AHRQ Grant Final Progress Report

Title of Project:

NLP to Improve Accuracy and Quality of Dictated Medical Documents

Principal Investigator:
Li Zhou, MD, PhD
Associate Professor
Brigham and Women's Hospital, Harvard Medical School

Team Members:

Co-Investigators	Foster Goss, Marie Meteer, David Bates
Data Coordinator/Manager	Frank Chang
Postdoc Research Fellows	Liqin Wang, Zfania Tom Korach, Jie Yang
Research Assistants	Kenneth Lai, Suzanne Blackley, Victor Lei, Kevin Lai, Raymond Doan, Warren Acker, Leigh Kowalski, Carlos Ortega, Sharmitha Yerneni, Valerie Schubert, Christine Eckhardt, Wasim Al Assad, Daniel Noar, Jessica Huynh, Erin MacPhaul
Other Collaborators	Adam Landman, Shira Hassid, Maxim Topaz, Evgeni Kontrient, Sheila Guston

Organization: Brigham and Women's Hospital

Inclusive Dates of Project: 09/30/2015-09/30/2019

AHRQ Project Officer: Christine Dymek

Grant Award Number: R01HS024264

Acknowledgement of Agency Support: This study was funded by the Agency for Healthcare Research and Quality grant number R01HS024264.

Structured Abstract

Purpose: This study's goal was to study the nature of speech recognition-generated errors in a large corpus of clinical documents and to develop and evaluate innovative methods for automatic error detection and correction.

Scope: The project was conducted at two large integrated health care systems located in the Boston, Massachusetts and Aurora, Colorado areas.

Methods: We analyzed speech recognition (SR) errors in documents created through multiple common SR workflows, during which we developed comprehensive guidelines for identifying and classifying SR errors. We also studied clinician users' experiences with and perceptions of SR technology and compared perceived error prevalence to the error rates observed in our analysis. We conducted the largest systematic review of SR for clinical documentation to date. Finally, we employed natural language processing and machine learning to develop a system to SR identify errors in notes.

Results: We observed an overall error rate of 7.4% in dictated notes prior to editing. The error rate dropped 0.3% following editing by a transcriptionist and the dictating physicians' final review. Likewise, 96.3% of initial SR transcriptions contained errors, compared to only 42.4% of finalized notes. Our best performing error detection model, based on a statistical language model, achieved an F1 score of 81%. A recurrent neural network-based model achieved an F1 score of 77%, while a topic model-based classifier achieved a lower F1 score of 24%. We found that physicians have generally positive opinions about SR, but multiple issues remain.

Key Words: speech recognition, dictation, documentation quality, electronic health records

Purpose

The goal of this study was to study the nature of speech recognition (SR)-generated errors in a large corpus of clinical documents and to develop and evaluate innovative methods for automatic error detection and correction. An NLP-based approach to error detection can be implemented not only as a post-processor for SR products to improve recognition accuracy, but also as a component of NLP applications to handle errors when used to process SR-generated documents. During the course of this project, we developed innovative methods for identifying and analyzing SR-generated errors through the completion of the following specific aims:

Specific Aim 1: Build a large corpus of clinical documents dictated via SR across different healthcare institutions and clinical settings
Specific Aim 2: Conduct error analysis to estimate the prevalence and severity of SR errors
Specific Aim 3: Develop automated, robust methods to detect SR errors in medical documents
Specific Aim 4: Evaluate the performance of the proposed methods and tool
Specific Aim 5: Distribute our methods and tool

Scope

Background and Context

High-quality, accurate medical documents are critical for effective inter-provider communication and patient care. Documentation is also important to both the patient and the provider when medical errors occur, legal issues arise, or when the patient is viewing their medical record through a personal health record system. Electronic health record (EHR) technology has evolved to offer clinicians a range of input methods, including speech recognition (SR).

There are two main ways in which SR is used to assist documentation: 1) *back-end* SR, in which the physician dictates into a telephone, their voice is transcribed by a remote SR engine, and the resulting transcription is edited by a professional medical transcriptionist and returned to the dictating physician for final review and signature; and 2) *front-end* SR, where physicians dictate directly into free-text fields of the EHR and must edit the resulting transcribed text themselves prior to signing the document. In the time since this study began, front-end SR has become increasingly more wide-spread and is now the primary type of SR used at both study sites.

Physician use of SR has risen in recent years due to its ease of use and efficiency at the point of care. However, high error rates have historically been observed in SR-generated medical text. Preventing SR errors necessitates careful proofreading and editing, a time-consuming task for physicians already feeling overburdened by documentation requirements. As such, an increasing number of errors may be entered into the EHR as a result of this technology, potentially jeopardizing the quality and accuracy of medical documents and, ultimately, patient care.

Most existing efforts to identify and/or correct errors in free-text documentation are designed for typed text and as such are not applicable to text entered via speech recognition software. For example, many of the errors that occur in typed text are non-word errors (i.e., misspellings or typos), which can generally be detected with traditional spellchecking techniques. On the other hand, the errors found in SR-generated text are “real-word” errors (i.e., words that are spelled correctly but are incorrect given the context in which they appear), because SR software cannot transcribe a word that does not exist in its dictionary. Further, SR errors can be divided into two types: 1) words that are incorrect given the local context in which they occur (e.g., "There is no facial Miami" [corrected sentence: "There is no facial hypomimia"]), and 2) words that are incorrect given the broader context of the document as a whole (e.g., "She is on lasix" [corrected sentence "She is on lithium"] in a note about a patient with bipolar disorder). Different methods are therefore required to identify errors present in SR-generated medical documents.

Setting and Participants

The project was conducted at two large integrated health care systems: Partners HealthCare System (PHS) located in the Boston, Massachusetts area, and 2) University of Colorado Health (UCHealth) in the Aurora, Colorado area. Both sites use Epic EHR software. During the study period, both sites transitioned from eScription and Dragon Medical 360, products of Nuance Communications, for SR to Dragon Medical One, also a product of Nuance Communications.

Methods

Data Sources and Collection

Aim 1

For back-end SR, we received 8,000 notes from Nuance Communications that had been dictated using eScription, a back-end SR service. We developed semi-automated methods for aligning the original SR transcriptions with the transcriptionist-edited versions in order to efficiently collect error examples. We also conducted annotation on a separate random sample of 217 notes (described further in Aim 2). For front-end SR, we conducted observations of 15 PHS physicians as they dictated real patient notes, and of 10 PHS physicians as they dictated fictional patient notes developed by our team. All observations and simulations were recorded so that the videos could be analyzed for errors. For the simulated dictations, we also asked the participating physicians to create notes via keyboard and mouse in order to compare the accuracy of SR and typing. More details about the data collection for both the observations and the simulations are provided in the *Study Design and Measures* section. We also received 1,349 front-end SR error examples from staff at Spectrum Health, one of our collaborators.

Aim 2

For back-end SR error analysis, we collected a random sample of 217 notes dictated by 144 providers from both sites (167 notes from PHS and 50 notes from UCHealth) stratified by note type and provider specialty. For each note, we collected three versions: 1) the original, unedited transcription (subsequently referred to as the "SR version"); 2) the note after being edited by a medical transcriptionist ("MT version"; and 3) the final signed note ("SN version"). We also conducted a survey to assess the role of SR technology in clinicians' documentation workflows by examining their use of, experience with, and opinions about this technology. One of the survey's main focuses was on clinicians' perceptions of SR accuracy and the impact of SR errors on their documentation habits and preferences.

Aim 3

All language models were trained on a set of 137,247 office visit notes collected between July 1, 2012 and June 30, 2014 from two outpatient clinics associated with Brigham and Women's

Hospital (BWH), part of PHS. In total, the notes contained 8,227,203 sentences, 79,949,783 tokens (space separated words, digits, punctuation marks, etc.).

Aim 4

Evaluation was conducted on a held-out test set of 598 sentences (10,439 tokens), 293 with errors and 305 without, taken from 97 randomly selected notes dictated by 11 unique providers between January 1, 2017 and April 28, 2017.

Study Design and Measures

Aim 1

For back-end SR, we developed a semi-automated method of aligning sentences and identifying differences between the original SR transcription and its corrected version, allowing us to efficiently extract error examples from collected notes. For front-end SR error analysis, we observed and recorded videos of clinicians as they dictated patient notes using Dragon, an SR system used at PHS, including any changes they made to the transcript, to better understand their use of Dragon and the conditions under which errors are generated during dictation. During each observation, a researcher was present to complete an observation checklist which was used to ensure that all the desired data (e.g., the level of background noise, whether or not the clinician was interrupted or took breaks while dictating) were collected. The researcher also distributed a short demographics survey and conducted a semi-structured interview following the observation.

We conducted 15 observations lasting between 1 and 4 hours each. During each dictation, an iPad was used to record audio of the physician's voice and video of the computer screen plus relevant hand movements, such as typing and mouse-clicks. For each video, we measured the total time needed to create a note and the percentage of time devoted to each of five predefined tasks: 1) speaking/dictating, 2) editing mistakes, 3) typing, 4), navigating with the mouse, and 5) thinking (e.g., mentally preparing for speech, re-reading previously dictated sections). We also analyzed trends and identified correlations within and across the observation checklist and video data with respect to medical specialty, total note time, time spent navigating or typing, speaker accent, time spent editing mistakes made by the SR system, interruptions, and time spent thinking.

Aim 2

For back-end error analysis, we first created a gold standard version for each of the 217 notes. A PharmD candidate or medical student, under the supervision of 2 practicing physicians, created a transcription of the note while listening to the original audio recording and using the MT version as a reference. Chart review was then conducted to validate each note's content (e.g., by checking the patient's medication list to verify a medication name that was partially inaudible in the recording).

We then created an annotation schema for identifying and classifying errors. The schema was developed by a team of clinical informaticians, computational linguists, and clinicians iteratively over multiple annotation rounds. It includes 12 general error types (e.g., insertion, deletion), 14 semantic types (e.g., medication, general English), and a binary classification of clinical significance, where an error was considered clinically significant if it could plausibly change a note's interpretation thereby potentially affecting a patient's future care, either directly (e.g., by influencing clinical decision making or possible treatment options) or indirectly (e.g., by resulting in billing errors or affecting potential litigation proceedings).

Knowtator, an open-source annotation tool, was used to annotate all three versions of each note. Two annotators (1 computational linguist and 1 medical student) independently annotated each note. Notes were then further annotated for the following non-error changes: automatic abbreviation expansion by the SR system, disfluencies (e.g., "um") by the dictating clinician, stylistic changes (e.g., rewording a grammatically incorrect sentence) by the transcriptionist or the signing clinician, rearranging of a note's contents by the transcriptionist or the signing transcription, and addition or removal of a note's content by the clinician prior to signing. Two practicing physicians independently evaluated each error for clinical significance, with disagreements reconciled via discussion.

Inter-annotator agreement was calculated using a subset of 33 notes (7 SR notes and 26 MT notes) which were randomly selected after considering each note version's variations in error complexity; for example, transcriptionists' edits often involve minor rewordings, which must be distinguished from true errors. Agreement was defined as the percentage of errors for which both annotators selected the same general and semantic type. For each error, we required only that the spans of text selected by each annotator overlap with one another to some degree rather than requiring exact span matches. For clinical significance, agreement was defined as the percentage of overlap between the 2 physicians' classifications.

For this portion of this aim, we determined the time required to dictate each note, along with each note's turnaround time (the length of time between completion of the original dictation and when the transcriptionist-revised document was sent back to the EHR) and clinician review time (the length of time between when the transcription was returned to the EHR and when the clinician signed the note). For each version of each note, we analyzed the differences between that note and the corresponding gold standard note. We determined the error rate (i.e., the number of errors per 100 words), the median error rate with interquartile ranges, the mean number of errors per note, the frequency of each error type, and the percentage of notes containing one or more errors. Analyses were conducted twice, once for all errors and once for just those errors deemed clinically significant.

Statistical analyses were conducted using R statistical software with *t*-tests used to identify significant differences in mean error rates at each sage by provider sex and specialty along with note type, with *p* values of less than 5 considered statistically significant. For comparisons involving 2 or more groups (e.g., specialty) each group's mean error rate was compared with that of all other groups combined. We calculated Pearson's correlation coefficient (*r*) to measure the strength of association between error rate and clinician age and between error rate and document length.

For front-end SR error analysis, we analyzed the videos of the recorded dictations to identify errors, error types, and the observed prevalence of errors compared to participants' perceptions about error frequency. Error types were classified based loosely on the error classification schema developed during the back-end error analysis. To further understand SR errors and the conditions in which they occur, we distributed a questionnaire prior to the simulations and conducted interviews following each simulated dictation. The interviews were conducted based on a predefined interview guide consisting of 4 open-ended questions about participants' SR usage habits, experience and preferences.

For the observations, we conducted similar, semi-structured interviews consisting of 18 open-ended questions about participants' SR usage habits, perceptions of SR accuracy, ability to integrate SR into their existing workflows, and general opinions about SR software.

Aim 3

We conducted multiple experiments to test the feasibility of applying both statistical and neural network-based language models to SR-generated transcriptions to identify sentences with errors. Because SR errors typically involve words that are spelled correctly but are incorrect contextually, identifying the specific word(s) involved in the error can be ineffective. We therefore chose to conduct sentence-level classification. Also, for the resulting error detection system, flagging an entire sentence as potentially containing an error provides enough context for the clinician to quickly ascertain whether there is actually an error.

Statistical language models were trained using SRILM, a toolkit for generating and evaluating language models which is freely available for noncommercial use. It provides implementations of state-of-the-art language modeling algorithms, including a wide range of smoothing and discounting methods to help account for words or phrases not found in the dataset on which the model was trained. Because our preliminary experiments showed that the Good-Turing and modified Kneser-Ney algorithms yielded the best performance given our data and task, we included only those models trained using one of these discounting methods in our final experiments. We also used n-grams of length 4 (i.e., all sets of four consecutive words in the document) in all of our models. Other adjustable parameters included whether or not to limit the words included in the model to a predefined vocabulary, whether or not unknown words should be discarded or treated as a separate "unknown word" token, and the frequency below which an n-gram should be treated as unknown (e.g., if a given sequence of words appears only once, it is rare enough that it can be treated as not having occurred at all in order to make the model less computationally expensive).

Recurrent neural networks (RNNs), a type of artificial neural network, were also used to build language models. Specifically, we trained an RNN model using a "long short-term memory" (LSTM) architecture which allows the network to "remember" important information from previous states and use this information when predicting the current state while discarding less important information. Our RNN-based language models were trained using the Keras functional API (François Chollet and others, 2015; http://keras.io). We trained both a standard model and a joint model. The standard model consists of a primary n-gram input (as with the statistical models described above) as well as any number of secondary inputs (e.g., part of speech tags, the section of the note in which the sentence appears, and so on), training on all provided inputs. The joint model takes as input any number of pre-trained standard models and, without any further training of those models, trains on the standard models' predictions. Various n-gram lengths were tested, including unigrams (single tokens), bigrams (2 consecutive tokens), trigrams (3 consecutive tokens), and 4-grams, as well as several combinations of pre-trained n-gram models.

In addition to language models, we also assessed the viability of using topic modeling to identify errors that are incorrect given the broader context of the document (as in the second example in the last paragraph of the *Background and Context* section). Where language models focus on a word's immediate context (i.e., the word(s) immediately preceding it), topic modeling looks at whether a word is likely to occur given the other words in the document, regardless of their order. We conducted latent topic analysis using Gensim, an open-source Python library for vector space and topic modeling (Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. Paper presented at: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks2010, Valletta, Malta; https://github.com/rare-technologies/gensim).

Aim 4

We treated error detection as a classification task, where each sentence either did or did not contain an error. All models were therefore evaluated in terms of their precision, recall, and F1 score. Precision, also called positive predictive value (PPV), is the percentage of correctly identified examples; here, this corresponds to the percentage of sentences that the system flagged as containing errors that in fact did contain errors. Recall, also called sensitivity, is the percentage of relevant examples that the system successfully identified. Out of all sentences in the test set known to contain errors, the recall is the percentage of those sentences that the system classified as erroneous. The F1 score is the harmonic mean of precision and recall.

Other Studies and Measures

We searched 10 scientific and medical literature databases to find articles about clinician use of SR for documentation published between January 1, 1990 and October 15, 2018. Databases searched included PubMed, the Cumulative Index to Nursing and Allied Health Literature, Web of Science, Association for Computing Machinery and Digital Library, IEEE Xplore, ScienceDirect, MEDLINE, the Cochrane Database of Systematic Reviews, PsycINFO, and Scopus. We iteratively built and refined the search statements between April and June of 2017 and subsequently reviewed the references of included articles to identify articles missed in the database searches.

Inclusion in the review required that all of the following criteria be met: 1) the article was written in English, the article included metadata (authors, title, publication year) and an abstract, 3) the article was published between January 1, 1990 and October 15, 2018 (as SR was not widely used for clinical documentation until the late 1980s), and 4) the abstract mentioned speech or voice recognition, a medical setting, and use of SR for documentation or similar purposes. Each included article was independently annotated by two reviewers for its research topic(s), medical domain, and SR system(s) evaluated, if applicable. Each article could be assigned up to three research topics. Disagreements between reviewers were resolved through discussion.

Results

Below we describe our study's major findings. We provide a list of publications and products from this study at the end of the report.

Aims 1 and 2

For back-end error analysis, the original audio recordings contained an average of 507 words (standard deviation [SD]: 296.9) and a median of 446 words (range: 59 to 1,911). The average dictation duration was 5 minutes and 46 seconds, with a median dictation duration of 4 minutes and 45 seconds (range: 21 seconds to 31 minutes and 35 seconds). The average turnaround time was 3 hours and 37 minutes, with a median turnaround time of 1 hour and 1 minute (range: 2 minutes to 38 hours and 45 minutes). The average clinician review time was 4 days, 13 hours, and 16 minutes, with a median clinician review time of 23 hours and 25 minutes (range: 0 minutes to 146 days, 4 hours, and 54 minutes).

For the 329 errors in the 33-note subset on which inter-annotator agreement was calculated, agreement was 71.9% for errors identified by both annotators. Each annotator failed to identify an average of 21.5% of errors identified by the other. Of 158 errors identified by only one annotator, 32 (20.3%) involved clinical information, while the remaining 126 (79.7%) involved minor changes to general English words. Agreement for clinical significance was 85.7%.

Errors were prevalent in SR transcriptions, with an overall error rate of 7.4%, or 7.4 errors per every 100 words dictated. The error rate dropped substantially following revision by professional medical transcriptionists, falling to 0.4%, and was further reduced in signed notes, with an overall error rate of 0.3%. The percentage of notes that contained at least one error also dropped at each stage, from 96.3% of SR notes, to 58.1% of MT notes, and finally to 42.4% of SN notes.

The effect of human review on note accuracy becomes more pronounced when considering just those errors that were clinically significant. Prior to human revision, 63.6% of notes had at least one clinically significant error; after revision by a transcriptionist, only 14.7% of notes had clinically significant errors, and only 7.8% of signed notes had clinically significant errors. However, the proportion of errors involving clinical information increased from 15.8% to 26.9% after transcriptionist review then fell slightly to 25.9% in signed notes. Similarly, the proportion of errors that were clinically significant increased from 5.7% in the original SR transcriptions to 8.9% after revision by a medical transcriptionist, then fell to 6.4% in signed notes.

At all processing stages (SR, MT, and SN), deletions were the most prevalent general error type, followed by insertions; the most frequent semantic type was general English. Medication was the most common clinical semantic type in original SR transcriptions, while diagnosis was most common in the transcriptionist-edited and signed versions.

For front-end error analysis, for the simulated notes, typed notes contained more uncorrected errors than dictated notes (57 vs. 30; p = 0.13). Typed notes also contained more corrected errors than dictated notes (678 vs. 82; p < 0.001), likely because many typing errors were minor typos that are both common and easy to correct. All 26 uncorrected errors caused by the SR system involved non-medical words, such as pronouns (e.g., he, she), articles (e.g., a, the), general English words, punctuation, negation, and so on. Of 37 corrected errors caused by the SR system, 8 (21.6%) involved medical terms such as diagnoses, procedures, or medication names.

In the post-simulation interview, all ten participants agreed that SR increases their efficiency and the accuracy of their documentation. The most common error types according to physicians were names (mentioned by 8 of 10 participants), medication names (n=6), grammatical errors (n=4), numbers (n=4), and pronouns (n=4). However, the most common errors observed in the video recordings were short words pronouns/articles (20 out of 63), general English (n=16), and medical terms/abbreviations (n=6).

For the observations, eight participants estimated that their SR system makes errors on only between 1% and 10% of words, while one estimated between 11% and 25% of words were transcribed incorrectly, and one estimated that more than 50% of words are transcribed incorrectly. The remaining 5 participants did not answer the question. When asked about what percentage of their documentation time is spent editing or correcting errors, eight participants answered 1-10% and four answered 11-25%. Of the remaining 3 participants, one estimated spending 1 minute per patient correcting errors, and another estimated spending 3 minutes per patient. The last participant did not answer this question.

For the clinician survey, we received responses from 348 out of 1,731 clinicians surveyed (20.1%), of whom 108 were excluded (reasons: only used back-end dictation, n = 72; did not use dictation, n = 12; did not complete the survey, n = 18; specialized in radiology, n = 1 [radiologists were excluded due to the field's unique workflow and early SR adoption]). The remaining 245 responses (95 from BWH and 150 from UCHealth) were included in our analysis. Most respondents were 35-54 years old (66.9%), white (84.4%), physicians (81.6%), native English speakers (93.1%), and received medical education in English (96.7%). Almost half (47.3%) specialized in general medicine, followed by emergency medicine (20.8%) and surgery (13.1%).

Perceived error incidence was low among respondents, with 43.7% reporting observing 5 or fewer errors per document and only 7.3% reporting observing more than 20. Most respondents (69.0%) estimated that only 25% or fewer errors would be considered clinically significant. Overall, 21.1% of respondents estimated spending 25% or more of their documentation time editing. Respondents had mixed opinions when asked if they had noticed improved accuracy with the latest version of the SR software: respondents strongly agreed, agreed, and felt neutral in roughly equal numbers, while a few strongly disagreed.

Users reporting more than 20 errors per document were 14 times more likely to be dissatisfied with their SR system than those reporting 5 or fewer. Native language was significantly associated with editing time, with native English speakers reporting less editing time than native speakers of other languages, as was specialty, with respondents specializing in emergency or general medicine reporting less editing time compared to those specializing in surgery.

Eighty-five respondents (34.7%) entered comments in the free-text field available at the end of the survey. Only 11 comments were positive, applauding improved accuracy, mobile phone integration, and increased time for patient care. The remaining comments were negative, referencing usability and technical issues, inadequate availability, and dictation errors. While most users still preferred dictation to typing, many desired improved SR accuracy, which they felt would increase usability. Some felt that front-end SR has not reduced documentation burden, as it requires clinicians to edit documents themselves, unlike workflows that involve editing by professional medical transcriptionists (i.e., back-end SR).

Aims 3 and 4

In general, the different model configurations for both the statistical and the RNN-based language models were comparable in their ability to detect errors and tended to have higher recall than precision. For the statistical language models, of the 8 model configurations tested, the best performing model achieved a precision of 72%, a recall of 93%, and an F1 score of 81%. For the RNN-based language models, the best performing model achieved a precision of 66%, a recall of 93%, and an F1 score of 77%. These findings demonstrate that use of language models for error detection is a promising avenue of research, but that further work is needed to reduce false positives so that they do not contribute to the already serious problem of alert fatigue. Additionally, we classified errors at the sentence level, but future work focusing on identifying the specific word(s) involved in an error may be useful.

For the topic models, of the five models we tested, the best performing model achieved a precision of 26%, a recall of 23%, and an F1 score of 24%. Although the topic models did not perform as well as the language models, this reflects the fact that detecting semantic errors is, in general, a more complex task than detecting syntactic errors. Topic models may prove more successful if they can be specifically tailored to certain medical domains. For example, when we tested a topic model trained on office notes on its ability to detect errors in rehabilitation notes, the best performance was only 14%, whereas the performance improved to 24% when the model was both trained and tested on rehabilitation data.

We submitted an Invention Disclosure Form to Partners HealthCare Innovation in which we declared our intent to develop an invention based on our error detection methods, however the invention has not been completed and registered at this time (as reflected in the Final Invention Statement and Certification Form).

Other Findings

For the systematic review, out of 1,343 records retrieved, 122 articles were included in the analysis. Most articles (89.3%) were published in or after the year 2000, with the annual number of articles fluctuating and peaking approximately every 7-8 years. The largest proportion of studies were conducted in the radiology department (39.3%), followed by emergency medicine (8.2%) and nursing (6.6%). Most research topics, such as comparison to transcription, error analysis, SR use and impact on clinical workflows, and SR implementation, we restudied throughout the review period. Since 2009, more studies have involved user surveys and interviews. The most common research topic was documentation time or cost and productivity analysis (39.3% of articles), followed by impact on clinical workflow (28.7%), error analysis (23.8% of articles), and comparisons between or assessments of the combination of SR-assisted dictation and traditional dictation and transcription (20.5% of articles). In general, articles within the same research topic employed a variety of methods and measures, making it challenging to compare findings across studies and over time. For example, most error analysis studies reported word error rate (i.e., the percentage of incorrect words), document error rate (i.e., the number of documents containing one or more errors), or mean number of errors per document, but rarely all three.

Publications and Products

1. Goss FR, Zhou L, Weiner SG. Incidence of speech recognition errors in the emergency department. *International Journal of Medical Informatics*. 2016 Sep 1;93:70-3.
2. Zhou L, Landman AB, Kontrient E, Doan R, Blackley SV, Mack D, Bates DW, Goss FR. An Error Analysis of Dictated Clinical Documents at Different Processing Stages. Paper presented at: American Medical Informatics Association 2016.
3. Zhou L, Blackley SV. Improving Health IT Through Use of NLP/AI in Documentation. Presented at: Healthcare Information and Management Systems Society 2018.
4. Zhou L, Blackley SV, Kowalski L, Doan R, Acker WW, Landman AB, Kontrient E, Mack D, Meteer M, Bates DW, Goss FR. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. *JAMA Network Open*. 2018 Jul 6;1(3):e180530-. [Viewed 25,181 times with an Altmetric score of 210]
5. Topaz, M, Schaffer, A, Lai, KH, Korach Z, Einbinder J, Zhou L. Medical malpractice trends: errors in automated speech recognition. *Journal of Medical Systems* (2018) 42: 153. https://doi.org/10.1007/s10916-018-1011-9.
6. Blackley SV, Huynh J, Wang L, Korach Z, Zhou L. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the American Medical Informatics Association*. 2019 Feb 11;26(4):324-38.
7. Goss FR, Blackley SV, Ortega CA, Kowalski LT, Landman AB, Lin CT, Meteer M, Bakes S, Gradwohl SC, Bates DW, Zhou L. A clinician survey of using speech recognition for clinical documentation in the electronic health record. *International Journal of Medical Informatics*. 2019 Oct 1;130:103938.
8. SUBMITTED: Blackley SV, Schubert VD, Goss FR, Al Assad W, Garabedian PM, Zhou L. Physician use of speech recognition versus typing in clinical documentation: a controlled observational study.
9. IN PREPARATION: Blackley SV, MacPhaul E, Noar D, Korach Z, Meteer M, Goss FR, Zhou L. Applying language models to identify errors in clinical notes created using speech recognition.
10. IN PERPARATION: Kowalski LT, Fowler SA, Hassid S, Blackley SV, Schubert VD, Bates DW, Zhou L. The integration of a speech recognition system in clinical documentation and workflow: a qualitative study.