the sentence level (eg, need to identify sections—'Current medication' vs 'Allergies'). This complexity of medication-related textual information indicates that simple methods such as the regular expression based approach may not be effective enough to capture all necessary medication-related fields. Our ultimate goal is to develop a medication parser that can accurately extract drug names, signatures, and contextual information. We are building the medication parser using a semantic-based approach, similar to MedLEE,[9] but using semantic types and patterns in a much finer granularity. As a first step toward that goal, we focused on accurately extracting drug names and signature information from clinical narratives. By integrating a semantic tagger and a Chart parser, we expect to capture medication names and major categories of signatures (eg, dose, route, and frequency) information with F-measures over 90%.

SYSTEM DESCRIPTION

A simple semantic representation model for prescription-type of medication findings was defined first and a medication extraction system (MedEx) was then developed to map medication text into structured representation using a sequential semantic tagger and a Chart parser. Figure 1 shows an overview of the MedEx system with an example.

A medication representation model

We define a 'medication finding' as all relevant medication information provided in the text—including the central finding of a medication name, and its modifiers such as signature information and contextual information such as status and temporal information. A complete medication representation model should include all above elements. In this study, we focused on algorithms to identify medication names and their signatures by conducting a manual analysis of clinical text in the training set. The analysis involves determining the underlying semantic categories and the semantic relationships among those categories necessary to model medication findings. Following a similar approach described in Friedman *et al*,[32] we identified 11 semantic categories that are related to medication findings, as shown in table 1. Semantic relations among those categories are also identified by manually reviewing medication sentences. A formal model was designed to represent the medication findings. Two main components of the model are 'Med finding' and 'Sig modifier', which represent the structures of medication findings and signature modifiers, respectively. Figure 2 shows the simplified representation of medication findings and signature modifiers, using the linear notation for Conceptual Graphs.[33] A concept is enclosed in square brackets and followed by the relations associated with it. Each relation appears in parentheses and its values are specified by another concept that follows the relations after an arrow ('→'). The model defined in figure 2 indicates that a 'Med finding' can be formed by a central concept (*Drug name*), zero or more signature modifiers (*Sig modifier*), and zero or more temporal modifiers (*Tem modifier*). The representation of 'Sig modifier' consists of different types of modifiers

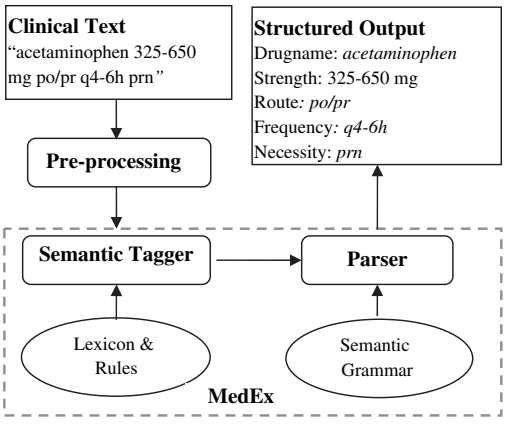

Figure 1 An overview of the MedEx system.

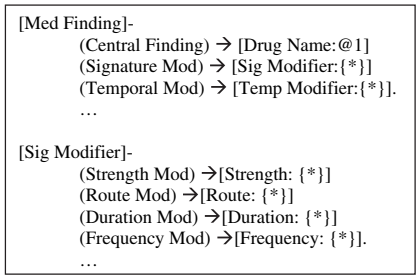

Figure 2 A simplified model to represent medication findings and signature modifiers in the notation of conceptual graphs. A concept is enclosed in square brackets and followed by the relations associated with it. Each relation appears in parentheses and its values are specified by another concept that follows the relations after an arrow ('→'). The number of values that a relation is permitted to have (its cardinality) is defined by following constraint:{*} means that the relation must have 0 or more values;{*}@>0 means that the relation must have 1 or more values; and:{*}@<2 means that the relation must have 0 or 1 values; the default cardinality is exactly 1.

including relations such as *Strength mod*, *Route mod*, and *Frequency mod*.

Pre-processing

The goal of the pre-processing step in MedEx is to determine the sentence boundaries in a clinical note. In this study, we assume a sentence is the basic unit for extracting information related to one drug. We used an existing sentence boundary detection program described in Denny *et al*,[34 35] which is a rule-based program.

Semantic tagging

Semantic tagging is one of the most important steps of MedEx and it significantly affects the performance of the system. A semantic tagger will break an input sentence into tokens and label proper words or phrases with a semantic category as described above. One widely used method for tagging is to look up terms in a predefined semantic lexicon file, which ideally contains all possible terms and their variants. Another simple tagging method involves using regular expressions to label terms. In this study, we developed a robust sequential tagger that consists of (1) an initial tagging step that combines lookup and regular expression tagging methods, (2) a disambiguation step that transforms the initial ambiguous tags into the final tags based on a set of pre-defined context-based rules.

The initial tagging step combines a lookup tagger and a regular expression tagger because different semantic pieces of a medication finding require different tagging methods. For drug names and forms, there are many lexicon sources; therefore a lookup tagger is very suitable. We generated lexicon files of drug names from RxNorm[28] by combining terms from normalized drug forms including IN (Ingredient, eg, *Fluoxetine*), BN (Brand name, eg, *Prozac*), SCDC (Ingredient+Strength, eg, *Fluoxetine 4 mg/ml*), SCDF (Ingredient+Form, eg, *Fluoxetine Oral solution*), and SCD (Ingredient+Strength+Form, eg, *Fluoxetine 4 mg/ml Oral solution*). If a drug finding is tagged as SCDC, SCDF, or SCD, it is straightforward to further decompose it into DrugName, Strength, and Form, based on relations within the RxNorm. For example, the DrugName 'Fluoxetine' (as an ingredient) can be directly obtained from a SCDC (Ingredient+Strength) drug 'Fluoxetine 4 mg/ml' based on the 'has ingredient' relation between 'Fluoxetine 4 mg/ml' and 'Fluoxetine'. Then the rest of the phrase '4 mg/ml' can also be obtained as the Strength of the drug. In this study, we will count both Drug name and Strength as correct if a SCDC drug is identified. Drug names from RxNorm contain some English words, such as 'air' and 'sleep', which are not true drug names occurring in clinical notes. Therefore we compared RxNorm drug names with a list of general English words (the SCOWL list[36]). Ambiguous words were manually reviewed by a physician and unlikely drug terms were removed from the drug name list. Some drug names (eg, abbreviations such as 'Vit-C' for 'Vitamin C') that we saw in the training set, but not in the RxNorm, were also added into the lexicon file. The lookup tagger maps a drug name to its longest match in the lexicon file. Other types of information are more suitable for a regular expression tagger. For example, frequency information such as 'q4h or q6h' can be easily captured by defining regular expressions such as 'q\dh'. Above two types of taggers are combined in a sequential manner (the lookup tagger followed by the regular expression tagger) and most of medication related terms can be tagged in this way.

The second step of tagging is to disambiguate tags that can be associated with two or more semantic categories. For example, a number tag (NUM) can be Strength (eg, 'Augmentin *875*'), Dose amount (eg, 'Take *2*'), or Dispense amount (eg, 'dispense # *30*'). Pre-defined rules that are based on the context around the ambiguous terms were used to determine the appropriate semantic categories. For example, a simple rule can be 'If a Num tag follows a DrugName tag, replace the Num tag with Strength'. Sometimes, drug names can also be semantically ambiguous (eg, drugs vs lab tests). For example, 'Potassium' can be a drug (eg, 'take Potassium') or be a lab test ('potassium level is normal'). Simple rules are also developed in this stage to remove false positive drug names, based on the contextual words around the possible drug names. For example, if words such as 'level', 'lab', and 'test' are found around a possible drug term, that term will not be labeled as DrugName.

Figure 3 shows an example of input, output, and intermediate steps of the sequential tagger. At the first step, the lookup tagger labels 'Augmentin' as 'DrugName', and the regular expression tagger labels '875' as 'Num', and 'q 8hrs' as 'Frequency'. At the disambiguation step, the Num tag was replaced by a 'Strength' tag based on a pre-defined rule.

Parsing

The parsing component of MedEx uses a context-free grammar to parse textual sentences into structured forms, via a Chart Parser,[37] a dynamic programming parsing method. We used an existing implementation of a Top-down Chart Parser in the Natural Language Tool Kit in Python.[38] The grammar is a semantic grammar that delineates semantic relations and structure, as revealed by the semantic representation model described above. A simplified version of the grammar is shown in figure 4 in Backus-Naur Form.[39] According to the partial grammar, a sentence (S) can contain a list (DRUGLIST) of drug findings (DRUG). One drug finding (DRUG) can be a drug with a single set of signatures (DGSSIG) or a drug with multiple sets of signatures (DGMSIG). For a drug with a single set of signatures (DGSSIG), it has to have

Input:	Augmentin 875 q 8hrs		
Initial:	Augmentin	875	q 8hrs
Tags	DrugName	Num	Frequency
Disambiguation:	Augmentin	875	q 8hrs
Tags	DrugName	Strength	Frequency

Figure 3 An example of the sequential semantic tagger.

```
<S> ::= <DRUGLIST>
<DRUGLIST> ::= <DRUG> | <DRUG> <DRUGLIST>
<DRUG> ::= <DGSSIG> | <DGMSIG>
<DGSSIG> ::= <DGN> | <DGN> <SIG>
<DGN> ::= <IN> | <BN> | <SCDC> | <SCDF> | <SCD>
<SIG> ::= <DOSE> | <FORM> | <RUT> | <FREQ> | <DOSE>
        <FORM> | <DOSE> <FREQ> …
```

Figure 4 Partial representation of the semantic grammar.

a drug name (DGN—from five RxNorm drug types), and zero or one set of signature modifiers (SIG), which can be DOSE (Dose), FORM (Form), RUT (Route), FREQ (Frequency), or their combinations.

If the Chart parser fails, a regular expression based Chunker in Natural Language Tool Kit is used to process the medication sentences. The Chunker finds medication phrases that consist of drug names and zero or more signatures using simple regular expressions. For example, medication phrases can be defined as regular expressions such as 'DrugName (DOSE|FORM|RUT|FREQ)*', which indicates a medication phrase can be composed by one drug name followed by zero or more signature items including 'Dose', 'Form', 'Route', and 'Frequency'. It improves the parser's capability to getting partial medication information, when the Chart parser cannot map the sentence to structured outputs defined by the grammar. Output from the parser is represented as a parse tree in Python. Final structured outputs (in figure 1) are extracted from the parse tree.

Evaluation

In this study, we used clinical notes from the Synthetic Derivative (SD) database, which is a de-identified copy of the EMR at Vanderbilt University Medical Center. Clinical notes were de-identified using DE-ID, a commercially available software package from University of Pittsburgh Medical Center, combined with custom pre-processing and post-processing algorithms. The system replaces identifiers (such as references to names, location, and identifying numbers), and shifts exact dates by a time period that is consistent within each record, but differs across records. We selected one month (January 2004) of notes titled as 'Discharge Summary' from the SD. It consisted of total 3510 notes, from which 50 were randomly selected as the test set for evaluation, and the rest were used for development. All developers were blinded to the test set and only had access to the development set. An internal medicine physician manually reviewed notes in the test set and annotated medication information in the text. The same set of notes was also processed by MedEx to generate structured output. The gold standard was generated by reconciling the results from manual annotation and MedEx's output. When there was a conflict between manual annotation and MedEx's output, another physician was consulted to determine the final gold standard. To evaluate the performance of MedEx on outpatient notes, 25 notes titled as 'Clinic visit notes' were randomly selected from the SD within January 2004 and manually reviewed in a similar way.

The structured outputs of MedEx were manually compared to the gold standard described above. Precision (P), Recall (R), and F-measure (F) were calculated for each type of data, where P=TP/(TP+FP), R=TP/(TP+FN), and F=2PR/(P+R), where TP stands for True Positive, FP stands for False Positive, and FN stands for False Negative. If a detected drug name contains strength or form (including SCDC, SCDF, and SCD types of drug names from RxNorm), such as 'TRAZODONE 50MG', both drug name and strength/form will be counted as true, because they can be easily obtained based on the RxNorm knowledge base as described above. Names of drug classes (eg, *Antibiotics*) and vaccines were excluded from this study. A similar evaluation was also done on a set of 25 clinic visit notes.

Table 2 Results of MedEx on 50 discharge summaries

Findings types	Total #	Precision (%)	Recall (%)	F-measure (%)
DrugName	377	95	92	93
Strength	179	99	91	95
Route	182	99	90	94
Frequency	192	99	94	96
Form	39	97	82	89
Dose Amount	36	100	78	88
IntakeTime	23	83	44	57
Duration	22	76	73	74
Dispense Amount	7	100	71	83
Refill	4	100	75	86
Necessity	42	100	83	91

STATUS REPORT

Based on the output of MedEx and expert manual review, the gold standard from the 50 discharge summaries contained 377 medication findings. Table 2 shows the evaluation results on discharge summaries in terms of precision, recall and F-measure. Drug names, strength, route, and frequency, which had enough samples (n>50), reached high F-measures of 93.2%, 94.5%, 93.9%, and 96.0% respectively. Evaluation results using a set of 25 clinic visit notes are shown in table 3. F-measure of drug names, strength, route, and frequency were also over 90%. We did not report results of other categories because they had very low total counts (n<30).

DISCUSSION

In this paper, we introduce an NLP system (MedEx) to extract structured medication information from discharge summaries. Several studies have worked on extracting medication names from clinical text using different approaches, including string matching methods[23] and rule-based methods.[24] A more recent study has focused on extracting both medication names and signatures using a regular expression based approach—the MERKI system.[27] Although it is efficient, the regular expression based approach has limitations when processing complicated medication text that contains multiple signatures and contextual level information such as status or temporal data. MedEx uses a new method to parse medication text, which consists of a sequential tagger and a combined parser. The sequential tagger, which combines lookup, regular expression, and rule-based disambiguation components, provides a robust tagging method, which highly improves the accuracy of semantic labeling of drug names and signatures. The parser that combines a Chart parser and a regular expression Chunker also improves the ability to parsing more complicated medication text. Evaluation showed that MedEx can accurately extract not only drug names, but also medication associated signature information, such as strength, route, and frequency, with high F-measures (93.2%–96.0%). For the task of medication signature information extraction, the

Table 3 Results of MedEx on 25 clinic notes

Findings types	Total #	Precision (%)	Recall (%)	F-measure (%)
DrugName	200	97	88	92
Strength	94	95	95	95
Route	54	96	87	91
Frequency	102	97	89	93

performance of MedEx is superior to systems reported in previous studies (F-measures of 85.3%, 80.3%, and 48.3% on strength, route, and frequency, respectively).[28]

We manually reviewed errors generated by MedEx and analyzed their causes. False negatives were usually caused by terms that were not in our lexicon files or not recognized by our regular expressions. For example, the low recall of IntakeTime was mainly caused by not recognizing terms like '06' (meaning 6am). Our RxNorm-derived lexicon had very good coverage on drug names. Causes of false positive drug names were multifactorial. Some involved ambiguous drug names, such as 'potassium is normal', where 'potassium' refers to a lab test, instead of a drug (supplement). As described previously, we defined rules in the tagging step based on contextual words and some of those errors were eliminated. Another cause of false positives occurs when drug allergy lists include drug names that appear to be current medications (eg, 'ALLERGIES: She is allergic to MORPHINE'). One of the solutions to prevent this type of errors is to use clinical note section header information (eg, 'Allergy' or 'Family history' section). In the future, we plan to integrate an existing section tagger (SecTag)[35] to the pre-processing step of MedEx, to reduce this type of false positives. As noted by Sirohi and Peissig,[29] drug names from databases (eg, First Data Bank's NDDF) could be English words as well. We also observed this in the lexicon derived from RxNorm when processing our training set. For example, 'Vital' was a drug name in RxNorm and it caused many false positives. By removing those English words using the approach described in the Methods section, we eliminated many of this type of errors, with some exceptions such as 'One daily' as a drug name. A more generalizable solution is to develop a disambiguation method that can determine the meaning of an ambiguous drug term (eg, 'Vital' can be a drug name or an adjective in English) based on the context around it. We also noticed some errors were caused by the sentence detection program, which sometimes breaks one medication finding into two sentences.

We applied MedEx to outpatient clinic visit notes without any algorithmic changes. Small drops on recall were observed in drug names, route and frequency, but overall the MedEx approach showed similar performance on discharge summaries and on clinic notes. This suggests potential generalizability on extracting medication information from various types of clinical notes. The performance drop could be due to several reasons. First, clinic visit notes are often typed, in contrast to discharge summaries, which are most often dictated and professionally transcribed. We noticed there were more spelling errors and more abbreviations of drug names in clinic visit notes, which caused the drop in recalls across all categories. Second, discharge summaries usually have a 'Medication' section that often contains semi-structured medication text, which is easier to parse. Last, we did not use a training set of outpatient clinic notes to expand our lexicon and tweak the algorithm.

Currently, we use a set of manually defined rules for disambiguation in the sequential tagger. In the future, we plan to further improve the semantic tagging step by implementing a transformation-based tagger (Brill Tagger[40]), which automatically learns rules from an annotated training set. We will also extend the MedEx by: (1) encoding the extracted medication names using RxNorm; (2) capturing contextual level information such as status of medications.

In this paper, we developed a medication information extraction system (MedEx) for clinical notes and we evaluated its performance using two types of data sets: discharge summaries and clinic visit notes. Results showed that MedEx can extract drug names and signature information such as strength, route, and frequency from discharge summaries and clinic visit notes with over 90% F-measure.

Funding This study was partially supported by grants from the US NIH: NHGRI U01 HG004603 and NLM R01-LM007995-05. The datasets used were obtained from Vanderbilt University Medical Center's Synthetic Derivative which is supported by institutional funding and by the Vanderbilt CTSA grant 1UL1RR024975-01 from NCRR/NIH.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Johnson KB,** Kiepek W. Outpatient e-prescribing at an academic medical center [case study]. In: Mansur JM (ed.). *A guide to the Joint Commission's Medication Management Standards*. Joint Commission Resources (JCR), 2009:120—7.
2. **Vira T,** Colquhoun M, Etchells E. Reconcilable differences: correcting medication errors at hospital admission and discharge. *Qual Saf Health Care* 2006;**15**:122—6.
3. **Moore C,** Wisnivesky J, Williams S, *et al*. Medical errors related to discontinuity of care from an inpatient to an outpatient setting. *J Gen Intern Med* 2003;**18**:646—51.
4. **JCAHO**. 2005 National patient safety goals. http://www.jointcommission.org/patientsafety/nationalpatientsafetygoals/05_npsgs.htm. 2005 (accessed 15 Nov 2009).
5. **Poon EG,** Blumenfeld B, Hamann C, *et al*. Design and implementation of an application and associated services to support interdisciplinary medication reconciliation efforts at an integrated healthcare delivery network. *J Am Med Inform Assoc* 2006;**13**:581—92.
6. **Cimino JJ,** Bright TJ, Li J. Medication reconciliation using natural language processing and controlled terminologies. *Stud Health Technol Inform* 2007;**129**(Pt 1): 679—83.
7. **Sager N,** Friedman C, Chi E, *et al*. The analysis and processing of clinical narrative. *Medinfo* 1986:1101—5.
8. **Sager N,** Friedman C, Lyman M. *Medical language processing: computer management of narrative data*. Reading, MA: Addison-Wesley, 1987.
9. **Friedman C,** Alderson PO, Austin J, *et al*. A general natural language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;**1**:161—74.
10. **Hripcsak G,** Friedman C, Alderson PO, *et al*. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995;**122**:681—8.
11. **Hripcsak G,** Austin JH, Alderson PO, *et al*. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology* 2002;**224**:157—63.
12. **Friedman C.** Towards a comprehensive medical language processing system: methods and issues. *Proc AMIA Annu Fall Symp* 1997:595—9.
13. **Friedman C,** Knirsch C, Shagina L, *et al*. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. *Proc AMIA Symp* 1999:256—60.
14. **Haug PJ,** Koehler S, Lau LM, *et al*. Experience with a mixed semantic/syntactic parser. *Proc Annu Symp Comput Appl Med Care* 1995:284—8.
15. **Haug PJ,** Christensen L, Gundersen M, *et al*. A natural language parsing system for encoding admitting diagnoses. *Proc AMIA Annu Fall Symp* 1997:814—18.
16. **Fiszman M,** Chapman WW, Evans SR, *et al*. Automatic identification of pneumonia related concepts on chest x-ray reports. *Proc AMIA Symp* 1999:67—71.
17. **Fiszman M,** Chapman WW, Aronsky D, *et al*. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc* 2000;**7**:593—604.
18. **Denny JC,** Smithers JD, Miller RA, *et al*. Understanding medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc* 2003;**10**:351—62.
19. **Denny JC,** Miller RA, Waitman LR, *et al*. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. *Int J Med Inform*. Published Online First: 19 Oct 2008. 2009;**78**(Suppl 1):S34—42.
20. **Hahn U,** Romacker M, Schulz S. MEDSYNDIKATE-a natural language system for the extraction of medical information from findings reports. *Int J Med Inform* 2002;**67**:63—74.
21. **Chapman WW,** Fiszman M, Dowling JN, *et al*. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. *Stud Health Technol Inform* 2004;**107**(Pt 1):487—91.
22. **Chapman WW,** Aronsky D, Fiszman M, *et al*. Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department. *Proc AMIA Symp* 2000:131—5.
23. **Ceusters W,** Spyns P, De Moor G. From natural language to formal language: when MultiTALE meets GALEN. Stud. *Health Technol Inform* 1997;**43**(Pt A):396—400.
24. **Zeng QT,** Goryachev S, Weiss S, *et al*. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006;**6**:30.
25. **Evans DA,** Brownlow ND, Hersh WR, *et al*. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proc AMIA Annu Fall Symp* 1996:388—92.

26. **Chhieng D,** Day T, Gordon G, *et al*. Use of natural language programming to extract medication from unstructured electronic medical records. *AMIA Annu Symp Proc* 2007:908.
27. **Levin MA,** Krol M, Doshi AM, *et al*. Extraction and mapping of drug names from free text to a standardized nomenclature. *AMIA Annu Symp Proc* 2007:438—42.
28. **National Library of Medicine**. RxNorm. http://www.nlm.nih.gov/research/umls/rxnorm/ (accessed 15 Nov 2009).
29. **Sirohi E,** Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. *Pac Symp Biocomput* 2005:308—18.
30. **Gold S,** Elhadad N, Zhu X, *et al*. Extracting structured medication event information from discharge summaries. *AMIA Annu Symp Proc* 2008:237—41.
31. **Jagannathan V,** Mullett CJ, Arbogast JG, *et al*. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *Int J Med Inform* 2008;**78**:284—91.
32. **Friedman C,** Huff SM, Hersh WR, *et al*. The Canon Group's effort: working toward a merged model. *J Am Med Inform Assoc* 1995;**2**:4—18.
33. **Sowa J.** *Conceptual structures: information processing in mind and machine*. Reading, MA: Addison-Wesley, 1984.
34. **Denny JC,** Miller RA, Johnson KB, *et al*. Development and evaluation of a clinical note section header terminology. *AMIA Annu Symp Proc* 2008:156—60.
35. **Denny J,** Spickard IA, Johnson K, *et al*. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc*. In press.
36. **SCOWL word list**. http://wordlist.sourceforge.net/ (accessed 15 Nov 2009).
37. **Kay M.** Algorithm schemata and data structures in syntactic processing. In: Allen St (ed.). *Text processing: text analysis and generation, text typology and attribution*. Stockholm, Sweden: Almqvist and Wiksell International, 1982:327—58.
38. **Natural Language Toolkit in Python**. http://www.nltk.org/Home. 2009 (accessed 15 Nov 2009).
39. **Chomsky N.** Three models for the description of language. *IRE Transactions on Information Theory* 1956;**2**:113—23.
40. **Brill E.** Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics* 1995;**21**:543—66.