# Medication Extraction from Electronic Clinical Notes in an Integrated Health System: A Study on Aspirin Use in Patients with Nonvalvular Atrial Fibrillation

Chengyi Zheng, PhD[1]; Nazia Rashid, PharmD[2]; River Koblick, BA[2]; and JaeJin An, PhD[3]

[1]*Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, California;* [2]*Drug Information Services, Kaiser Permanente Southern California, Downey, California; and* [3]*Department of Pharmacy Practice and Administration, Western University of Health Sciences, Pomona, California*

## ABSTRACT

**Purpose:** The purpose of this study was to investigate whether aspirin use can be captured from the clinical notes in a nonvalvular atrial fibrillation population.

**Methods:** A total of 29,507 patients with newly diagnosed nonvalvular atrial fibrillation were identified from January 1, 2006, through December 31, 2011, and were followed up through December 31, 2012. More than 3 million clinical notes were retrieved from electronic medical records. A training data set of 2949 notes was created to develop a computer-based method to automatically extract aspirin use status and dosage information using natural language processing (NLP). A gold standard data set of 5339 notes was created using a blinded manual review. NLP results were validated against the gold standard data set. The aspirin data from the structured medication databases were also compared with the results from NLP. Positive and negative predictive values, along with sensitivity and specificity, were calculated.

**Findings:** NLP achieved 95.5% sensitivity and 98.9% specificity when compared with the gold standard data set. The positive predictive value was 93.0%, and the negative predictive value was 99.3%. NLP identified aspirin use for 83.8% of the study population, and 70% of the low dose aspirin use was identified only by the NLP method.

**Implications:** We developed and validated an NLP method specifically designed to identify low dose aspirin use status from the clinical notes with high accuracy. This method can be a valuable tool to supplement existing structured medication data. (*Clin Ther*. 2015;37:2048–2058) © 2015 Elsevier HS Journals, Inc. All rights reserved.

**Key words:** aspirin, electronic clinical notes, electronic medical record, integrated health systems, medication status, natural language processing.

## INTRODUCTION

Atrial fibrillation (AF) is the most common cardiac arrhythmia, affecting between 2.7 million and 6.1 million people in the United States, and its prevalence is expected to increase in the next few decades.[1] Most AF patients develop nonvalvular atrial fibrillation (NVAF), which is associated with a 5-fold increase in risk of ischemic stroke.[2] One of the main goals of NVAF treatment is to prevent stroke with the use of antithrombotic therapy. Clinical guidelines recommend either oral anticoagulant or antiplatelet therapy based on a patient's risk of stroke (eg, $CHA_2DS_2$-VASc score) calculated using clinical factors.[1] However, there may be high-risk patients who take aspirin instead of anticoagulants in a real-world setting because of reasons such as high risk of bleed, falls, or patient preference. Patients may also be taking both an anticoagulant and antiplatelet agent at the same time. Understanding the real-world treatment patterns and outcomes are critical; however, often this is challenging because of data limitations.

Typically, clinicians and researchers access patients' medication history using "structured" medication

*Accepted for publication July 6, 2015.*
http://dx.doi.org/10.1016/j.clinthera.2015.07.002
0149-2918/$ - see front matter

databases. Most antithrombotic drugs are dispensed prescriptions and can be captured in structured medication databases. However, not all medications are available in the medication databases, such as over-the-counter (OTC) medications. Aspirin is an OTC drug commonly used as an antiplatelet therapy in patients with NVAF. One study estimated that 50 million people in the United States (36% of the adult population) are taking daily aspirin solely for cardiovascular disease and stroke prevention.[3] It has been challenging to capture OTC aspirin use in a systematic way other than survey questionnaires.

One potential method for capturing OTC aspirin use may be collecting information from electronic medical record (EMR) systems. Large amounts of medication data are available to both clinical care and research in EMRs. Aspirin use may be documented by health care professionals in the free text clinical notes, within the EMR system. Manual review of clinical notes or medical records can identify aspirin information; however, this can be expensive and time-consuming.

There has been a growing interest in natural language processing (NLP). NLP is a field of computer science and linguistics that aims to understand human (natural) languages. This technique has been used to identify and extract information from the free-text formatted data. Rule-based and statistical machine learning methods are often used together to deliver a robust system.[4] Compared with human review of medical records, NLP is more efficient and consistent.[4] In recent years, NLP has gained wider adoption in the biomedical field.[5]

Understanding the use of both anticoagulants and antiplatelet therapy is important for NVAF patients, but capturing aspirin use has been extremely difficult compared with other prescription anticoagulant or antiplatelet agents. The purpose of this study was to determine aspirin use by applying an NLP algorithm. In this study, we developed and applied an NLP method on a large NVAF population. We identified whether a patient was undergoing aspirin therapy and then determined the status of aspirin use (on/off aspirin therapy) and the aspirin dosage information. We also evaluated the contribution of NLP on reducing missing aspirin data in our current structured medication databases.

## MATERIALS AND METHODS

### Study Setting

Kaiser Permanente Southern California (KPSC) provides integrated, comprehensive medical services to 3.9 million members through its own facilities. Every member receives a medical record number, which allows the member to be linked to various clinical and administrative databases, such as enrollment, drug benefits, medical services and visits, laboratory results, and pharmacy services. The aspects of care and interactions within this integrated care system were captured in an EMR system (Epic Systems, Epic Systems Corporation, Madison, Wisconsin, US), which is available for research purposes.

### Study Population

The study population included all KPSC members aged $\geq$18 years with a new diagnosis of atrial fibrillation ($\geq$2 serial *International Classification of Diseases, Ninth Revision Clinical Modification* [ICD-9-CM] codes of 427.31, $\geq$30 days apart) from January 1, 2006, through December 31, 2011. The first diagnosis date is the index date. Patients had to have 12 months of continuous membership eligibility before the index date (Figure 1). Patients were followed up until disenrollment, outcomes of interest, death, or December 31, 2012, whichever occurred first. Patients who had prior anticoagulation prescriptions or an anticoagulation clinic visit were excluded. Additional exclusion criteria were applied to deliver the final incident NVAF cohort (n = 29,507). The institutional review board at KPSC approved this study.

### Data Collection

After the final cohort was created, all clinical notes from the index date through the end of follow-up date for each patient were extracted. All electronic clinical notes related to the study cohort (n = 29,507) were retrieved from our EMR system and available for this study. A total of 3,235,393 notes were retrieved for 29,425 patients (99.7% of the study cohort).

### NLP Training and Evaluation Data

Training and gold standard data sets were created from 2 different random samples to evaluate aspirin use status and aspirin dosage. A random sample of 2949 notes was selected for the training data set. These notes were used to refine the NLP method. A second random sample of 5339 notes was selected

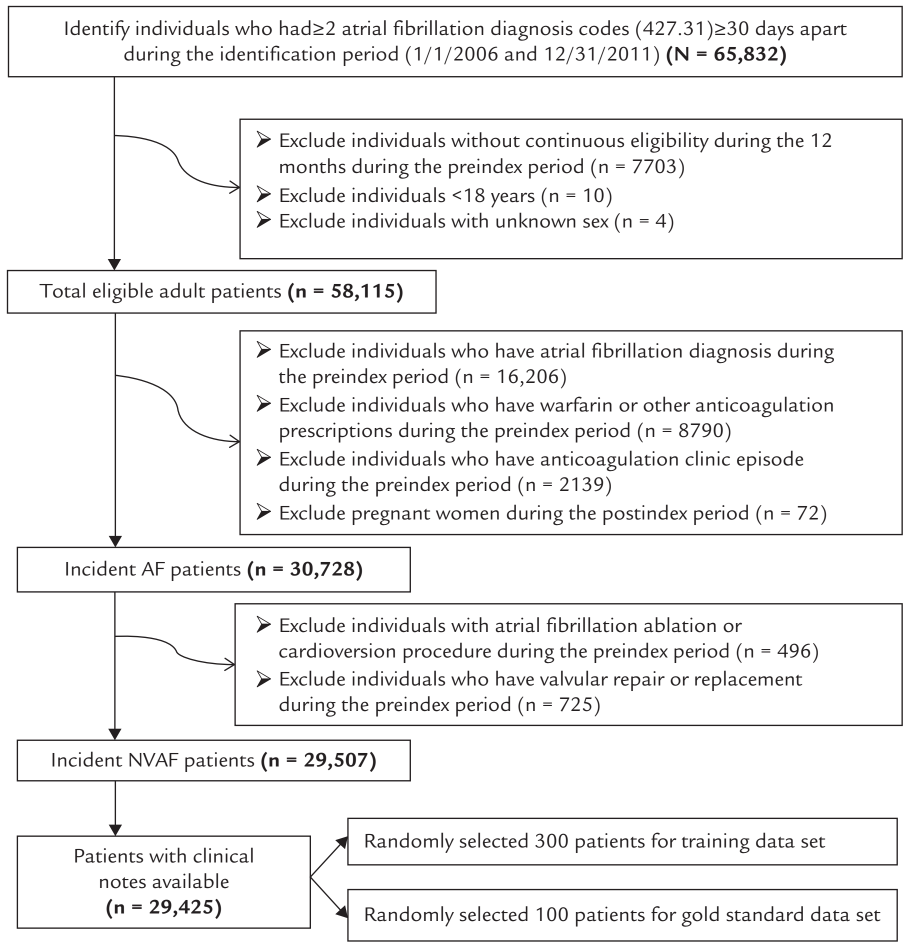


Figure 1. Study population diagram. AF = atrial fibrillation.

for the evaluation data set (gold standard). Notes from the evaluation data set were blinded and manually reviewed by 2 study investigators (N. Rashid and R. Koblick); aspirin-use status and dosage information was documented. At the end of the review, the results were compared between the 2 study investigators. Discrepancies were identified and discussed with the cardiologist, and a consensus was used as the final result for training and evaluation.

## Structured Medication Data

Two structured medication data sources, the prescription (Rx) table and the current medication (CM) table, were included in the analysis. The Rx table captured all medications dispensed by KPSC pharmacies, including OTC medications. The CM table captured medications a patient was taking at the time of each patient visit. These structured databases were used as comparators for the NLP data to determine the contribution of aspirin information from the NLP system.

## Definition of Low- to Medium-Dose Aspirin on/Off Therapy

### *Aspirin Use Status*

The aspirin status for a patient could change throughout the study period, providing periods of "on" and "off" aspirin therapy. Although the aspirin status can be described as a continuous time event, the on/off status captured by NLP is a discrete time point defined by the date of the clinical note.

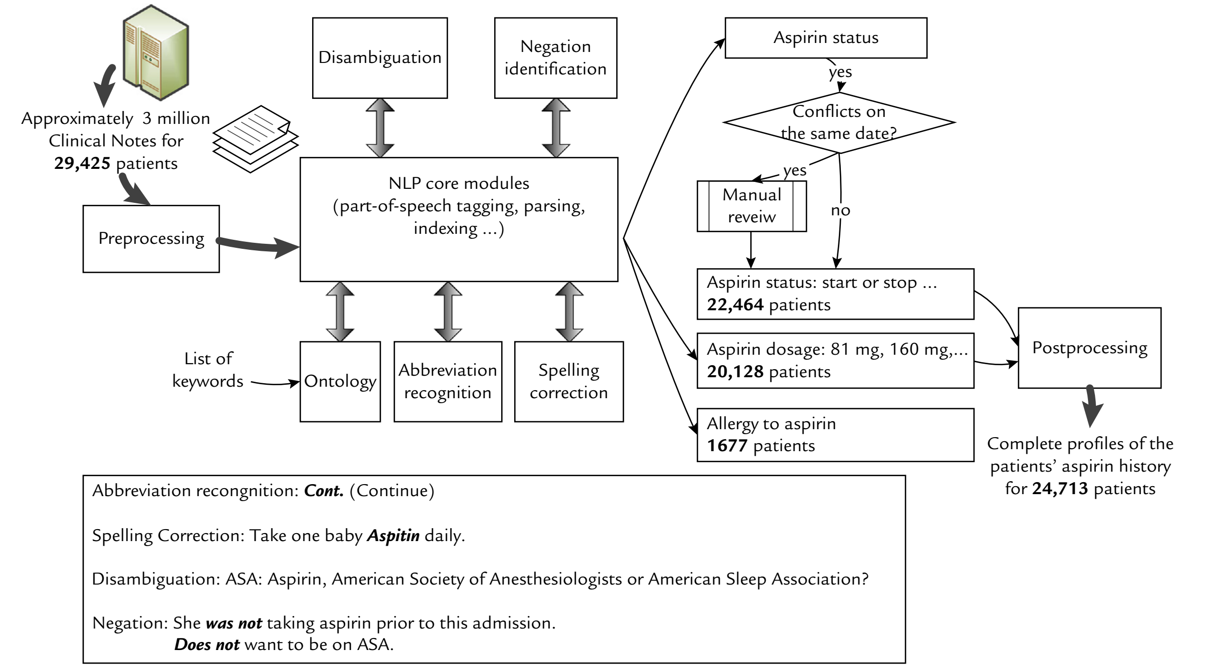


Figure 2. Overview of natural language processing (NLP) on identifying aspirin use data from clinical notes.

### *Aspirin Dosage Criteria*

For stroke prevention, NVAF patients typically use aspirin at low- to medium-dose regimens.[6] Aspirin doses of 75 to 325 mg were included in this analysis because these are the standard dose regimens for stroke prevention in NVAF patients. Higher doses (500–1500 mg) were excluded in this analysis. Aspirin was often used with caffeine and oxycodone in combination drugs as pain reliever. Therefore, those combination drugs were not included in this analysis.

## NLP Algorithms

We developed NLP algorithms for this study based on our existing NLP system.[7–9] The training data set used to develop the algorithm was composed of 2949 notes. A typical NLP system has many modules that process the document in a pipeline fashion. An overview of our NLP system for this study is provided in Figure 2. The underlying NLP core modules include the sentence splitting, tokenization, part-of-speech tagging, parsing, and indexing. The auxiliary modules include the word sense disambiguation, negation identification, abbreviation recognition, and spelling correction. The main steps and some essential processes are highlighted in the following sections.

### *Building Medication Ontologies*

The following ontologies were created: aspirin name ontology (see appendix in the online version at http://dx.doi.org/10.1016/j.clinthera.2015.07.002), medication modifiers (buffered, coated, etc), dosage (amount with or without the unit), duration and frequency (how long and often administrated, such as every other day), status (on, off, resume, etc), form (tablet, pill, drip, etc), mode (route of administration, such as oral, intravenous, sublingual), and reason (medical indications, such as headache or pain).

### *Preprocessing*

The preprocessing step prepared the text for later NLP steps. This step includes (1) removal of markup tags, such as any HTML (HyperText Markup Language) or XML (Extensible Markup Language) tags; and (2) sentence splitting, which breaks up long text into individual sentences.

### *Section detection*

Clinical notes are often organized into sections. Common sections include medical history, encounters, orders, and prescriptions. Each section has its own medical context implications that can be used in NLP searches. For example, *family history*, *social history*, *past medical history*, and *history of present illness* all refer to a patient's history, but with different meanings. Each section often has a section head that is easily distinguished in its original format but hard to identify in the plain text formatted notes with which we were dealing. We applied an algorithm that automatically identifies the section head and segments the text into different sections. Any remaining text was treated as nonsection text.

#### *NLP indexing*

In this step, notes were indexed for NLP searching. Title and sections identified in the previous step were recorded in the index. We applied linguistic processing before indexing, including assigning each word a grammatical tag, and analyzed the syntactic structure of sentences.

#### *NLP searching*

Medication use was identified using NLP. Simple keyword searches did not have high specificity on identifying aspirin use. Additional algorithms were applied to increase overall accuracy.

*Disambiguate medication name*

As a commonly used abbreviation of aspirin, ASA is also an acronym or abbreviation for other medical terms, such as American Society of Anesthesiologists. We designed an algorithm to disambiguate the use of ASA and ignore all uses with a nonaspirin meaning.

*Identify medication allergy*

We identified patients who were allergic to aspirin. However, we did not use the allergy data in the analysis because these data were not a direct indication of patient's aspirin use or status.

*Exclude history information*

We were only interested in the current medication use information. Aspirin information that appeared in the history sections or documented as a prior history was ignored.

*Exclude other Uses*

Aspirin has many indications. Beside its use as an antiplatelet medication in low doses, aspirin is commonly used to treat pain, fever, and inflammatory conditions in higher doses. We were only interested in its antiplatelet use but not the latter indications. We excluded the following use cases:

1. Indications other than antiplatelet, such as for headache or pain relief.
2. Instructions were not for long-term use, such as for as needed (PRN) use.
3. Aspirin dosage >325 mg/d.
4. Aspirin taken in an intravenous format, which is mainly used in treating pain.

### Identify Medication Status Using NLP

Medication status can be documented in medication-related sections or other parts of the notes. Different NLP search strategies were applied to these 2 types of text.

#### *Identify Medication Status from Medication Sections*

Each section had a section heading and text. When a section heading had medication-related terms, such as *medication* or *prescription*, we treated them as the medication section. Medications mentioned in the section text were often not written as a complete sentence and lacked a corresponding verb. Aspirin status was determined by a combination of searching both the section head and text.

#### *Identify Medication Status from Text Other Than the Medication Section*

Aspirin status appearing in text other than the medication section was confirmed by identifying specific terms in the clinical notes. Status indication terms included verbs such as *add* and *administer*, and their different tenses; adjectives such as *on* and *off*; and phrases such as *stay with* and *resumption of* (Table I). We searched for the status indication terms before and after the aspirin terms. Medication status could be deduced from statements such as "aspirin is sufficient" and "increase the dosage to 325 mg." Depending on the grammatical tense, the medication status could be a past, current, or future statement. Negation identification was applied at the same time to avoid search errors for cases such as "was not taking aspirin."

Table I. Sample keywords for the medication status.

| Status Category | Type of Key words | Examples |
|---|---|---|
| On | Verb | add, administer, begin, commence, continue |
| Off | Verb | avoid, discontinue, hold |
| On | Phrase | change to, initiation of, resumption of |
| Off | Phrase | avoidance of, leaving off, refrain from |
| On | Adjective | on, satisfied, tolerant, tolerable |
| Off | Adjective | off, unsatisfied, intolerant, intolerable, untolerable |
| Reduce | | cut, decrease, reduce |
| Change | | change, change from, switch, switch to |

### Identifying Medication Dosages Using NLP

We searched for both explicit and implicit dosage amounts: (1) explicit dosage amount, such as 81 mg or asa 325; and (2) implicit dosage amount, such as asa baby dose or low-strength asa (see appendix).

### Postprocessing

Note-level search results were combined based on the patient medical record number (MRN) and note date. Each record had MRN, note date, medication name, dosage, and status information. If there were different results for the same MRN and note date, the results were selected for manual review. A series of such records for each patient allowed a longitudinal overview of their aspirin use history.

### Implementation

The NLP algorithm was applied to all the 3,253,393 notes.

### Analysis Method

We evaluated how well NLP identified aspirin data based on a single note (note level). NLP results were validated against the manually reviewed gold standard data set. The numbers of true-positive, false-positive, true-negative, and false-negative results were calculated. Sensitivity, specificity, positive predictive value, and negative predictive value were then derived based on those numbers.

We also compared NLP results with the structured medication data. We measured the volume of patients identified by NLP and the reduction of missing information on the complete study population by comparing to the structured data source (CM and Rx tables).

## RESULTS

### Evaluation of NLP Performance Using the Gold Standard

A 2 × 2 table (Table II) summarizes the true-positive, false-positive, true-negative, and false-negative results by comparing NLP results with the gold standard. NLP achieved 95.5% sensitivity and 98.9% specificity when compared with the gold standard data set (n = 5339). The positive predictive value was 93.0%, and the negative predictive value was 99.3%. Positive and negative likelihood ratios were 82.9 and 0.05, respectively (Table III).

### Analysis of NLP Results for the Complete Study Population

The mean (SD) and median follow-up periods were 2.94 (1.71) and 2.84 years, respectively, for the study

Table II. Note-level performance of NLP on aspirin status identification based on the gold standard (n = 5339).

| | Gold standard | | |
|---|---|---|---|
| 2 × 2 table | Positive | Negative | Total |
| NLP positive | 705 | 53 | 758 |
| NLP negative | 33 | 4548 | 4581 |
| Total | 738 | 4601 | 5339 |

NLP = natural language processing.

Table III. Sensitivity, specificity, PPV, NPV, and LRs for NLP.

| Sensitivity, % (95% CI) | Specificity, % (95% CI) | PPV, % (95% CI) | NPV, % (95% CI) | Positive LR (95% CI) | Negative LR (95% CI) |
|---|---|---|---|---|---|
| 95.5 (93.7–96.9) | 98.9 (98.5–99.1) | 93.0 (90.9–94.7) | 99.3 (99.0–99.5) | 82.9 (63.4–108.4) | 0.05 (0.03–0.06) |

PPV = positive predictive value; NPV = negative predictive value; LR = likelihood ratio; NLP = natural language processing.

population (n = 29,507). NLP identified 203,697 numerical dosage mentions with 68.5% low-dose (75–150 mg/d) regimens and 31.5% medium-dose (160–325 mg/d) regimens from the 3.3 million notes. There were a total of 156,356 on/off statuses identified, with 92.4% having an on status and 7.6% having an off status. On the patient level, numerical dosage was identified in 22,464 patients (76.3%), and aspirin status was explicitly documented in 20,128 patients (68.2%) (Figure 2). Overall, there were 226,572 aspirin note-level results identified for 24,713 patients (83.8% of the study population). The data indicated that most of our study population had tried aspirin at least once during our study period (the mean [SD] and median days from the index date to first aspirin therapy date are 243 [392] and 36 days, respectively). NLP also identified 1677 patients (5.7%) with a documented allergy to aspirin.

### Comparing NLP Results with Structured Medication Data

There were 28,471 patients (96.5% of the study population) identified with aspirin use on combining the results of NLP and the structured medication tables. NLP identified 86.8%, and structured medication tables identified 30.2%. Of the 28,471 patients, 69.8% were identified only by NLP and 13.2% were identified only by the structured medication tables. Depending on the location of text in the clinical notes, the NLP results were separated into 2 groups: the nonmedication section and medication section. The structured medication tables were separated into the CM table and the Rx table. Within the 4 data sources, the NLP had much larger contribution (77.9% NLP (a) and 57.2% (NLP (b)) than the structured medication tables (23.6% (CM) and 8.6% (Rx)) (Figure 3). When comparing their unique contribution to the identification of aspirin patients, NLP also had much a greater contribution (24.3% (NLP (a)) and 7.5% (NLP (b)) than the structured medication tables (10% (CM) and 2.4% (Rx)).

## DISCUSSION

In this study, we developed and applied an NLP method to a large NVAF population without any restriction on the type of clinical notes. We evaluated our method on a manual annotated large data set (5339 notes) and then further applied it to 3.3 million notes. The results of this study indicate the viability of using NLP to identify and extract aspirin profiles from clinical notes. We also found that clinical notes are a high-quality data source for medication use. NLP identified that 96.5% of patients in our study population had aspirin use and 5.7% of patients were allergic to aspirin in our study population.

With an integrated EMR system, we were able to compare the completeness of aspirin data obtained via NLP with the information collected through pharmacy systems and physician-entered current medications. NLP captured far more aspirin use data (86.8% aspirin users identified) when compared with data from the structured medication data sources (30.2% aspirin users identified) in our EMR. This finding indicated that NLP can identify large amounts of missing aspirin data from clinical notes compared with existing pharmacy tables.

Dosage is another important component of medication use data. In this study, NLP identified dosage data: 68.5% were taking a low dose (75–150 mg/d), and 31.5% were taking a medium dose (160–325 mg/d). It appears that our dose distribution is similar to the US population, with slightly more patients taking low-dose aspirin.[3]

NLP has previously been used to extract medication information from unstructured data. One example is the Informatics for Integrating Biology and the Bedside medication extraction challenge,[10–13] which

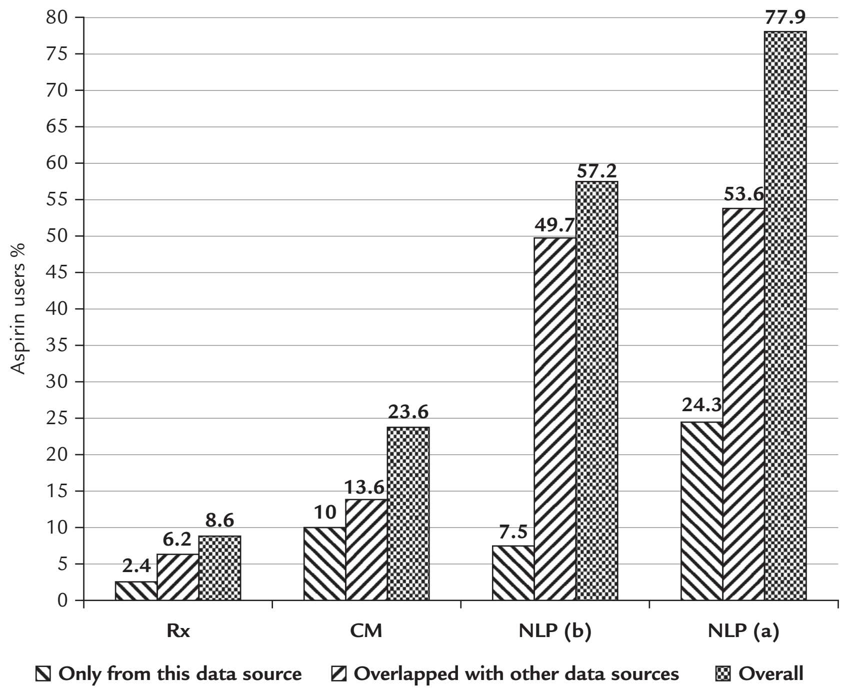


Figure 3. Contributions of different data sources on identifying patients who used aspirin. Aspirin users were identified from the free text formatted clinical notes: Aspirin users identified from the nonmedication section in the clinical notes by natural language processing (NLP [a]), aspirin users identified from the medication section in the clinical notes by natural language processing (NLP [b]), aspirin users identified from the current medication (CM) table, and aspirin users identified from the prescription (Rx) table.

tested the identification of medication, dosage, route, frequency, duration, and reason for administration on approximately 500 discharge summaries. In addition, there have been prior efforts to identify medication status from the clinical notes. Sohn et al[14] developed a rule-based system and a machine learning classifier to automatically classify the medication status using the indication words in the clinical notes. Liu et al[15] developed an NLP and machine learning combined system to determine warfarin exposure at hospital admissions, with 79% sensitivity and 87% specificity. Pakhomov et al[16] developed a rule-based system using regular expressions to identify aspirin exposure and contraindication information for 499 patients with type 2 diabetes and achieved high sensitivity and good specificity. Their study, however, captured all aspirin exposure and did not differentiate between on or off status or the specific dosage. Overall, these prior studies were developed and tested on small numbers of patients or restricted type of clinical notes. Our study did not put any restriction on the type of clinical notes we searched. The clinical notes were from various data sources with heterogeneous characteristics. We developed and applied the NLP method on a large NVAF cohort. The NLP results were then used in other studies to evaluate the clinical outcomes among patients with NVAF and to understand patient characteristics and other factors associated with their antithrombotic treatment patterns.

There are >100,000 OTC medications,[17] and these numbers are increasing as more prescription medications are reclassified as OTC.[18] OTC medications are widely used, and their concurrent use with prescription medications put people at risk for drug-drug interactions.[19] However, OTC medications are not typically captured by the medication system, and their accuracy is even worse than the prescription medications. Because aspirin is mostly OTC, its data are usually not captured in the prescription table. Our pharmacy database is not able to capture many of the aspirin data because most of our members purchased it from outside pharmacies because of its wide availability and low price. The CM table is supposed to be updated by the health care professional during patient encounters; however, our findings concur with what was previously published—data are often incomplete or not updated.[20,21] This is a gap we are trying to address internally as a good practice. We also decided to complete this NLP project so we could take our findings internally and spread changes to reduce existing gaps in current documentation. Our results indicate that structured medication databases (CM table and the Rx table) did not capture the OTC medication well. Before we fill this gap with better documentation, NLP could be used to identify OTC medication use.

Medication discrepancies are prevalent throughout patient care, and up to 67% of cases have at least 1 prescription medication–history error.[22] Medication adherence is another common problem when patients do not adhere to their recommended medication regimen.[23] In this study, we used NLP to identify the medication dosage, on/off status, and allergy history. Integrated with the structured medication data sources, the information identified by NLP could be used to identify medication discrepancies[24] and medication nonadherence. Aspirin use information as identified by the NLP method could also be used to measure the adherence with current guidelines, to support clinical decisions, and to assist perioperative antithrombotic management.

This study has several limitations. First, although it included a large sample of patients and different types of clinical notes, the patients all had NVAF and a tendency to take aspirin for stroke prevention. Second, the notes were all from a single integrated health care system. It is possible that this algorithm would need to be adapted for external applications. Third, this algorithm was only applied to one medication, and its portability to other medications needs to be tested. Finally, this tool was based on a proprietary NLP software system that creates barriers to dissemination. For institutions without the capability to implement an NLP algorithm, the keyword-based search could be used to screen the notes and find those sentences containing keywords. If resources are available, manual review of those sentences is quick and straightforward. Manual review efforts could be affordable for research studies with a limited study population.

Despite these limitations, our study has a number of unique strengths. It was conducted on a large population within an integrated care system with its own pharmacy system. The size of our study population is much larger compared with prior clinical NLP studies. Our model of care enables us to capture the complete medical history of our members and enables us to study the medication use on a longitudinal scale. Even though we only focused on 1 OTC medication, the concept and principle are applicable to other nonprescription or prescription medications.

High- and low-risk NVAF patients can be identified by calculating $CHA_2D_2$-VASc scores using clinical variables from our EMR system. Anticoagulant is an important pharmacotherapy for NVAF patients, and the use of anticoagulants is commonly determined using structured medication data (eg, pharmacy claims). This particular study was focused on identifying aspirin use by developing an NLP algorithm for use on EMRs. Future studies may be able to determine a complete picture of both anticoagulant and antiplatelet use for patients with NVAF and outcomes associated with these treatments.

## CONCLUSIONS

With more EMRs becoming readily available in different health care systems, there is a great need for new methods to evaluate medications that are not easily found in structured pharmacy databases. In this study, we developed and validated an NLP algorithm specifically designed to identify low-dose aspirin use information from the clinical notes in an NVAF population. This NLP algorithm accurately identified aspirin use based on the information documented in the clinical notes. This method can be a valuable tool to supplement existing structured medication data.

## ACKNOWLEDGMENTS

All authors were involved in drafting the article and approved the final version to be published. Dr. Zheng had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Chengyi Zheng, Nazia Rashid, JaeJin An were responsible in study conception

and design. Chengyi Zheng is responsible in acquisition of data, design and implementation of NLP system. and Chengyi Zheng, Nazia Rashid, River Koblick, JaeJin An were responsible in analysis and interpretation of data.

## CONFLICTS OF INTEREST

This study was sponsored by a research grant from Bristol-Myers Squibb Company (BMS). BMS had no role in the study design or in the collection, analysis, or interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript for publication. Publication of this article was not contingent upon approval by BMS. The authors have indicated that they have no conflicts of interest regarding the content of this article.

## SUPPLEMENTARY MATERIAL

Supplementary data accompanying this article can be found in the online version at http://dx.doi.org/10.1016/j.clinthera.2015.07.002.

## REFERENCES

1. January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: A report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the heart rhythm society. *Circulation*. 2014;130:e199–e267.
2. You JJ, Singer DE, Howard PA, et al. Antithrombotic therapy for atrial fibrillation: Antithrombotic therapy and prevention of thrombosis, 9th ed: American college of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2 Suppl):e531S–e5375.
3. Campbell CL, Smyth S, Montalescot G, et al. Aspirin dose for the prevention of cardiovascular disease: A systematic review. *JAMA*. 2007;297:2018–2024.
4. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. *J Am Med Inform Assoc*. 2011;18:544–551.
5. Friedman C, Rindflesch TC, Corn M. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *J Biomed Inform*. 2013;46:765–773.
6. Fuster V, Ryden LE, Cannom DS, et al. 2011 ACCF/AHA/HRS focused updates incorporated into the ACC/AHA/ESC 2006 guidelines for the management of patients with atrial fibrillation: A report of the american college of cardiology foundation/american heart association task force on practice guidelines. *Circulation*. 2011;123:e269–e367.
7. Zheng C, Rashid N, Wu YL, et al. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. *Arthritis care & research*. 2014;66:1740–1748.
8. Danforth KN, Early MI, Ngan S, et al. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. *J Thorac Oncol*. 2012;7:1257–1262.
9. Thomas AA, Zheng C, Jung H, et al. Extracting data from electronic medical records: Validation of a natural language processing program to assess prostate biopsy results. *World J Urol*. 2014;32:99–103.
10. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc*. 2010;17:514–518.
11. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc*. 2010;17:524–527.
12. Doan S, Bastarache L, Klimkowski S, et al. Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc*. 2010;17:528–531.
13. Spasic I, Sarafraz F, Keane JA, et al. Medication information extraction with linguistic pattern matching and semantic rules. *J Am Med Inform Assoc*. 2010;17:532–535.
14. Sohn S, Murphy SP, Masanz JJ, et al. Classification of medication status change in clinical narratives. *AMIA Annu Symp Proc*. 2010;762–766.
15. Liu M, Jiang M, Kawai VK, et al. Modeling drug exposure data in electronic medical records: an application to warfarin. *AMIA Annu Symp Proc*. 2011;815–823.
16. Pakhomov SV, Shah ND, Hanson P, et al. Automated processing of electronic medical records is a reliable method of determining aspirin use in populations at risk for cardiovascular events. *Inform Prim Care*. 2010;18:125–133.
17. Rolita L, Freedman M. Over-the-counter medication use in older adults. *J Gerontol Nurs*. 2008;34:8–17.
18. Jacobs LR. Prescription to over-the-counter drug reclassification. *Am Fam Physician*. 1998;57:2209–2214.
19. Qato DM, Alexander GC, Conti RM, et al. Use of prescription and over-the-counter medications and dietary supplements among older adults in the united states. *JAMA*. 2008;300:2867–2878.
20. Breydo EM, Chu JT, Turchin A. Identification of inactive medications in narrative medical text. *AMIA Annu Symp Proc*. 2008;66–70.
21. Hasan S, Duncan GT, Neill DB, et al. Automatic detection of omissions in medication lists. *J Am Med Inform Assoc*. 2011;18:449–458.
22. Tam VC, Knowles SR, Cornish PL, et al. Frequency, type and clinical importance of medication history errors at admission to hospital: a systematic review. *CMAJ*. 2005;173:510–515.

23. Brown M.T., Bussell J.K. Medication adherence: Who cares? Mayo Clinic proceedings. 2011; 86: 304–14.
24. Pronovost P, Weast B, Schwarz M, et al. Medication reconciliation: A practical tool to reduce the risk of medication errors. *Journal of critical care*. 2003;18:201–205.

**Address correspondence to:** Chengyi Zheng, PhD, Department of Research and Evaluation, Kaiser Permanente Southern California, 100 S Los Robles Ave, Second Floor, Pasadena, CA 91101. E-mail: Chengyi.X.Zheng@kp.org

## SUPPLEMENTARY MATERIALS

Appendix. List of aspirin terms.

A.S.A.
ACETYLSALICYLIC ACID
ADULT ASA
ADULT ASPIRIN
ASA, BUFFERED
ASA, CHEW TAB
ASA, COATED
ASA, DR
ASA, EC
ASA, ENTERIC COATED
ASA, ENTERIC-COATED
ASA, TAB
ASA, TBEC
ASA BUFFERED
ASA CHEW TAB
ASA CHEWABLE
ASA COATED
ASA DR
ASA EC
ASA ENTERIC COATED
ASA ENTERIC-COATED
ASA TAB
ASA TBEC
ASA.
ASA-BUFFERED
ASA-CHEW TAB
ASA-COATED
ASA-DR
ASA-EC
ASA-ENTERIC COATED
ASA-ENTERIC-COATED
ASA-LOW DOSE
ASA-LOW STRENGTH
ASA-TAB
ASA-TBEC
ASCRIPTIN
ASP.
ASPERGUM
ASPIRIN
ASPIRIN, BUFFERED
ASPIRIN, CHEW TAB
ASPIRIN, COATED
ASPIRIN, DR
ASPIRIN, EC
ASPIRIN, ENTERIC COATED
ASPIRIN, ENTERIC-COATED
ASPIRIN, TAB
ASPIRIN, TBEC
ASPIRIN BUFFERED
ASPIRIN CHEW TAB
ASPIRIN CHEWABLE
ASPIRIN COATED
ASPIRIN DR
ASPIRIN EC
ASPIRIN ENTERIC COATED
ASPIRIN ENTERIC-COATED
ASPIRIN TAB
ASPIRIN TBEC
ASPIRIN-BUFFERED
ASPIRIN-CHEW TAB
ASPIRIN-COATED
ASPIRIN-DR
ASPIRIN-EC
ASPIRIN-ENTERIC
ASPIRIN-ENTERIC COATED
ASPIRIN-ENTERIC-COATED
ASPIRIN-LOW DOSE
ASPIRIN-LOW STRENGTH
ASPIRINS
ASPIRIN-TAB
ASPIRIN-TBEC
BAYER ASA
BAYER ASPIRIN
BUFFERED ASA
BUFFERED ASPIRIN
BUFFERIN
CHEWABLE ASA
CHEWABLE ASPIRIN
EC ASA
ECASA
ECOTRIN
EMPIRIN
ENTERIC-ASPIRIN
ENTERICIN
EXTREN
HALFPRIN
MEASURIN
ZORPRIN

List of Terms for Low Strength or Baby Aspirin:

ADULT LOW DOSE ASA
ADULT LOW DOSE ASPIR
ADULT LOW DOSE ASPIRIN
ADULT LOW STRENGTH ASA
ADULT LOW STRENGTH ASPIR
ADULT LOW STRENGTH ASPIRIN
ASA, LOW

ASA, LOW DOSE
ASA, LOW STRENGTH
ASA 81
ASA LOW
ASA LOW DOSE
ASA LOW STRENGTH
ASA-81
ASA-LOW
ASPIR 81
ASPIR LOW
ASPIR-81
ASPIRIN, LOW
ASPIRIN, LOW DOSE
ASPIRIN, LOW STRENGTH
ASPIRIN 81
ASPIRIN LOW
ASPIRIN LOW DOSE
ASPIRIN LOW STRENGTH
ASPIRIN-81
ASPIRIN-LOW
ASPIR-LOW
BABY ASA
BABY ASPIRIN
BABY-ASA
BABY-ASPIRIN
BASA
B-ASA
BAYER 81 MG
BAYER BABY ASA
BAYER CHILDRENS ASPIRIN
BAYER LOW DOSE
BAYER LOW DOSE ASA
BAYER LOW DOSE ASPIRIN
BAYER LOW STRENGTH
CHILDREN'S ASA
CHILDREN'S ASPIRIN
ECOTRIN LOW DOSE
ECOTRIN LOW STRENGTH
List of misspelled terms for *aspirin*
aaspirin
aslirin
asopirin
aspiorin
aspirijn
aspirion
aspirni
aspoirin
asppirin
asspirin
saspirin
aspiirin
aspirijn
aspiriun
asspirin
aspiin
aspilrin
aspiron
aspoirin
asoirin
aspitin
aspitrin
spirin
asirin
aspiirn
aaspirin
aspirinn
apsirin
aspiriin
asiprin
asapirin
aspriin
apirin
aspiri
aspirn
asprin
aapirin
adpirin
adspirin
aepirin
aespirin
aqspirin
as0irin
as0pirin
asdpirin
asepirin
aslpirin
asp0irin
asp8irin
asp8rin
asp9irin
asp9rin
aspi4in
aspi4rin
aspi5in
aspi5rin
aspi8rin
aspi9rin
aspidin