Case Report ■

Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes

ALEXANDER TURCHIN, MD, MS, NIKHEEL S. KOLATKAR, MD, MPH, RICHARD W. GRANT, MD, MPH, ERIC C. MAKHNI, MERRI L. PENDERGRASS, MD, PHD, JONATHAN S. EINBINDER, MD, MPH

Abstract This case study examined the utility of regular expressions to identify clinical data relevant to the epidemiology of treatment of hypertension. We designed a software tool that employed regular expressions to identify and extract instances of documented blood pressure values and anti-hypertensive treatment intensification from the text of physician notes. We determined sensitivity, specificity and precision of identification of blood pressure values and anti-hypertensive treatment intensification using a gold standard of manual abstraction of 600 notes by two independent reviewers. The software processed 370 Mb of text per hour, and identified elevated blood pressure documented in free text physician notes with sensitivity and specificity of 98%, and precision of 93.2%. Anti-hypertensive treatment intensification was identified with sensitivity 83.8%, specificity of 95.0%, and precision of 85.9%. Regular expressions can be an effective method for focused information extraction tasks related to high-priority disease areas such as hypertension.

■ **J Am Med Inform Assoc.** 2006;13:691–695. DOI 10.1197/jamia.M2078.

Introduction

By some estimates free text physician notes contain over 50% of the data in the patient's medical record.[1] Across the United States healthcare organizations are increasingly moving toward electronic medical record systems.[2] As part of this process, physician notes are frequently becoming available in digital format and thus potentially amenable to computational processing.[3] Information extracted from narrative medical documents has been used for populating structured electronic medical records databases,[4] billing,[5] identification of potential subjects for research studies,[6] and epidemiological research.[7–9]

A number of software tools for extraction of information from narrative medical documents have been described in the literature.[4,5,10–14] At this time, most of the tools developed in the academic settings are not easily available to other researchers, while the commercial ones are costly. Regular expressions—a metalanguage that describes finite-state automata used to recognize string patterns[15]—have been employed in information extraction both in and outside of medicine,[16–18] and could provide an alternative approach to more complex syntactic/semantic parsers.

Regular expressions were first described by Kleene in 1956.[19] Their advantages include speed and ease of use: the tools for interpreting regular expressions already exist in multiple implementations and over the years have been fine-tuned for performance.[20] At the same time regular expression syntax is mostly standard across all implementations and regular expressions developed for one application can usually be transferred to any of the others with minimal modification.[15] The main drawback of regular expressions when compared to the syntactic/semantic parsers is their lack of flexibility. However, medical narrative documents have been shown to be lexically less ambiguous than unrestricted documents.[21] Consequently, regular expressions possess many of the qualities necessary for successful extraction of information from free text medical documents, such as physician notes.

Case Description

In this paper we report on an example of application of regular expression techniques to extraction of blood pressure values and anti-hypertensive medication intensification information from narrative medical documents using regular expressions. Information derived from this application can potentially be used for future epidemiologic studies of

Affiliations of the authors: Division of Endocrinology, Brigham and Women's Hospital, Boston, MA (AT, NSK, MLP); Division of General Medicine, Massachusetts General Hospital, Boston, MA (RWG); Harvard Medical School, Boston, MA (AT, NSK, RWG, ECM, MLP, JSE); Clinical Informatics Research and Development, Partners HealthCare System, Boston, MA (AT, JSE); Division of General Medicine, Brigham and Women's Hospital, Boston, MA (JSE).

The authors would like to express our gratitude to Dr. Maria Shubina at the Center for Clinical Investigations at the Brigham and Women's Hospital for her advice on statistical evaluation.

This research was supported in part by the Partners HealthCare IS Research Council (AT, JSE), Diabetes Trust Foundation (AT), NHLBI training grant T32HL007609 (NSK), and NIDDK Career Development Award K23 DK067452 (RWG).

Correspondence and reprints: Alexander Turchin, MD, MS, Clinical Informatics Research and Development, Partners HealthCare System, 93 Worcester Street, Suite 201, Wellesley, MA 02481; e-mail: <aturchin@partners.org>.

Received for review: 02/07/06; accepted for publication: 08/08/06.

hypertension treatment. Many large-scale epidemiologic investigations are currently carried out using manual review of patient charts, which typically requires the expensive labor of trained professionals over long periods of time. In contrast, information extraction software based on regular expressions can process many thousands of notes per hour,[22] drastically reducing the cost and time required to complete the study.

In order to successfully apply computation information extraction for epidemiologic research it is necessary to show that the technique employed has high accuracy and a significant gain in speed over the manual methods. In this study we therefore evaluated the accuracy and speed of a regular expression-based software tool that abstracts the documented blood pressure values and anti-hypertensive medication regimen intensification from the text of narrative physician notes in the electronic medical record.

Methods

Algorithm

The program used to extract the data from physician notes was implemented in Perl and used extended regular expressions to detect word patterns empirically determined to be specific for the concepts sought. The program identified two sets of concepts: blood pressure values and treatment intensification. Treatment intensification was defined as initiation of a new or an increase in the dose of an existing anti-hypertensive medication (the definition used in previously reported studies on the subject[23]). Substitutions of one anti-hypertensive medication for another were included; decreases in the dose of an existing anti-hypertensive medication were excluded.[24] The documented blood pressure values were identified using the following algorithm:

1. A blood pressure "tag"—a word pattern/phrase whose presence indicates that the remainder of the sentence is likely to contain a blood pressure value—was identified using 23 regular expressions.
2. The remainder of the sentence where a blood pressure tag was found was then analyzed for presence of a blood pressure value using four regular expressions that detect both blood pressure ranges as well as single measurements.
3. The blood pressure values identified in the sentence were then checked for possible errors and were discarded if:
   a) diastolic blood pressure is greater than systolic;
   b) systolic blood pressure is greater than 300 or less than 50;
   c) diastolic blood pressure is greater than 200 or less than 20.

Documentation of anti-hypertensive treatment intensification was identified using the following algorithm:

1. Sentences that contained one of 204 names and abbreviations of anti-hypertensive medications and medication classes were identified.
2. These sentences were subsequently analyzed for the presence of one of the patterns empirically found to identify documentation of initiation of a new medication or increase in the dose of an existing medication using 24 regular expressions.
3. The sentences with documentation of medication initiation or a dose increase were then checked for the presence of one of the patterns that indicated that the treatment intensification was conditional rather than definitive (only documentations of definitive anti-hypertensive treatment intensification were recorded).
4. If a sentence that contained the name of an anti-hypertensive medication did not contain a pattern that indicated that this medication was initiated or its dose increased, the next sentence in the note was analyzed to determine whether it contained documentation for medication initiation or a dose increase referring to the medication mentioned in the previous sentence.

The actual regular expressions used to detect blood pressure and anti-hypertensive treatment intensification can be found in Appendices 1 and 2, respectively (both available as JAMIA online supplements at www.jamia.org). If more than one blood pressure value was documented in the note, the blood pressure value with the lowest mean arterial blood pressure (diastolic blood pressure + one-third of the difference between systolic and diastolic blood pressure) was recorded.

Evaluation

We assessed the accuracy of extraction of blood pressure values documented in the text of the note and of identification of anti-hypertensive medication intensification. Two non-overlapping sets of 300 primary care physician notes randomly selected from the electronic medical record of two academic medical centers were used to evaluate each of these outcomes. The notes in the medication intensification set were randomly selected from the notes with documented elevated blood pressure (identified using the technique validated in the first phase of the evaluation). Elevated blood pressure was defined as either systolic blood pressure above 129 mm Hg, or diastolic above 84 mm Hg, in accordance with the guidelines published prior to the beginning of the study period.[25]

For the first phase of the evaluation, blood pressure values were manually abstracted from each of the notes by two independent reviewers who did not participate in the design of the software and did not know the word patterns that the software identified. The reviewers' results were then compared and inter-reviewer consensus was established after joint review of the notes for which the original abstractions differed. This consensus was subsequently used as the gold standard to which the software results were compared to determine sensitivity, specificity, and overall agreement of automatic extraction of documented blood pressure values. For the second phase of the evaluation, a similar procedure was followed to establish inter-reviewer consensus between manual abstractions of anti-hypertensive treatment intensification by two independent reviewers. This consensus was subsequently used as the gold standard to which the software results were compared to determine sensitivity, specificity and overall agreement of automatic extraction of documented intensification of anti-hypertensive medication regimen.

This study was approved by the Partners HealthCare Human Research Committee.

Table 1 ▪ Agreement between Automatic and Manual Concept Extraction

Measure	Overall Agreement	Sensitivity	Specificity	Positive Predictive Value
Numeric value of documented blood pressure	93.0% (± 2.9%)	90.6% (± 3.3%)	96.1% (± 2.2%)	N/A
Documentation of elevated blood pressure (Boolean)	98.3% (± 1.5%)	98.2% (± 1.5%)	98.4% (± 1.4%)	93.2% (± 6.4%)
Documentation of anti-hypertensive treatment intensification (Boolean)	92.0% (± 3.1%)	83.8% (± 4.2%)	95.0% (± 2.5%)	85.9% (± 7.7%)

95% confidence interval is indicated in parentheses.

Example

Inter-reviewer agreement for manual abstraction of the numeric value of documented blood pressure from 300 randomly selected physician notes was 94.0% with kappa of 0.94. Inter-reviewer agreement for manual abstraction of documentation of anti-hypertensive treatment intensification from a second set of 300 randomly selected physician notes that documented elevated blood pressure was 91.7% with kappa of 0.79.

The same 600 physician notes were subsequently processed by the software and the results were compared to the inter-reviewer manual abstraction consensus. The software processed over 370 Mb of text per hour. Sensitivity (recall), specificity, and positive predictive value (precision) were calculated for identification of elevated blood pressure, numeric value of the blood pressure documented in the note, and documentation of anti-hypertensive treatment intensification. Sensitivity of the automated data extraction ranged from 83.8% for treatment intensification to 98.2% for documentation of elevated blood pressure, and specificity ranged from 95.0% for treatment intensification to 98.4% for documentation of elevated blood pressure (Table 1). Positive predictive value of the automated data extraction ranged from 85.9% for treatment intensification to 93.2% for documentation of elevated blood pressure (Table 1).

Examples of the word patterns correctly and incorrectly identified by the software are given in Table 2. For blood pressure identification most common false negatives were encountered in sentences that only documented the systolic but not the diastolic blood pressure, and most common false positives were blood pressure values that did not represent the patient's blood pressure (e.g., goal blood pressure). Many of the false positive identifications of treatment intensifications were due to missed conditionals or references to the past, while most of the false negatives were caused by word patterns not captured by the set of regular expressions used.

Discussion

In this study, we present the evaluation of utility of using regular expressions for computational extraction of blood pressure and anti-hypertensive medication intensification information from narrative physician notes. The software achieved accuracy rates comparable to the rates of agreement between human abstractors for all categories of information it extracted, while processing data at speeds several orders of magnitude higher. This approach could make possible the use of the software in large-scale epidemiologic studies where it could potentially replace months of manual work by many highly trained human abstractors.

Our software used a set of regular expressions to accomplish the task. This approach has both advantages and disadvantages. Its obvious limitation is the lack of generalizability: a new set of regular expressions has to be developed and validated for each particular task. Applications of regular expressions are also limited to the extraction of data items that have a constrained lexical scope, and complex synonyms have to be manually generated. On the other hand, a set of regular expressions can be developed much faster than a full-fledged natural language processing engine. This is particularly important because most of the academic natural language processing engines are not publicly available, and the commercial ones frequently bear a price tag unaffordable to researchers. The only freely available natural language processing engine for medical documents—MetaMap—has significant limitations, including lack of implementation of negations.[4]

Both benefits and shortcomings of using regular expressions for information extraction were well illustrated in our study. The software tool was designed, implemented and validated over the period of only six months—a significant advantage

Table 2A ▪ Examples of Word Patterns Correctly and Incorrectly Identified in the Text of Physician Notes to Detect Documentation of Blood Pressure

True Positives	False Positives	True Negatives	False Negatives
Patient's blood pressure was measured to be 170/105 and on repeat 160/95	BP – increase Zestril to 20 mg qd with the goal of 130/80.	Blood pressure was last measured on 12/05.	In addition, his SBP was 70 in the office.
BP was 135/70			Blood pressure by palpation 130.
Vitals: 68 150/90 180 lbs			
P.E. 120/80			
120/70, 68, wt 190			
BP is 130-140/70-80			

Table 2B ■ Examples of Word Patterns Correctly and Incorrectly Identified in the Text of Physician Notes to Detect Documentation of Treatment Intensification

True Positives	False Positives	True Negatives	False Negatives
Will try increasing her lisinopril	She believes she started the Cozaar I prescribed to her in mid-September	If his BP doesn't come down by next visit, would recommend starting minoxidil	Rx given for lisinopril 10 mg po qd
Need to start her back on glyburide and low-dose captopril	Will increase lisinopril if elevated at next visit	Likely will recommend labetalol next time	Again adding 25 mg of HCTZ to the regimen
She currently takes losartan 50 mg daily. Will increase to 100 mg.			high on diovan 160-incr to 320
He will switch from 0.1 to 0.2 mg of clonidine			
We are going to change it to atenolol			

over syntactic and semantic parsing systems that frequently takes years to develop. The software achieved high accuracy rates while maintaining the speed of data processing necessary for handling large data sets.

Conversely, the accuracy of the information extraction by the software, while high, was not perfect. At the point where the design and implementation of the software were completed, there remained a number of patterns that the software misinterpreted either as false positives or false negatives. Typically these patterns occurred rarely (e.g., once or twice in the entire dataset of several thousand documents) and it was therefore not practical to design additional regular expressions to capture them. Concepts whose expression in the narrative medical documents is less lexically constrained than the ones we chose to study may not be suitable for this technique.

In conclusion, our case study demonstrates that regular expressions can be effectively used to extract focused information from narrative medical documents. When general purpose NLP software is not available, regular expressions provide an alternative approach for abstraction of lexically constrained data elements that can be quickly designed and validated and can potentially be used in a number of clinical applications.

References ■

1. Hicks J. The potential of claims data to support the measurement of health care quality. [PhD]. San Diego, CA, RAND; 2003.
2. de Lusignan S, Teasdale S, Little D, et al. Comprehensive computerised primary care records are an essential component of any national health information strategy: report from an international consensus conference. Inform Prim Care. 2004; 12(4):255–64.
3. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Fam Pract. 2006;23(2):253–63.
4. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. J Biomed Inform. [Epub ahead of print.] Dec 5 2005.
5. Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc. 2003:420–4.
6. Embi PJ, Jain A, Clark J, Bizjack S, Hornung R, Harris CM. Effect of a clinical trial alert system on physician participation in trial recruitment. Arch Intern Med. 2005;165(19):2272–7.
7. Hazlehurst B, Sittig DF, Stevens VJ, et al. Natural language processing in the electronic medical record assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med. 2005;29(5):434–9.
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(3):254–66.
9. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005;12(4):448–57.
10. Aronsky D, Kasworm E, Jacobson JA, Haug PJ, Dean NC. Electronic screening of dictated reports to identify patients with do-not-resuscitate status. J Am Med Inform Assoc. 2004;11(5): 403–9.
11. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med. 1998;37(1):1–7.
12. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc. 2005;12(5):517–29.
13. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. J Am Med Inform Assoc. 2000;7(6):593–604.
14. Sequist TD, Gandhi TK, Karson AS, et al. A randomized trial of electronic clinical reminders to improve quality of care for diabetes and coronary artery disease. J Am Med Inform Assoc. 2005;12(4):431–7.
15. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Upper Saddle River, NJ: Prentice Hall, 2000.
16. Grishman R. Information Extraction: Techniques and Challenges. Paper presented at: International Summer School SCIE-97, 1997.
17. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc. 2001;8(6):598–609.
18. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34(5): 301–10.
19. Kleene SC. Representation of events in nerve nets and finite automata. In: Shannon C, McCarthy J, Automata Studies. Princeton, NJ: Princeton University Press, 1956:3–41.
20. Friedl JEF. Mastering Regular Expressions. 2nd ed. Sebastopol, CA: O'Reilly and Associates; 2002.
21. Ruch P, Baud R, Geissbuhler A, Rassinoux AM. Comparing general and medical texts for information retrieval based on natural language processing: an inquiry into lexical disambiguation. Medinfo. 2001;10(Pt 1):261–5.

22. Turchin A, Pendergrass ML, Kohane IS. DITTO—a tool for identification of patient cohorts from the text of physician notes in the electronic medical record. AMIA Annu Symp Proc. 2005:744–8.
23. Berlowitz DR, Ash AS, Hickey EC, et al. Inadequate management of blood pressure in a hypertensive population. N Engl J Med. 1998;339(27):1957–63.
24. Okonofua EC, Simpson KN, Jesri A, Rehman SU, Durkalski VL, Egan BM. Therapeutic inertia is an impediment to achieving the Healthy People 2010 blood pressure control goals. Hypertension 2006;47(3):345–51.
25. The sixth report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. Arch Intern Med. 1997;157(21):2413–46.