JAMIA Open, 4(3), 2021, 1–9
doi: 10.1093/jamiaopen/ooab071
Research and Applications

Research and Applications

Natural language processing for automated annotation of medication mentions in primary care visit conversations

Craig H Ganoe[1], Weiyi Wu[1], Paul J Barr[2], William Haslett[1], Michelle D Dannenberg[2], Kyra L Bonasia[2], James C Finora[2], Jesse A Schoonmaker[2], Wambui M Onsando[2], James Ryan[3], Glyn Elwyn [2], Martha L Bruce[2], Amar K Das[2] and Saeed Hassanpour [1,4,5]

[1]Biomedical Data Science Department, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire, USA, [2]The Dartmouth Institute for Health Policy & Clinical Practice, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire, USA, [3]Ryan Family Practice, Ludington, Michigan, USA, [4]Computer Science Department, Dartmouth College, Hanover, New Hampshire, USA, and [5]Epidemiology Department, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire, USA

Corresponding Author: Saeed Hassanpour, PhD, One Medical Center Drive, HB 7261, Lebanon, NH 03756, USA (Saeed.-Hassanpour@dartmouth.edu)

Received 29 March 2021; Revised 27 July 2021; Editorial Decision 30 July 2021; Accepted 4 August 2021

ABSTRACT

Objectives: The objective of this study is to build and evaluate a natural language processing approach to identify medication mentions in primary care visit conversations between patients and physicians.
Materials and Methods: Eight clinicians contributed to a data set of 85 clinic visit transcripts, and 10 transcripts were randomly selected from this data set as a development set. Our approach utilizes Apache cTAKES and Unified Medical Language System controlled vocabulary to generate a list of medication candidates in the transcribed text and then performs multiple customized filters to exclude common false positives from this list while including some additional common mentions of the supplements and immunizations.
Results: Sixty-five transcripts with 1121 medication mentions were randomly selected as an evaluation set. Our proposed method achieved an F-score of 85.0% for identifying the medication mentions in the test set, significantly outperforming existing medication information extraction systems for medical records with F-scores ranging from 42.9% to 68.9% on the same test set.
Discussion: Our medication information extraction approach for primary care visit conversations showed promising results, extracting about 27% more medication mentions from our evaluation set while eliminating many false positives in comparison to existing baseline systems. We made our approach publicly available on the web as an open-source software.
Conclusion: Integration of our annotation system with clinical recording applications has the potential to improve patients' understanding and recall of key information from their clinic visits, and, in turn, to positively impact health outcomes.

Key words: clinic visit recording, medication information extraction, natural language processing

LAY SUMMARY

In this work, we built a natural language processing approach to identify medication mentions in primary care visit conversations between patients and physicians to allow patients to easily find important elements of their recorded conversations with their physicians. This method annotates medication mentions in the text transcribed from office visits. Our approach utilizes a repository of common medication names to generate a list of medication candidates in the transcribed text, and then exclude common false positives from this list while including some additional common mentions of supplements and immunizations in the medication list for a transcript. We evaluated our method on a test set of 65 clinic visit transcripts with 1121 medication mentions. In this evaluation, our proposed method achieved a high performance for identifying the medication mentions, significantly outperforming existing medication information extraction systems for medical records. Integration of this annotation system with clinical recording applications has the potential to improve patients' understanding and recall of key information from the clinic visits, and their health outcomes.

BACKGROUND AND SIGNIFICANCE

Forty to 80% of healthcare information is forgotten *immediately* by patients postvisit.[1–4] Poor recall and understanding of medical concepts have been identified as significant barriers to self-management, a central component of the Chronic Care Model, resulting in poorer health outcomes.[5–7] These barriers are amplified in older adults with multimorbidity,[8–11] where reduced cognitive capacity,[12–14] low health literacy,[15,16] and complex treatment plans are common.[17–19] Older adults with multimorbidity account for 96% of Medicare expenditures, and in the absence of optimal self-management, they experience a lower quality of life and greater functional decline.[10,11,20–26]

An after-visit summary, shared via a patient portal, is a common strategy to improve recall of visit information.[27–29] Open notes is a current trend in healthcare that encourages clinicians to share the visit notes with patients. Sharing visit notes with patients not only increases patients' confidence in their ability to manage their health and understanding of their care but also enhances the communication efficiency. Through accessing visit notes, patients can take medications as prescribed and remember their healthcare plan better.[30,31] However, summaries impose a significant burden on clinicians who must document the entire visit in terms that are understandable to patients, with low health literacy being common.[32,33] Alternatively, audio recordings can provide a full account of the clinic visit and are an effective modality—71% of patients listen to recordings and 68% share their recording with a caregiver.[34] Clinic recordings improve patient understanding and recall of visit information, reduce anxiety, increase satisfaction, and improve treatment adherence.[34–40] As patient demand for recordings increases,[41,42] a growing number of clinics across the United States are offering audio recordings of clinic visits, and a recent survey reveals that almost a third of clinicians in the United States have shared a recording of a clinic visit with patients.[43]

Yet, unstructured clinic recordings may overwhelm patients.[41,44] Advances in data science methods, such as natural language processing (NLP), can be used to identify patterns in unstructured data and extract clinically meaningful information. These methods have been used to predict hospital readmissions[45] and future radiology utilization,[46] and to characterize the significance, change, and urgency of clinical findings in medical records.[47–51] As such, we have developed a recording system for patients that applies NLP methods to unstructured clinic visit recordings.[52]

In this article, we describe an approach to extract mentions of medication names in transcripts of clinic visit audio recordings. Annotating mentions of medications discussed during a clinic visit recording can provide added value to the audio-recorded health information. We use NLP to highlight medication mentions in transcripts of clinic recordings. These annotations can be utilized to index the audio and aid visit recall by enabling key visit information to be easily accessed. In addition, the indexed medical concepts can be linked to credible and trustworthy online resources. These resources would provide additional information about medications to aid in patient understanding. Such an approach could potentially increase patient self-management, and, when shared with caregivers, could increase their confidence in delivering care.

At the time of this work, no prior work focused on extracting medication information from clinic visit conversations and their transcriptions. There has been some work on the extraction of medication names and also prescription-related attributes such as dosage and frequency from the medical text (primarily clinical notes). These systems have mainly focused on the extraction of medication information from written clinical notes. In 2009, the Third i2b2 Shared-Task on Challenges in Natural Language Processing for Clinical Data Workshop focused on medication information extraction. The challenge was to extract and label medication-related terms (medication name, dosage, frequency, etc.) from discharge summaries.[53] Teams were given 696 summaries for development, and then 547 summaries were used for evaluation. Twenty teams submitted entries to the challenge, with the top result for annotating medication names being an F-score of 90.3% on the evaluation data set, utilizing a combination of a rule-based approach with two machine learning models (conditional random field and support vector machine). This top approach also achieved an F-score of 90.81% on an internal test set of 30 clinical records when evaluated by the system's authors.[54]

Since the 2009 i2b2 challenge, additional work has been done to improve medication information extraction methods. Sohn et al.[55] developed Medication Extraction and Normalization (MedXN) to extract medication information and map it to the most specific RxNorm concept possible. This group reported an F-score of 97.5% for medication name on a test set of 26 clinical notes containing 397 medications. In 2014, MedEx, the system with the second-best results in the i2b2 challenge, was reimplemented using Unstructured Information Management Architecture (UIMA) to extract drug names and map them to both generalized and specific RxNorm concepts.[56] This system, named MedEx-UIMA, achieved an F-score of 97.5% for extracting and mapping to the most generalized concept and an F-score of 88.1% for mapping to the most specific concept, evaluating on a set of 125 discharge summaries from the original i2b2 challenge. The authors concluded that the new MedEx-UIMA implementation was consistent with and sometimes outperformed the original MedEx method. Most recently, PredMed was developed to extract medication names and related terms from office visit notes.[57] The comparison of PredMed for extracting medication names to earlier versions of MedEx and MedXN on a test set of 50 visit encounter notes showed F-scores of 80.0% for PredMed, 74.8% for MedEx, and 83.9% for MedXN. Since MedEx-UIMA and MedXN are available as open-source systems, we used these systems as baselines for comparison in our study.

In another related work, Kim et al.[58] developed a method for retrieval of biomedical terms in tele-health call notes. Their team identified two types of noise in these records, explicit—including “spelling errors, unfinished sentences, omission of sentence marks, etc.”—and implicit—“non-patient information and a patient’s untrustworthy information”—and sought to remove that noise as part of their method. Utilizing a bootstrapping-based pattern learning process to detect variations related to the explicit noise, and dependency path-based filters to remove the implicit noise, their system achieved an F-score of 77.33% for detecting biomedical terms on evaluation data from 300 patients. This tool and its corresponding codebase are not publicly available for comparison for this study. Furthermore, recently, there has been additional work on the analysis of medical conversations based on deep learning models.[59–63] However, unlike our open-source tool, the presented proprietary tools and their corresponding test sets are not publicly available for comparison to our approach. Of note, some of these previous works are focused on relation extraction and were evaluated for identifying relations between medications and their properties,[59] rather than finding medication mentions themselves. Also, the proposed deep learning models require a large amount of data for training and fine-tuning, including tens of thousands of doctor–patient annotated conversations.[60,61] On the other hand, our approach is developed using only a fraction of those deep learning models’ training sets. Considering the finite list of possible medications, our approach could achieve high performance (F-score: 85%) by efficiently using the proposed rules and filters without requiring large data sets and computational resources.

MATERIALS AND METHODS

Our NLP pipeline was developed and validated to extract medication mentions in clinic visit transcripts. We define medication mentions as any place in the text that a term refers to a medication by a specific or general name or common lay term. Our pipeline takes advantage of Apache clinical Text Analysis and Knowledge Extraction System (cTAKES)[64] to generate a primary candidate list of medication mentions. Subsequently, our approach filters out false-positive medical mentions in this list and adds the medication mentions that cTAKES misses in visit transcripts. Our workflow took the original visit text transcripts and processed them through the cTAKES default clinical pipeline resulting in a set of corresponding UIMA CAS XMI output files with the sentences, parts of speech, and all clinical concepts annotated by cTAKES. The software we developed for our approach utilizes the CAS XMI output from cTAKES and outputs our final annotated medication mentions in a Knowtator file format. Our approach and cTAKES baseline pipeline for identifying medications in this study do not utilize the outputted part of speech tags from cTAKES. eHOST was used in this study to compute metrics for our evaluation. Outputs from MedEx-UIMA and MedXN were also converted to Knowtator format to compute evaluation metrics using eHOST.

Visit transcripts data set

Transcripts of 85 patient visits with a primary care physician were used as our data set in this study. These visits were audio-recorded and transcribed by a HIPAA compliant commercial medical transcription service. These recordings, which came from eight clinicians, were 31 min long on average, ranging from 5.5 to 70.5 min. This study and the use of human subject data in this project were approved by the committee for the Protection of Human Subjects at Dartmouth College (CPHS STUDY#30126) with informed consent. Table 1 shows the demographics of the participants who had their clinical visit recordings used in our study.

Ten transcripts were randomly selected from this data set as a development set. Another ten of the visit transcripts were randomly selected as a validation set for our model. The remaining 65 transcripts were reserved as a held-out test set for evaluation.

Annotation for medication mentions

All the transcripts were independently annotated for medication mentions by two second-year medical students using the Extensible Human Oracle Suite of Tools (eHOST) software.[65] The two annotators initially worked through blocks of 5 or 10 transcripts, meeting after annotating each block to track inter-annotator agreement (IAA) on the identified medication mentions, discuss disagreements, and improve their accuracy in this annotation task, which led to steadily higher IAA over time. Our IAA calculation considers overlapping annotations as a match, allowing a flexible annotation arrangement for compound medication names. Once the annotators reached over 80% IAA, we considered them trained in this annotation task. Subsequently, they annotated the entire set of transcripts. Inter-annotator agreement for medication mentions between our annotators for the 65 transcripts in the evaluation data set was 84.6%. In that data set, Annotator 1 annotated 1076 instances of medication mentions, and Annotator 2 annotated 1048 instances of medication mentions.

For evaluation, we created a set of gold standard medication mentions in our evaluation data set based on the work of our expert annotators. Our labels are based on overlapping annotations of two annotator experts. All medication mentions in our evaluation set that were agreed upon by the two expert annotators were kept in this gold standard set. A physician, trained in the method used by the annotators, served as an adjudicator to resolve disagreements between our annotators. A disagreement in the annotations would occur when one annotator had annotated a medication mention while the other had not. Disagreements were resolved by the adjudicating physician either choosing to keep the annotation from a single annotator in the gold standard set, or choosing to reject it. The adjudicating physician also reviewed disagreements between the output from our model and the set of annotations from the human adjudicator to identify true positives and false positives for evaluating our model by either choosing to keep the annotation from either source or rejecting it. As a result, a small number of medication mentions that were missed by both annotators were thus added to our gold standard set. The resulting gold standard evaluation data set contained 1121 medication mentions.

cTAKES baseline for annotating medications in transcripts

Our baseline approach was to utilize Apache cTAKES[64] to identify the medication mentions in the transcripts. cTAKES is an open-source widely-used NLP system for biomedical text processing. As

Table 1. Participant demographics for transcribed visit recordings (SD: standard deviation)

	Development data set (%)	Validation data set (%)	Evaluation data set (%)	Total (%)
Number of recordings in data set	10	10	65	85
Participants with demographic data[a]	9[a] (90.0)	10 (100.0)	54[a] (83.1)	73[a] (85.9)
Gender				
Female	4 (40.0)	6 (60.0)	34 (52.3)	44 (51.8)
Male	5 (50.0)	4 (40.0)	20 (30.8)	29 (34.1)
Mean age (SD) [range]	50.00 (18.57) [23–87]	58.60 (18.95) [20–77]	54.65 (15.61) [25–92]	54.62 (16.35) [20–92]
Race				
White	9 (90.0)	10 (100.0)	54 (83.1)	73 (85.9)
Ethnicity				
Not Hispanic or Latino	9 (90.0)	10 (100.0)	52 (80.0)	71 (83.5)
Declines to list	–	–	2 (3.1)	2 (2.4)
Language spoken				
English	9 (90.0)	10 (100.0)	54 (83.1)	73 (85.9)
Recording length (SD) [range]	36.46 (17.37) [17.55–70.39]	37.07 (10.00) [20.16–49.41]	28.36 (11.95) [5.42–55.33]	30.55 (12.85) [5.42–70.39]
Visit type				
Annual physical established patient	3 (30.0)	6 (60.0)	8 (12.3)	17 (20.0)
Established patient follow-up	2 (20.0)	3 (30.0)	29 (44.6)	34 (40.0)
Same day add-on	2 (20.0)	1 (10.0)	11 (16.9)	14 (16.5)
New patient workup	2 (20.0)	–	1 (1.5)	3 (3.5)
History and physical	–	–	2 (3.1)	2 (2.4)
Other[b]	–	–	3 (4.6)	3 (3.5)

[a] Demographic data was not captured for 12 of the 85 transcripts.

[b] "Other" includes "Res-visit 20" and diabetic follow-up.

one of its NLP capabilities, cTAKES is able to annotate and extract medical information from the free text of clinical reports. We utilized the Default Clinical Pipeline of cTAKES (version 4.0.0) and its Language System (UMLS) Metathesaurus[66] fast dictionary lookup functionality. cTAKES' UMLS fast dictionary lookup, by default, uses sentences as a lookup window for matching, covering the text of the entire document. For our dictionary, we used the provided prebuilt cTAKES dictionary, which includes RxNorm and SNOMED-CT. SNOMED-CT provides extensive coverage of laboratory tests and clinical measurements, while RxNorm focuses on drug names and codes. Our only modification to the default cTAKES configuration was to utilize its PrecisionTermConsumer function, which refines annotations to the most specific variation (eg, if it finds the text "colon cancer" in a report, it only annotates "colon cancer" but not "colon" nor "cancer"). Since cTAKES is designed to work with medical record-free text, there is an assumption that input text is a clinical note, written by an individual with a medical background. In contrast, the visit transcripts are typically a dyadic conversation between a patient and their physician.

Our model for annotating medications in transcripts

After initial experiments with cTAKES and UMLS as a means to find medications mentioned in transcribed clinic visit conversations, we explored additional methods to filter out common false positives from the output generated by cTAKES. For this purpose, we took an iterative approach, looking at the most common errors in cTAKES outcomes for identification of medication mentions in our development set and developed new rule-based filters to detect and remove those from the cTAKES output. As our accuracy on the development set improved by filtering out many types of false positives (described in detail below), we ran our model against our validation set, finding that immunizations along with herbs and supplements persisted as typical errors. cTAKES had difficulty differentiating immunizations from diagnoses (eg, chickenpox vaccine vs chickenpox). Also, cTAKES did not annotate some commonly used herbs and supplements. In the next sections, we describe how our approach adds annotations for immunizations, herbs, and supplements, while filtering out false positives for medication mentions. An overview of this approach is shown in Figure 1. We have made our code for this approach publicly available on GitHub (https://github.com/BMIRDS/HealthTranscriptAnnotator).

Common word filtering

Since many of the words appearing as false positives in the cTAKES output for medication annotations are common conversational words that have second meanings as medication names or acronyms (eg, "today" is also ToDAY, a name for an antibiotic primarily in veterinary use that appears in UMLS), we decided to utilize a large dictionary of common words to filter out these occurrences. We chose to use a dictionary of the 10 000 most common English words from Google's Trillion Word Corpus (https://github.com/first20hours/google-10000-english).[67] If any of those 10 000 words were annotated by cTAKES as a medication, our model removes that annotation, with a small subset of exceptions. From the 10 000 common words list, there were 24 words that are considered as exceptions and are allowed to remain annotated as medications. These words fit into three categories: (1) names of common medications (eg, "Ambien", "Insulin", etc., which accounted for 17 of the

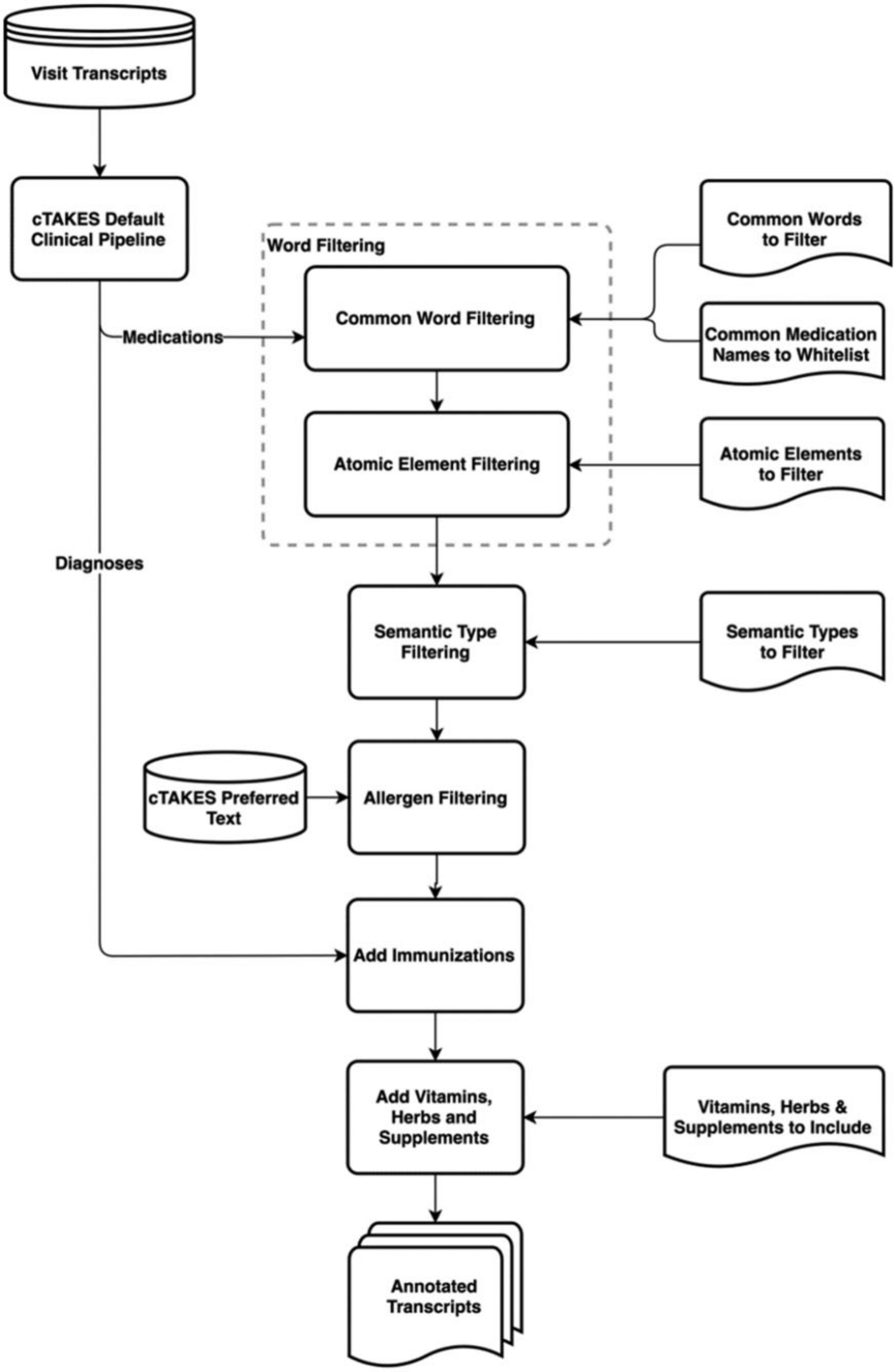

Figure 1. Overview of our approach to annotate medication mentions in clinic visit transcripts.

24); (2) generic terms (eg, “herb”, “supplement”, and “vitamin”, along with their plurals); and (3) the word “flu”, which can refer to either a diagnosis or an immunization.

Chemical element filtering

cTAKES also annotates all chemical elements as medication mentions; “gold,” for example, can also be taken as a medication. In our approach, we systematically remove annotations for those chemical elements that are not typically taken as a medication or as a supplement. These chemical elements include actinium, aluminum, antimony, argon, arsenic, astatine, barium, beryllium, bromine, cadmium, carbon, cerium, cesium, chlorine, cobalt, copper, dysprosium, erbium, europium, fluorine, francium, gallium, germanium, gold, hafnium, helium, hydrogen, indium, iridium, krypton, lanthanum, lead, lutetium, mercury, molybdenum, neodymium, neon, nickel, niobium, osmium, palladium, phosphorus, platinum, polonium, potassium, promethium, protactinium, radium, radon, rhenium, rubidium, ruthenium, samarium, scandium, silicon, silver, strontium, tantalum, tellurium, terbium, thorium, thallium, tin, titanium, tungsten, uranium, vanadium, xenon, ytterbium, yttrium, and zirconium.

UMLS semantic type filtering

In our error analysis for cTAKES outputs, we also examined UMLS semantic types for the terms that cTAKES annotated as medication mentions. The six types shown in Table 2 generally produced false positives and few to no true positives. Our approach removes these semantic types as medication annotations from the cTAKES output where they occur.

Allergen filtering

cTAKES annotates a number of food and food ingredient-related terms (eg, "coconut") as medication mentions, denoting them as an allergenic. We identify those annotations that have the word "allergenic" included in their preferred cTAKES text metadata, and we remove those annotations from the output of cTAKES in our model's output.

Immunization additions

A small number of medication-related UMLS terms are considered as both diagnoses and immunizations/vaccinations (eg, "flu" and "pertussis"). As a result, cTAKES annotation outputs were inconsistent about annotating these terms as immunizations/vaccinations or diagnoses. To improve the annotation of immunizations as medications, we also investigated the cTAKES diagnosis annotations. Since cTAKES segments the input text into sentences, we searched for the words "vaccine," "shot," "booster," and "pill" in the same sentence as a diagnosis annotation, and if both co-occurred, we annotated the diagnosis text as a medication.

Vitamin, herb, and supplement additions

cTAKES also produces inconsistent results for annotating herbs and supplements. Our approach adds an additional dictionary of common herbs and supplements from MedlinePlus (https://medlineplus.gov/druginfo/herb_All.html) to capture these.[68]

Evaluation

We applied our model on the evaluation data set containing 65 transcripts to annotate medication mentions, in addition to capturing the original medication mention annotation output from cTAKES 4.0.0's default clinical pipeline. We also applied publicly available MedEx-UIMA 1.3.7 and MedXN 1.0.1 software on the evaluation data set to compare our results with their medication name annotations as the baselines.

Table 2. UMLS semantic types in cTAKES annotations that are filtered out in our approach

TUI	Semantic type
T114	Nucleic acid, nucleoside, or nucleotide
T122	Biomedical or dental material
T123	Biologically active substance
T125	Hormone
T130	Indicator, reagent, or diagnostic aid
T197	Inorganic chemical

RESULTS

We calculated the standard evaluation metrics of precision, recall, and F-score for our proposed approach and the baseline methods using the medication mention gold standards in our validation and evaluation sets. These evaluation metrics are shown in Table 3. We compared the results from cTAKES, MedEx-UIMA, MedXN, and our proposed model for identification of the gold standard medication mentions for the 65 transcripts in the evaluation set. Table 4 shows this comparison.

DISCUSSION

Our results indicate that the proposed approach significantly reduced the number of false positives, with a relatively small drop in the number of true positives and false negatives, in comparison to the best of three baseline models. As highlighted in Table 4, our proposed model has the best overall performance in comparison to the other baseline methods, with all of its evaluation metrics falling in the range of 83–87%. Overarching the finer aspects of our work is the observation that extracting medical terms from conversational dialogue between patients and their primary care physician has distinct challenges, such as more informal medical terms and unstructured content, in comparison to extracting terms from typical clinical, note-like reports. To the best of our knowledge, the proposed work in this article is the first attempt to extract medical terminology from conversations between a patient and their physician. Prior work for finding medication mentions has focused on written clinical reports.[47–51]

Our error analysis suggests that baseline approaches, which rely on dictionaries, struggle with patient-clinician conversational text because of language like filler words (eg, "aha" and "hmm") matching with abbreviations for medications, and the fact that common conversational words are often used as medication names. We also observed, among the filters that we applied to the original cTAKES outputs, that filtering out "hormone" semantic type had the most impact on the improvement of the results. The most common ($n > 10$) false negatives by cTAKES were "flu shot" (36), "tetanus" (14), and "inhaler" (11). Among annotations that were missed by one of our two annotators, the most common (>10) cases were "Vitamin D" (21), "flu shot" (13), and "Mirena" (11). The most common (>10) false positives in the evaluation set annotated by our approach were "clot" (15 occurrences) and "over-the-counter" (11 occurrences), and the most common (>10) false negatives missed by our approach were "inhaler," "calcium," and "tetanus" (11 occurrences each). A slight but consistent majority of false positives in our data set were from the discussion of lab test results, which will be a focus of our future work to improve the current results.

One advantage of our approach is that each portion of our pipeline was designed to generalize addressing issues seen during development, so our approach was able to recognize terms outside the development/validation data sets. Other rule-based and dictionary-based systems have often relied on whitelisting/blacklisting terms from their development data sets, which limits how they generalize outside their development data. For example, our use of the 10 000 most common English words from Google's Trillion Word Corpus allows us to recognize and filter many common words. Our solu-

Table 3. The performance of our approach on the validation and evaluation sets

Data set	No. of true positives	No. of false positives	No. of false negatives	Precision (%)	Recall (%)	F-score (%)
Validation set	291	46	40	86.4	87.9	87.1
Evaluation set	1062	168	206	86.3	83.8	85.0

Table 4. The comparison of our proposed approach to existing baseline models for identification of gold standard medication mentions in our evaluation set

Model	No. of true positives	No. of false positives	No. of false negatives	Precision (%)	Recall (%)	F-score (%)
cTAKES	1119	2814	163	28.5	**87.3**	42.9
MedEx-UIMA	830	1215	292	40.6	74.0	52.4
MedXN	832	318	432	72.3	65.8	68.9
Our approach	1062	168	206	**86.3**	83.8	**85.0**

tions to chemical elements, UMLS semantic types, allergens, immunizations, and vitamins/herbs/supplements were also all designed to potentially recognize terms outside what appeared in our development data set.

Of note, our evaluation has limitations. Foremost, our evaluation data set is relatively small and is from a single medical institution. We plan to extend our evaluation data set in future work to test the generalizability of the proposed approach. In addition, because our gold standard was created by reaching consensus between two medical annotators and carrying out our approach, it is possible that other baseline methods, such as cTAKES, found a small number of true positives that were not accounted for by any of the annotators or our proposed method. That said, the sheer number of false positives generated by cTAKES makes adjudication of its medication mention output impractical. Also, our approach has been developed to detect only medication mentions in primary care visit notes. Identifying other types of medical words and their properties in these notes can significantly increase and broaden the utility of our approach. Especially, detecting additional information about medications, such as frequency, dose, refill, modifications, and side effects, can benefit the patients. We plan to extend our approach to identify additional information about medications and other semantics types, such as disorders, in future work. Another limitation is that clinical visit transcripts are more complex if English is not the patient's first language or if an interpreter is involved. Transcripts do not reflect non-verbal communication, such as visible emotions and body language. The transcripts do not include the assessment or plan section of the visit note, which reflect the clinician's summary and reflection that may occur after the visit itself. Finally, our approach, which is based on controlled vocabulary and rule-based filtering, does not consider word context and the corresponding contextual semantics in different circumstances. Since one of our goals is using these annotations to index segments of clinic visit conversations for end-users to review postvisit, we plan to conduct future work with end-users to determine how these limitations may impact the usability of the system. Future plans to integrate the proposed information extraction methods in this study with a digital library of clinic visit recordings is expected to make patients and caregivers more knowledgeable and confident of their health care needs, resulting in greater self-management capabilities.

Notably, as we fine-tuned our model on the validation set, we observed that context words in a sentence can be critical in our task, for example, for determining mentions of immunizations/vaccinations. Our result suggests that although the dictionary- and rule-based methods can achieve a promising result (F-score = 85%) for identification of medication mentions in clinic visit conversations, additional improvements in this domain will be gained through considering contextual semantics and machine learning models, which our team will pursue in future work.

CONCLUSION

In this work, we developed an NLP pipeline for finding medication mentions in primary care visit conversations. The proposed model achieved promising results (Precision = 86.3%, Recall = 83.8%, F-Score = 85.0%) for identification of medication mentions in 65 clinic visit transcripts in our evaluation set. Since this is a first-of-a-kind study with clinic visit transcripts, we compared our approach to three existing systems used for extracting medication mentions from clinical notes. This comparison shows our approach can extract about 27% more medication mentions while eliminating many false positives in comparison to existing baseline systems. Integration of this annotation system with clinical recording applications has the potential to improve patients' understanding and recall of key information from their clinic visits, and, in turn, behavioral and health-related outcomes. We plan to explore this potential in future trials of our system.

CONTRIBUTORS

All authors reviewed and edited the manuscript and contributed to the study concept and design of the experiments. CHG, PJB, WH, and MDD collected the data. KLB, JCF, JAS, WMO, and JR contributed to data annotation. CHG, WW, and SH analyzed the data and wrote the manuscript. SH and PJB acquired the funding, and SH supervised the study.

FUNDING

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01 LM012815 and the Gordon & Betty Moore Foundation under award number GBMF-4952. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Gordon and Betty Moore Foundation.

Conflict of interest statement

GE: Glyn Elwyn has edited and published books that provide royalties on sales by the publishers: the books include Shared Decision Making (Oxford University Press) and Groups (Radcliffe Press). Glyn Elwyn's academic interests are focused on shared decision making and coproduction. He owns copyright in measures of shared decision making and care integration, namely collaboRATE, integRATE (measure of care integration, consideRATE (patient experience of care in serious illness), coopeRATE (measure of goal setting), incorpoRATE (clinician attitude to shared decision making, Observer OPTION-5 and Observer OPTION-12 (observer measures of shared decision making). He has in the past provided consultancy for organizations, including: (1) Emmi Solutions LLC who developed patient decision support tools; (2) National Quality Forum on the certification of decision support tools; (3) Washington State Health Department on the certification of decision support tools; (4) SciMentum LLC, Amsterdam (workshops for shared decision making). He is the Founder and Director of &think LLC which owns the registered trademark for Option GridsTM patient decision aids; Founder and Director of SHARPNETWORK LLC, a provider of training for shared decision making. He provides advice in the domain of shared decision making and patient decision aids to: (1) Access Community Health Network, Chicago (Adviser to Federally Qualified Medical Centers); (2) EBSCO Health (Consultant); (3) Bind On-Demand Health Insurance (Consultant), (4) PatientWisdom Inc (Adviser); (5) abridge AI Inc (Chief Clinical Research Scientist). Other authors declare no conflict of interests.

DATA AVAILABILITY

The data set utilized in this study contains patient health information and is not publicly available. This data set can be shared with potential collaborators upon reasonable request to the corresponding author in compliance with in-place institutional policies and protocols to protect the data privacy and intellectual property.

ACKNOWLEGMENTS

The athors would like to thank Sheryl Piper and Roger Arend for their input on the project and thank Lamar Moss for his feedback on the manuscript.

REFERENCES

1. Watson PW, Mckinstry B. A systematic review of interventions to improve recall of medical advice in healthcare consultations. *J R Soc Med* 2009; 102 (6): 235–43.
2. Kessels RPC. Patients' memory for medical information. *J R Soc Med* 2003; 96 (5): 219–22.
3. Jansen J, Butow PN, van Weert JCM, *et al.* Does age really matter? Recall of information presented to newly referred patients with cancer. *J Clin Oncol* 2008; 26 (33): 5450–7.
4. Ley P. Memory for medical information. *Br J Soc Clin Psychol* 1979; 18 (2): 245–55.
5. Wagner EH. Chronic disease management: what will it take to improve care for chronic illness? *Eff Clin Pract ECP* 1998; 1 (1): 2–4.
6. Hibbard JH, Greene J. What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs. *Health Aff (Millwood)* 2013; 32 (2): 207–14.
7. Bayliss EA, Ellis JL, Steiner JF. Barriers to self-management and quality-of-life outcomes in seniors with multimorbidities. *Ann Fam Med* 2007; 5 (5): 395–402.
8. Violan C, Foguet-Boreu Q, Flores-Mateo G, *et al.* Prevalence, determinants and patterns of multimorbidity in primary care: a systematic review of observational studies. *PLoS One* 2014; 9 (7): e102149.
9. Rocca WA, Boyd CM, Grossardt BR, *et al.* Prevalence of multimorbidity in a geographically defined American population: patterns by age, sex, and race/ethnicity. *Mayo Clin Proc* 2014; 89 (10): 1336–49.
10. Bayliss EA, Bayliss MS, Ware JE, *et al.* Predicting declines in physical function in persons with multiple chronic medical conditions: what we can learn from the medical problem list. *Health Qual Life Outcomes* 2004; 2: 47.
11. Condelius A, Edberg A-K, Jakobsson U, *et al.* Hospital admissions among people 65+ related to multimorbidity, municipal and outpatient care. *Arch Gerontol Geriatr* 2008; 46 (1): 41–55.
12. Bopp KL, Verhaeghen P. Aging and verbal memory span: a meta-analysis. *J Gerontol B Psychol Sci Soc Sci* 2005; 60 (5): P223–P233.
13. Brown SC, Park DC. Roles of age and familiarity in learning health information. *Educ Gerontol* 2002; 28 (8): 695–710.
14. Grady CL, Craik FIM. Changes in memory processing with age. *Curr Opin Neurobiol* 2000; 10 (2): 224–31.
15. McCarthy DM, Waite KR, Curtis LM, *et al.* What did the doctor say? Health literacy and recall of medical instructions. *Med Care* 2012; 50 (4): 277–82.
16. Kutner M, Greenberg E, Jin Y, *et al.* The Health Literacy of America's Adults: results from the 2003 National Assessment of Adult Literacy. 2006. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006483.
17. Lehnert T, Heider D, Leicht H, *et al.* Review: health care utilization and costs of elderly persons with multiple chronic conditions. *Med Care Res Rev* 2011; 68 (4): 387–420.
18. Boyd CM, Darer J, Boult C, *et al.* Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *JAMA* 2005; 294 (6): 716–24.
19. Hughes LD. Using clinical practice guidelines in multimorbid older adults—a challenging clinical dilemma. *J Am Geriatr Soc* 2012; 60 (11): 2180–2.
20. Fortin M, Bravo G, Hudon C, *et al.* Relationship between multimorbidity and health-related quality of life of patients in primary care. *Qual Life Res* 2006; 15 (1): 83–91.
21. Fortin M, Lapointe L, Hudon C, *et al.* Multimorbidity and quality of life in primary care: a systematic review. *Health Qual Life Outcomes* 2004; 2: 51.
22. Fortin M, Bravo G, Hudon C, *et al.* Psychological distress and multimorbidity in primary care. *Ann Fam Med* 2006; 4 (5): 417–22.
23. Marengoni A, Winblad B, Karp A, *et al.* Prevalence of chronic diseases and multimorbidity among the elderly population in Sweden. *Am J Public Health* 2008; 98 (7): 1198–200.
24. Moffat K, Mercer SW. Challenges of managing people with multimorbidity in today's healthcare systems. *BMC Fam Pract* 2015; 16: 129.
25. McPhail SM. Multimorbidity in chronic disease: impact on health care resources and costs. *Risk Manag Healthc Policy* 2016; 9: 143–56.
26. Wallace E, Salisbury C, Guthrie B, *et al.* Managing patients with multimorbidity in primary care. *BMJ* 2015; 350: h176.
27. Hummel J, Evans P. Providing clinical summaries to patients after each office visit: a technical guide. 2012. http://hit.qualishealth.org/sites/default/files/hit.qualishealth.org/Providing-Clinical-Summaries-0712.pdf.
28. Emani S, Healey M, Ting DY, *et al.* Awareness and use of the after-visit summary through a patient portal: evaluation of patient characteristics and an application of the theory of planned behavior. *J Med Internet Res* 2016; 18 (4): e77.
29. Pavlik V, Brown AE, Nash S, *et al.* Association of patient recall, satisfaction, and adherence to content of an electronic health record (EHR) - generated after visit summary: a randomized clinical trial. *J Am Board Fam Med* 2014; 27 (2): 209–18.
30. Wolff JL, Darer JD, Berger A, *et al.* Inviting patients and care partners to read doctors' notes: OpenNotes and shared access to electronic medical records. *J Am Med Inform Assoc* 2017; 24 (e1): e166–72–e172.
31. Nazi KM, Turvey CL, Klein DM, *et al.* VA opennotes: Exploring the experiences of early patient adopters with access to clinical notes. *J Am Med Inform Assoc* 2015; 22 (2): 380–9.
32. Gaston CM, Mitchell G. Information giving and decision-making in patients with advanced cancer: a systematic review. *Soc Sci Med* 2005; 61 (10): 2252–64.
33. Weiss L, Gany F, Rosenfeld P, *et al.* Access to multilingual medication instructions at New York city pharmacies. *J Urban Health* 2007; 84 (6): 742–54.

34. Tsulukidze M, Durand M-A, Barr PJ, *et al.* Providing recording of clinical consultation to patients – a highly valued but underutilized intervention: a scoping review. *Patient Educ Couns* 2014; 95 (3): 297–304.
35. Good DW, Delaney H, Laird A, *et al.* Consultation audio-recording reduces long-term decision regret after prostate cancer treatment: a non-randomised comparative cohort study. *Surgeon* 2016; 14 (6): 308–14.
36. Ford S, Fallowfield L, Hall A, *et al.* The influence of audiotapes on patient participation in the cancer consultation. *Eur J Cancer* 1995; 31 (13-14): 2264–9.
37. Scott JT, Entwistle VA, Sowden AJ, *et al.* Giving tape recordings or written summaries of consultations to people with cancer: a systematic review. *Health Expect* 2001; 4 (3): 162–9.
38. McClement SE, Hack TF. Audio-taping the oncology treatment consultation: a literature review. *Patient Educ Couns* 1999; 36 (3): 229–38.
39. Krackow KA, Buyea CM. Use of audiotapes for patient education, medical record documentation, and informed consent in lower extremity reconstruction. *Orthopedics* 2001; 24 (7): 683–5.
40. Santo A, Laizner AM, Shohet L. Exploring the value of audiotapes for health literacy: a systematic review. *Patient Educ Couns* 2005; 58 (3): 235–43.
41. Elwyn G, Barr PJ, Grande SW. Patients recording clinical encounters: a path to empowerment? Assessment by mixed methods. *BMJ Open* 2015; 5 (8): e008566.
42. Tsulukidze M, Grande SW, Thompson R, *et al.* Patients covertly recording clinical encounters: threat or opportunity? a qualitative analysis of online texts. *PLoS One* 2015; 10 (5): e0125824.
43. Barr PJ, Bonasia K, Verma K, *et al.* Audio-/videorecording clinic visits for patient's personal use in the United States: cross-sectional survey. *J Med Internet Res* 2018; 20 (9): e11308.
44. Tang PC, Ash JS, Bates DW, *et al.* Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. *J Am Med Inform Assoc* 2006; 13 (2): 121–6.
45. Bayati M, Braverman M, Gillam M, *et al.* Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS One* 2014; 9 (10): e109264.
46. Hassanpour S, Langlotz CP. Predicting high imaging utilization based on initial radiology reports: a feasibility study of machine learning. *Acad Radiol* 2016; 23 (1): 84–9.
47. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. *Artif Intell Med* 2016; 66: 29–39.
48. Hassanpour S, Bay G, Langlotz CP. Characterization of change and significance for clinical findings in radiology reports through natural language processing. *J Digit Imaging* 2017; 30 (3): 314–22.
49. Huhdanpaa HT, Tan WK, Rundell SD, *et al.* Using natural language processing of free-text radiology reports to identify type 1 Modic endplate changes. *J Digit Imaging* 2018; 31 (1): 84–90.
50. Meng X, Ganoe CH, Sieberg RT, *et al.* Assisting radiologists with reporting urgent findings to referring physicians: a machine learning approach to identify cases for prompt communication. *J Biomed Inform* 2019; 93: 103169. doi:10.1016/j.jbi.2019.
51. Hassanpour S, Langlotz CP. Unsupervised topic modeling in a large free text radiology report repository. *J Digit Imaging* 2016; 29 (1): 59–62.
52. Barr PJ, Dannenberg MD, Ganoe CH, *et al.* Sharing annotated audio recordings of clinic visits with patients—development of the open recording automated logging system (ORALS): study protocol. *JMIR Res Protoc* 2017; 6 (7): e121.
53. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010; 17 (5): 514–8.
54. 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 (5): 524–7.
55. Sohn S, Clark C, Halgrim SR, *et al.* MedXN: an open source medication extraction and normalization tool for clinical text. *J Am Med Inform Assoc* 2014; 21 (5): 858–65.
56. Jiang M, Wu Y, Shah A, *et al.* Extracting and standardizing medication information in clinical text – the MedEx-UIMA system. *AMIA Jt Summits Transl Sci Proc* 2014; 2014: 37–42.
57. Wang Y, Steinhubl SR, Defilippi C, *et al.* Prescription extraction from clinical notes: towards automating EMR medication reconciliation. *AMIA Jt Summits Transl Sci* 2015; 2015: 188–93.
58. Kim M-Y, Xu Y, Zaiane OR, *et al.* Recognition of patient-related named entities in noisy tele-health texts. *ACM Trans Intell Syst Technol* 2015; 6 (4): 1–59.23.
59. Du N, Wang M, Tran L, *et al.* Learning to infer entities, properties and their relations from clinical conversations. *EMNLP-IJCNLP 2019 - 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process Proc Conf* 2020; 4979–90. doi:10.18653/v1/d19-1503.
60. Selvaraj SP, Konam S. Medication regimen extraction from medical conversations. *Stud Comput Intell* 2021; 914: 195–209.
61. Patel D, Konam S, Selvaraj SP. Weakly supervised medication regimen extraction from medical conversations. Published Online First: 11 October 2020. http://arxiv.org/abs/2010.05317. Accessed March 24, 2021.
62. Mani A, Palaskar S, Konam S. Towards understanding ASR Error Correction for Medical Conversations. 2020; 7–11. doi:10.18653/v1/2020.nlpmc-1.2
63. Enarvi S, Amoia M, Del-Agua Teba M, *et al.* Generating medical reports from patient-doctor conversations using sequence-to-sequence models. 2020; 22–30. doi:10.18653/v1/2020.nlpmc-1.4.
64. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
65. South BR, Shen S, Leng J, *et al.* A Prototype Tool Set to Support Machine-assisted Annotation. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012: 130–9. http://dl.acm.org/citation.cfm?id=2391123.2391141.
66. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–D270.
67. Kaufman J. 10,000 most common English words from Google's Trillion Word Corpus. 2018. https://github.com/first20hours/google-10000-english. Accessed August 9, 2021.
68. Herbs and Supplements: MedlinePlus. https://medlineplus.gov/druginfo/herb_All.html. Accessed August 9, 2021.