Asynchronous Speech Recognition Affects Physician Editing of Notes

Kevin J. Lybarger[1] Mari Ostendorf[1] Eve Riskin[1] Thomas H. Payne[2] Andrew A. White[2] Meliha Yetisgen[3]

[1] Department of Electrical Engineering, University of Washington, Seattle, Washington, United States
[2] Division of General Internal Medicine, University of Washington, Seattle, Washington, United States
[3] Department of Biomedical & Health Informatics, University of Washington, Seattle, Washington, United States

Address for correspondence Kevin J. Lybarger, MS, Department of Electrical Engineering, University of Washington, Campus Box 352500, 185 Stevens Way, Paul Allen Center–Room AE100R, Seattle, WA 98195-2500, United States (e-mail: lybarger@uw.edu).

Appl Clin Inform 2018;9:782–790.

CME/MOC-II*

Abstract

Objective Clinician progress notes are an important record for care and communication, but there is a perception that electronic notes take too long to write and may not accurately reflect the patient encounter, threatening quality of care. Automatic speech recognition (ASR) has the potential to improve clinical documentation process; however, ASR inaccuracy and editing time are barriers to wider use. We hypothesized that automatic text processing technologies could decrease editing time and improve note quality. To inform the development of these technologies, we studied how physicians create clinical notes using ASR and analyzed note content that is revised or added during asynchronous editing.

Materials and Methods We analyzed a corpus of 649 dictated clinical notes from 9 physicians. Notes were dictated during rounds to portable devices, automatically transcribed, and edited later at the physician's convenience. Comparing ASR transcripts and the final edited notes, we identified the word sequences edited by physicians and categorized the edits by length and content.

Results We found that 40% of the words in the final notes were added by physicians while editing: 6% corresponded to short edits associated with error correction and format changes, and 34% were associated with longer edits. Short error correction edits that affect note accuracy are estimated to be less than 3% of the words in the dictated notes. Longer edits primarily involved insertion of material associated with clinical data or assessment and plans. The longer edits improve note completeness; some could be handled with verbalized commands in dictation.

Conclusion Process interventions to reduce ASR documentation burden, whether related to technology or the dictation/editing workflow, should apply a portfolio of solutions to address all categories of required edits. Improved processes could reduce an important barrier to broader use of ASR by clinicians and improve note quality.

Keywords

- electronic health records and systems
- clinical documentation and communications
- natural language processing
- notes
- workflow

Background and Significance

Clinical documentation is a critical component of patient care, and communicating accurately and comprehensively through clinical notes is important to achieving positive health outcomes. Creating notes within electronic health record (EHR) systems is time-consuming, affects documentation accuracy, negatively affects the career satisfaction of clinicians, and causes lost labor productivity.[1–5] Dictation using transcriptionists and automatic speech recognition (ASR) has the potential to improve

* To earn credit, visit AMIA for details.

received
January 26, 2018
accepted after revision
August 26, 2018

DOI https://doi.org/10.1055/s-0038-1673417.
ISSN 1869-0327.

the note creation process, but these methods entail significant trade-offs.[6,7] Human transcription is expensive and typically delays document availability by hours to days. This delay prevents the timely use of clinical notes by other clinicians and by point-of-care clinical decision-support systems that utilize natural language processing (NLP). Conversely, ASR has shorter document turn-around-time (delay between dictation and transcript availability) and lower marginal costs than human transcription.[8] ASR is commonly used for speech transcription, and it has a long history in the clinical setting.[6] However, broader use of ASR in the clinical setting is limited by editing time, uncaught errors, as well as administrative overhead.[9] ASR lacks the accuracy of traditional dictation,[10] requiring clinicians, rather than transcriptionists, to either spend time editing ASR transcripts or accept flawed documentation.[11] Interactive editing and dictation with ASR can provide increased speed, longer notes, and improved physician mood over typing alone.[12] Unfortunately, current real-time editing tools for ASR are not compatible with handheld devices, and thus are difficult for clinicians to incorporate in inpatient settings.

Previous work within the clinical setting has evaluated ASR note quality (e.g., accuracy), turn-around time, and documentation time.[6,10,11,13–16] Additionally, studies have evaluated the prevalence and clinical significance of ASR errors within dictated notes and explored mechanisms for detecting ASR errors.[17,18] A 2008 study of radiology reports created using ASR found that 22% of the finalized notes contained "potentially significant errors,"[19] while a 2017 study involving ASR-created radiology reports found 1.9% of the notes contained "material" errors.[20] However, performance for other specialties is less well studied. The existing literature has not yielded a sufficiently clear understanding of how clinicians edit ASR transcripts to inform the development of technologies and workflows that overcome these barriers.

In this work, we studied physician editing of clinical notes created using a commercially available ASR system in a noninteractive (asynchronous) setting. The study utilized ASR in portable wireless recording devices, untethering physicians from workstations and allowing mobile dictation while on rounds. Physicians dictated to a recording device, the dictation was automatically transcribed, and they later edited the transcripts at a workstation, at their convenience. The study focused on inpatient progress notes because the short hospital stay and patient acuity make note timeliness even more important than in the outpatient setting. In both the inpatient and outpatient setting, physicians feel progress notes take too long to create.[1,21,22]

We compared ASR transcripts with the associated final edited notes to determine edited regions. The resulting corpus of edits was analyzed to identify opportunities for improving the note creation process. Our hypothesis was that many edits could be automated, which would reduce physician editing time. In this article, we present findings related to the types of edits observed, which included short modifications, as well as longer edits associated with a continuation of the note creation process. Possibilities for automation are discussed.

Objective

The objective was to understand how physicians use ASR to create clinical notes, including how much of the finalized note is created while editing the ASR transcripts and characteristics of the edits. This objective is motivated by the longer-term goal of improving the note creation process. This article focuses solely on the notes generated using ASR with physician editing in an asynchronous setting and does not explore physician satisfaction.

Materials and Methods

Setting and Participants

The clinical notes analyzed in this work were created through the voice-generated enhanced electronic note system (VGEENS) study, which was an effort to improve timeliness, quality, and physician satisfaction.[23–25] The VGEENS study developed and implemented a process for creating inpatient progress notes for patients hospitalized on the internal medicine service of two teaching hospitals (University of Washington Medical Center and Harborview Medical Center).[23–25] Study participants were internal medicine physicians, including resident physicians and attendings (but not nurses, medical students, or other health professionals). The study included an unblinded randomized controlled trial, consisting of: a control group of physicians that created progress notes by typing text into a note template (typical approach at these medical centers) and an intervention group that used the VGEENS to create notes. The VGEENS used a commercial ASR software, Dragon Medical Edition 2 by Nuance Inc., in a noninteractive manner as described above. The ASR software was not configured to utilize verbalized commands for abbreviations, due to the limited duration of the study and a desire to avoid ambiguous abbreviations.[26] Both medical centers use the Cerner Millennium Powerchart electronic medical record, and physicians only have access to human transcriptionists for some admission, discharge, and operative notes.

Physicians in the intervention group created their notes after the patient visit on hospital rounds by dictating the note to a Wi-Fi-connected recording device, including section headings (e.g., "Assessment and Plan") and verbalized punctuation. The privacy and confidentiality of dictation in open spaces were managed using the same internal policies that govern provider conversations. The recorded audio was transmitted to a server where it was transcribed using ASR, and the transcription was postprocessed and placed in the EHR inbox. Postprocessing routines automatically formatted section headings, converted the note to rich text format, and executed custom verbalized commands for automatically inserting clinical data related to blood counts, vitals, liver function, electrolytes, and coagulation studies. For example, the verbalized command, "Please insert vitals" was automatically replaced with the applicable test result, "Vitals: T 36.4, P 77, RR 16, BP 119/89 (02/07/16 12:13)." After postprocessing, the physician used EHR tools to manually edit and format the transcript as needed to create the final EHR

progress note and sign it. We refer to the output of the automatic postprocessing as the "dictated note" and the signed note in the EHR as the "final note." Development of the postprocessing routines was not completed at the start of the trial. They were rolled out during the study, in the following order: formatting of section headings, conversion to rich text format, and inclusion of verbalized commands. All participants were provided written instructions regarding the use of verbalized commands. Technical details of this system are described elsewhere.[24]

Data

We analyzed the corpus of notes created by the VGEENS intervention group using ASR. The intervention corpus consists of 669 dictated final note pairs created by 15 physicians. Only notes created by physicians who dictated at least 10 notes were used in this study, resulting in a corpus of 649 notes created by 9 physicians.

The text of each note was split into a sequence of tokens based on whitespace and punctuation, where tokens consisted of words, punctuation symbols, and line breaks. To identify edited regions, each dictated note was automatically aligned with the corresponding final note using Gestalt pattern matching.[27] This alignment algorithm recursively finds the longest sequence of matching tokens within the note pairs, then the next longest sequence of matching tokens to the left and right, and so forth. Capitalization of tokens was ignored during the alignment of the note pairs, and alignment spans were not constrained by sentence boundaries or line breaks. The original audio file was neither hand transcribed nor archived, so it was not possible to distinguish between ASR-related edits and other edits. In addition, any errors in the dictated notes that were unchanged in the final notes were not captured by this analysis.

Categorization of Edit Spans

To analyze editing practices, we developed a two-level taxonomy of edits using labels produced by the alignment algorithm (equal, replace, delete, insert). At the top level, categories were designed for estimating the relative amount of editing associated with transcript correction versus continued note creation. Subcategories were defined for analyses designed to inform future work on editing tools.

Top-Level Categorization of Spans

Word tokens associated with equal spans are labeled in both versions of the notes as *original.* Any spans in the final notes associated with verbalized commands in the dictated notes were categorized as *command*; other unequal spans were categorized as either a *short edit* or a *continuation.* The *short edit* category was intended to include edits associated with: (1) the correction of speech recognition errors, (2) correction of disfluencies (repeated phrases, such as "the the," or self-corrections, such as "the left leg, correction the right leg"), (3) incorporation of standard phrasing and formatting, and (4) other minor rephrasing. The *continuation* category was intended to reflect edits associated with a continuation of the note creation process, including the removal of dictation that was no longer relevant or the incorporation of information that was not dictated. Edit spans having four or fewer tokens in both dictated and final spans were categorized as *short edits*. All other edit spans were categorized as *continuation.*

The threshold of four tokens for identifying *short edits* was chosen based on an analysis of ASR errors in a separate set of eight dictated notes, where two physicians each dictated (read) the same two note templates (ground truth) twice. We aligned the recognized and ground truth transcripts as described above to identify word differences. The word differences were treated as errors, though a few differences were probably due to reading errors. The ASR error rate was approximately 5%, and 98% of all error spans involved four or fewer consecutive tokens.

Continuation spans separated only by a single-token *equal* span consisting of punctuation or a line break were merged to form a single *continuation* span. This merging was necessary due to the prevalence of punctuation and line break tokens within the notes, which leads to incorrect splits of some long edits. Merged *continuation* spans with different span labels (e.g., *delete* and *insert*) were labeled *replace.*

►Table 1 contains alignment examples with the associated edit type and span label. Deleted tokens are denoted by strikethrough font and inserted tokens are denoted by

Table 1 Examples of edit types and span labels

No.	Edit type	Span label	Example
1	Short edit	Replace	Pain in the right shoulder and upper extremity is fairly well-controlled, improves with the ~~novolin lower~~ now lowered dose of dilaudid 0.1 mg via patient-controlled analgesia
2	Short edit	Insert	I suggested transitioning to oral pain medication today, but he would like to wait until tomorrow
3	Continuation	Replace	He denies lightheadedness ~~when sitting up from a laying position or going from a seated to standing position~~ associated with positional changes
4	Continuation	Delete	1. Depression, worse in the setting of progressive cancer and having no family in the area, continue lexapro ~~pending visit from the psychiatric clinical nurse specialist for additional emotional support~~
5	Continuation	Insert	... liver cirrhosis with ascites and edema. The ultrasound showed nodular appearance to the liver...

underline font. In this table, capitalization is removed. The first example is likely due to an ASR error. The second example may be related to an ASR error but could simply be an editing addition. The third and fourth examples illustrate how the *continuation* spans tend to reflect continued note creation. The fifth example includes two *continuation* spans separated by a single-token *equal* span, as an example of spans that were merged.

Distinguishing between *short edit* and *continuation* spans based on word sequence length is imperfect. For example, physicians inserted the sentences "No clubbing or cyanosis" and "No nausea or vomiting" during editing. The spans were labeled as a *short edit*, even though they incorporated information not originally dictated and thus were part of continued note creation. *Short edits* also include changes associated with formatting and nomenclature, such as deleting nonstandard section headings (e.g., "CODE STATUS":), inserting standard section headings (e.g., "ASSESSMENT & PLAN":), and replacing dictated phrases with acronyms (e.g., "present on admission" → "POA"). Our inspection of roughly 200 *short edit* and *continuation* spans indicated these types of mislabeling are infrequent. Similarly, some of the long *replace* edits involve simple rephrasing, which might not be considered continued note creation. However, less than 2% of the *continuation* spans had a similar number of dictated and final tokens (length difference less than 25% or fewer than 3 tokens), so the *replace continuation* edits reflect more changes than would be associated with formatting or ASR errors, as in the examples in ►Table 1. Most of the results in this article are presented in terms of token count, rather than span count, so we do not expect the infrequent mislabeling of shorter spans to significantly affect either physician-level or corpus-level results.

Subcategorization of Edits

Continuation spans are subcategorized according to whether they are associated with the dictated note versus the final note. Dictated *continuation* spans that are aligned as delete or replace are categorized as *omit* content, and final note *continuation* spans that are aligned as insert or replace are categorized as *new* content. The average contribution of these categories to each note pair was calculated for the major topical sections to determine which are most affected by continued editing. Additionally, we investigated the possibility that *new* spans reflect text copying from previous notes by comparing sequential note pairs created within 48 hours of each other for the same patient and physician (110 note pairs created by 7 physicians).[a] Copied text was identified by aligning the *new* spans in the current note with the text of the previous note using Gestalt pattern matching.

We defined subcategories of *short edits* motivated by how the edit might be detected and whether the category could be automatically extracted by word look-up tables and pattern matching tests. ►Table 2 describes the *short edit* categories.

[a] The number of note pairs for which copying would have been possible is low because not all patient notes are captured in the intervention data set.

Table 2 Short edit categories

Short edit category	Description
Equivalent	Equivalent dictated and final word sequences (e.g., "rehab" versus "rehabilitation" or "present on admission" versus "poa")
Punctuation	All tokens were punctuation or line break (e.g., comma insertion)
Numbers	Word sequences include at least one number
Medical terms	Word sequences include at least one medical term
Gender	Different gender in dictated and final word sequence ("she" versus "he")
Function words	All tokens were function words (e.g., "and" versus "in," insertion of "but")
Negation	At least one negative inserted or deleted (e.g., insertion of "not")
Other edits	Edits that did not fall into previous categories

The categories are listed in order of the pattern matching tests, so a "she" versus "he" substitution was classified as a gender edit rather than a function word edit. The edit types expected to represent clinically significant errors include: *numbers*, *medical*, *gender*, and *negation*. In counting *short edits*, we follow conventions used in ASR, including deletions from the dictated note, and insertions and replacements (substitutions) from the final (reference) note. However, the standard ASR word error rate (WER) formula is normalized by the number of reference tokens, which would be misleading for documents with large amounts of *continuation* editing. Thus, we adapted the WER formula to calculate the *short edit* rate (SER) as: $SER = (D_D + I_F + R_F)/N_D$, where D_D is the number of dictated tokens deleted, I_F is the number of final tokens inserted, R_F is the number of final tokens in replacement spans, and N_D is the total number of dictated tokens.

Because capitalization was ignored in alignment, some *equal* spans involved capitalization edits. Two types of capitalization edits are distinguished based on whether or not they were associated with the start of a sentence, for example, for words occurring after a period, line break, or colon. Examples of sentence internal capitalization edits include acronyms (all upper case) and medical specialties.

Results

Note Composition

The average token count of the dictated and final notes from the intervention group was 779 (min = 18, max = 2,831, standard deviation [SD] = 495) and 1,151 (min = 62, max = 4,061, SD = 633), respectively. In a random sample of 101 manually typed notes from the control group, the average note length was 985 tokens (min = 228, max = 1,650, SD = 343). All subsequent results pertain only to the intervention group

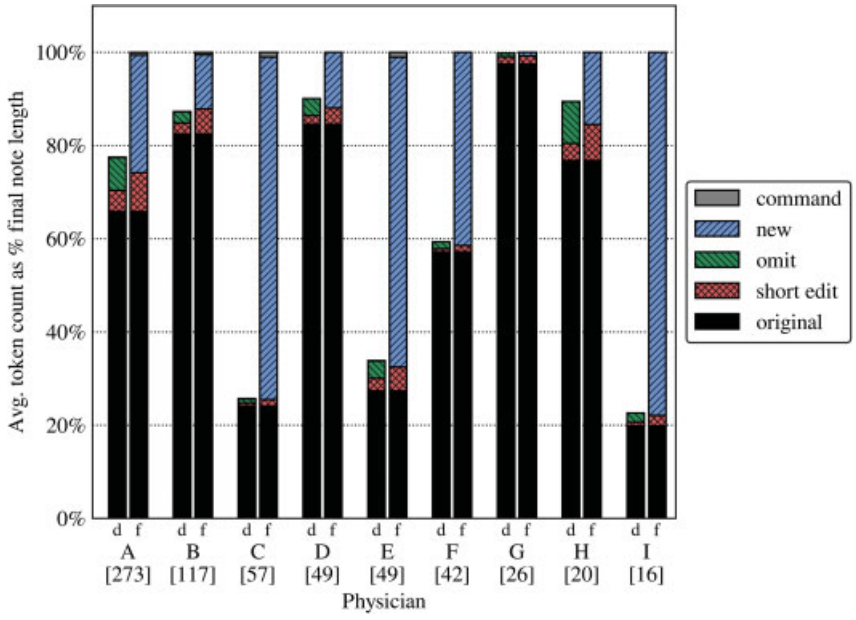

Fig. 1 Breakdown of dictated (d) and final notes (f) by physician, in terms of unedited dictation (*original*), revised dictation (*short edit*), omitted dictation (*omit*), new content (*new*), and content inserted through verbalized commands (*command*). *omit* and *new* categories are associated with *continuation* edits. The note breakdowns are based on average of token counts normalized by the average length of the physician's final notes. The number of notes created by each physician is indicated by the bracketed numbers under the physician labels.

notes. Within the final notes, 60% of the words (691 tokens) in the final note were from *original* spans, and 40% (460 tokens) were manually inserted (34% from *continuation* spans and 6% from *short edit* spans, corresponding to 391 and 69 tokens, respectively). On average, 13% of the dictated tokens (101 tokens) were deleted or replaced during editing. Fewer than 2% of *continuation* spans had a similar number of dictated and final tokens and were of medium length (5–7 tokens), suggesting that at least 98% of the *continuation* spans are unlikely to be simple rephrasing.

►**Fig. 1** contains a breakdown of the dictated and final notes by physician in terms of the span labels. The *short edits* in the dictated bar are words that were deleted or replaced during editing. We observed three primary note creation strategies: (1) dictated a majority of the note with heavy editing of transcription (physicians A, B, D, F, H), (2) dictated a minority of the note with significant additions during editing (physicians C, E, I), and (3) dictated the entire note with very little editing (physician G).

We did not find a simple relationship between the number of notes created by a physician and the percentage of the final note that was dictated. This observation suggests that increased experience with the ASR system did not result in physicians dictating a larger portion of the note.

Manually Inserted Information

Manually inserted information incorporates new information and findings that were not dictated. ►**Fig. 2** contains a breakdown of the dictated and final notes by topical note section in terms of average token count. The most editing, including *short edit* and *continuation* spans, occurs in the Assessment & Plan section, with 53% of the *short edit* tokens in the final notes (average of 37 tokens per note) and 58% of the *new continuation* tokens (average of 227 tokens per note) in the final notes occurring in this section. Approximately 26% of the *new* tokens (average of 102 tokens per note) were inserted in sections that typically incorporate patient clinical data available through the EHR (e.g., test results), including the ID/Chief Complaint, Scheduled Meds, PRN Meds, Physical Exam, Labs, Microbiology, and Imaging sections. The Physical Exam and Labs sections accounted for 17% of the *new* tokens (average of 66 tokens per note).

The insertion of new information during editing raises the question of whether the dictated intervention notes are more informative than the control notes. The analysis of the VGEENS dictated intervention notes and manually typed control notes from Payne and colleagues[23,25] included measures of note quality using the 9-Item Physician Documentation Quality Instrument (PDQI-9).[28] These studies found no significant differences in the PDQI-9 scores between the VGEENS intervention and control notes.

Verbalized Commands

Of the nine participants, four physicians used VGEENS before the verbalized command functionality was available. After the introduction of verbalized commands, four physicians used the commands and one physician did not utilize the functionality. The physicians who used verbalized commands (physicians A, B, C, and E) collectively created 371 notes without verbalized commands and 125 notes with verbalized commands. All of the results automatically inserted through verbalized commands were longer than four tokens. In the subset of notes that included verbalized commands, an average of 32 tokens was automatically inserted per note within the Physical Exam and Labs sections. The average number of tokens added in manual editing

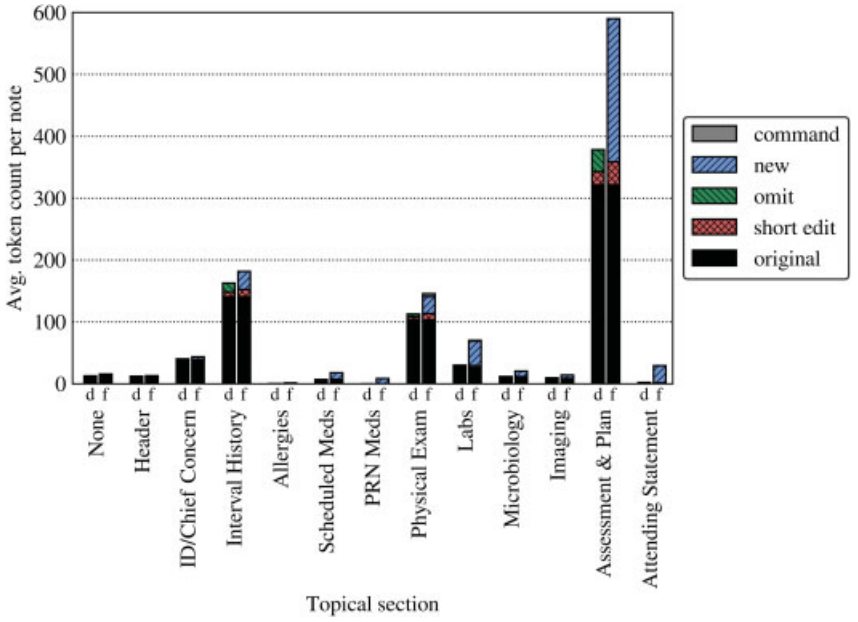

Fig. 2 Breakdown of dictated (column d) and final notes (column f) by note section, in terms of unedited, dictation (*original*), revised dictation (*short edit*), omitted dictation (*omit*), new content (*new*), and content inserted through verbalized commands (*command*). *omit* and *new* categories are associated with *continuation* edits. The note breakdowns are presented in terms of average token counts.

decreased from 66 to 63 for *continuation* spans, and the average final length of these sections increased from 226 to 246 tokens (9% increase).

New Spans

We investigated two possible reasons for the *new* spans: copying text from previous notes, and availability of additional information associated with a delay between note creation and editing. In the notes that included copied text, an average of 141 tokens per note was copied from the previous note, with 86% of the copied tokens occurring in the Assessment & Plan section. Within the Assessment & Plan section, 62% of the tokens added through *continuation* spans were copied from the previous notes, and these accounted for 42% of the tokens added through *continuation* spans in the entire note. The copying was primarily utilized by three physicians, who accounted for 96% of the copied text of the Assessment & Plan section.

We hypothesized that a delay between patient visit and dictation might increase the percentage of the final note that is dictated, because more information (e.g., new laboratory results or consultation with other physicians) would be available during dictation. We also hypothesized that a delay between dictation and note editing (note finalization) might decrease the percentage of the final note that is dictated, because more information would be available during editing than during dictation. Trends supporting these hypotheses were observed for some physicians, but overall there were no significant differences attributable to delay.

Short Edit and Capitalization Edit Distribution

There are approximately 53 *short edit* spans per note (SD of 45 *short edit* spans per note) corresponding to an average 38 dictated tokens (5% of dictated note) and 71 final note tokens (6% of the final note). The large variance reflects differences in note length and note creation strategies. Four *short edit* types were frequent in the data: punctuation (44%), medical terms (23%), function words (9%), and numbers (5%). The *short edit* types that were most likely to be associated with errors that have a significant effect on note accuracy include numbers, medical, gender, and negatives. Of these, gender and negatives only account for 1% of the *short edit* spans. In contrast to the span level (29%), the *short edits* that affect note accuracy account for 58% of *short edit* tokens in the dictated notes, corresponding to an average of 22 dictated tokens (3% of the dictated note). For all *short edits*, the SER was 10%, and for the subset of *short edits* that affect accuracy, the SER was 4%. Physicians A and B created 60% of the notes and collectively account for 87% of the *short edit* tokens in the dictated and final notes. Physicians edited the capitalization of an average of 6 tokens per note: 51% were associated with sentence boundaries and 49% were sentence internal.

Discussion

Most of the physicians in the study heavily edited the dictated notes, making both *short edits* and *continuation* edits. *Short edits* represent 81% of the edit spans, and they may be an important source of frustration for physicians because locating shorter edits requires time and attention to detail. The SER for all *short edits* was 10%. While many of these edits reflect format changes as opposed to actual errors, the finding suggests that the ASR error rate is higher than the advertised rate of 1%,[29] consistent with recent real-world trials of ASR demonstrating higher error rates and time demands compared with manual entry.[7] The results

indicate there are on average 22 *short edit* tokens per dictated note (3% of the dictated note) that affect note accuracy, suggesting that careful review of the dictated notes is required to avoid medically significant errors. There were an additional 6 capitalization edits per note.

If we count changes in the document by number of words, the *continuation* edits dominate. We found that clinical note creation using ASR in a noninteractive, asynchronous setting involved a substantial amount of editing by physicians—beyond what would be expected from edits associated with ASR errors, disfluencies, and formatting. Within the VGEENS framework, physicians continued the note creation process during editing, adding 34% of the final note tokens through *continuation* spans. For some physicians, the *new* content appeared to be pasted in and was primarily in the Assessment & Plan section. A smaller number of inserted tokens was associated with physical exam and laboratory results. Physicians may have found manual entry of the *continuation* span content faster or more accurate than dictation; other *continuation* span information may have become available after dictation.

The findings raise questions to explore in user studies on improving the note creation process, including both timing of dictation and editing and technological interventions (e.g., commands for inserting information, automatic flagging of likely errors). Based on the distribution of *short edit* and capitalization subcategories, we estimate that roughly 80% are amenable to automatic detection algorithms that could be used in an enhanced editing tool to alert physicians to spans of text to check, optionally with proposed corrections. Advances in NLP algorithms for ASR error detection,[30–32] disfluency detection,[33] sentence segmentation,[34] true casing,[35] and entity recognition[36] are relevant here. Such algorithms also benefit from incorporating additional resources, such as patient data within the EHR and biomedical knowledge sources, as shown for edit detection.[37] For certain aspects of the note, such as the Physical Exam and Labs sections, both dictation and editing time can be reduced through the use of verbal commands, which led to increased length of notes for the physicians who used this capability in VGEENS. Properly designed, the use of such editing tools should decrease editing time and improve note quality by reducing the possibility that the clinician will miss an error.

The predominance of manual entry in the Assessment & Plan section is likely a consequence of the asynchronous workflow, since such behaviors have not been noted in papers discussing the use of ASR in an interactive dictation setting (to our knowledge). However, the asynchronous setting has some workflow benefits. Future development efforts should explore technology enabling physicians to maintain and interact with problem lists and care plans including both verbal commands in dictation and interactive tools for editing. There are potential advantages to manually retyping or inserting vital signs and other patient data into the note, because the act of doing so may make the physician more aware of these findings. This manual insertion must be balanced against the disadvantages, which include reduced time for other more important tasks. In practice, we find that physicians use templates that automatically insert these data, rather than typing them.

In this study, three primary note taking strategies emerged, but the physician sample size was too small to identify specific characteristics (typing speed, familiarity with dictation, etc.) that correlate with these strategies. With a larger participant pool, more note taking strategies may emerge, and the relationship between specific personal characteristics and note taking strategies may be identified and used to implement more personalized interventions to improve efficiency, quality, and satisfaction. In addition, a larger pool and/or longer study would make it possible to learn whether physicians adapt to the ASR system to improve note creation efficiency.

The analyses looked only at statistics available through automatic text analysis; it does not include assessments of note quality, which requires human evaluation of a large number of notes. Additional work is needed to understand the effect editing has on documentation time and note quality. The analysis of note quality could be extended to leverage guidelines from the Healthcare Documentation Quality Assessment and Management Best Practices Web site.[38]

This study has important limitations. We focused on inpatient progress notes rather than hospital admission notes, discharge summaries, or outpatient clinic notes. Our findings may differ from what would be learned using the same techniques for these other note types. However, outpatient progress notes—one of the most common note types physicians in the outpatient setting create—are similar in structure to hospital progress notes. Both generally follow a Subjective, Objective, Assessment, and Plan format, in which patient history, physical exam findings, laboratory and other results, and the plan of action is included. In addition, the findings may not apply to other health care organizations or specialties, and they may change as ASR performance improves. Finally, the note creation strategies of both the control group, which manually typed notes, and intervention group, which dictated notes, may have been influenced by the unblinded nature of this study, resulting in physicians being more diligent in their documentation process than in a real-world setting.

Conclusion

As others have observed, we found that physicians made many short, potentially time-consuming, edits throughout the note. The *short edits* implement changes related to formatting, rephrasing, and the correction of disfluencies and ASR errors. While *short edits* were frequent, a majority of the editing involved extended changes to the Assessment & Plan, which can be viewed as a continuation of the note creation process in this asynchronous editing setting. In an extensive survey of publications on ASR for EHR, Kumah-Crystal et al conclude that further research is needed to understand the impact of ASR tools on EHR workflow and safety.[9] Our findings raise additional considerations related to asynchronous editing and tools to facilitate both short and extended edits.

Clinical Relevance Statement

When creating clinical notes using ASR, clinicians edit transcripts with both short modifications that improve note accuracy and formatting, as well as longer edits that improve note relevance and completeness. Interventions to reduce ASR documentation burden should apply a portfolio of solutions to address all categories of required edits, including mechanisms for automatically inserting different types of edits as well as automatic formatting and error detection.

Multiple Choice Questions

1. This study explored the creation of clinical notes using automatic speech recognition (also known as voice recognition) in a noninteractive (asynchronous) setting. Dictated notes and the associated finalized notes were compared with identify physician edits. Edits were categorized as either a "short edit" (edits associated with error correction, rephrasing, and formatting) or a "continuation" (edits associated with a continuation of the note creation process). Which of the following is an example of a "short edit"?
   a. Inclusion of recent laboratory results.
   b. Removal of speech disfluencies, where the dictation includes repeated phrases.
   c. Removal of dictated information that is no longer relevant.
   d. Copying of content from the previous note.

   Correct Answer: The correct answer is option b, removal of speech disfluencies, where the dictation includes repeated phrases. "Short edits" included the correction of speech recognition errors, correction of disfluencies, incorporation of standard phrasing and formatting, and other minor rephrasing. Answers A, C, and D are examples of "continuation" edits, including the removal of dictation that is no longer relevant or the incorporation of information that was not dictated. Please refer to Pezzullo et al[11] for further study.

2. This study found that clinical note creation using automatic speech recognition in a noninteractive, asynchronous setting involved a substantial amount of editing (both short and extended edits) by physicians. Which of the following best describes this editing?
   a. Editing was primarily associated with the correction of speech recognition errors.
   b. Editing was associated with multiple factors, related to accuracy, completeness, and formatting.
   c. Editing was limited to the inclusion information that was not dictated.
   d. Editing was primarily associated with EHR-related formatting requirements.

   Correct Answer: The correct answer is option b, editing was associated with multiple factors, related to accuracy, completeness, and formatting. Notes were edited to implement a range of changes, related to the correction of ASR errors, the correction of disfluencies, implementation of standardized nomenclature and formatting, and inclusion of information not dictated. Please refer to Johnson et al[10] for further study.

Protection of Human and Animal Subjects

The University of Washington Human Subjects Division approved the VGEENS study, and this work was performed in compliance with the approved study design and procedures.

Funding

This study was supported by grant number R21HS023631 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Conflict of Interest

None.

References

1 Oxentenko AS, West CP, Popkave C, Weinberger SE, Kolars JC. Time spent on clinical documentation: a survey of internal medicine residents and program directors. Arch Intern Med 2010;170(04): 377–380
2 Yadav S, Kazanji NK C N, et al. Comparison of accuracy of physical examination findings in initial progress notes between paper charts and a newly implemented electronic health record. J Am Med Inform Assoc 2017;24(01):140–144
3 Friedberg MW, Chen PG, Van Busum KR, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Rand Health Q 2014;3(04):1
4 Sinsky CA, Willard-Grace R, Schutzbank AM, Sinsky TA, Margolius D, Bodenheimer T. In search of joy in practice: a report of 23 high-functioning primary care practices. Ann Fam Med 2013;11(03): 272–278
5 Lam JG, Lee BS, Chen PP. The effect of electronic health records adoption on patient visit volume at an academic ophthalmology department. BMC Health Serv Res 2016;16:7
6 Hodgson T, Coiera E. Risks and benefits of speech recognition for clinical documentation: a systematic review. J Am Med Inform Assoc 2016;23(e1)e169–e179
7 Hodgson T, Magrabi F, Coiera E. Efficiency and safety of speech recognition for documentation in the electronic health record. J Am Med Inform Assoc 2017;24(06):1127–1133
8 Borowitz SM. Computer-based speech recognition as an alternative to medical transcription. J Am Med Inform Assoc 2001;8 (01):101–102
9 Kumah-Crystal YA, Pirtle CJ, Whyte HM, Goode ES, Anders SH, Lehmann CU. Electronic health record interactions through voice: a review. Appl Clin Inform 2018;9(03):541–552
10 Johnson M, Lapkin S, Long V, et al. A systematic review of speech recognition technology in health care. BMC Med Inform Decis Mak 2014;14:94
11 Pezzullo JA, Tung GA, Rogg JM, Davis LM, Brody JM, Mayo-Smith WW. Voice recognition dictation: radiologist as transcriptionist. J Digit Imaging 2008;21(04):384–389
12 Vogel M, Kaisers W, Wassmuth R, Mayatepek E. Analysis of documentation speed using web-based medical speech recognition technology: randomized controlled trial. J Med Internet Res 2015;17(11):e247

13 Kauppinen T, Koivikko MP, Ahovuo J. Improvement of report workflow and productivity using speech recognition–a follow-up study. J Digit Imaging 2008;21(04):378–382
14 Mohr DNT, Turner DW, Pond GR, Kamath JS, De Vos CB, Carpenter PC. Speech recognition as a transcription aid: a randomized comparison with standard transcription. J Am Med Inform Assoc 2003;10(01):85–93
15 Weiss DL, Kim W, Branstetter BF IV, Prevedello LM. Radiology reporting: a closed-loop cycle from order entry to results communication. J Am Coll Radiol 2014;11(12 Pt B):1226–1237
16 Hammana I, Lepanto L, Poder T, Bellemare C, Ly MS. Speech recognition in the radiology department: a systematic review. Health Inf Manag 2015;44(02):4–10
17 Goss FR, Zhou L, Weiner SG. Incidence of speech recognition errors in the emergency department. Int J Med Inform 2016;93:70–73
18 Zhou L, Shi Y, Sears A. Third-party error detection support mechanisms for dictation speech recognition. Interact Comput 2010;22:375–388
19 Quint LE, Quint DJ, Myles JD. Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. J Am Coll Radiol 2008;5(12):1196–1199
20 Ringler MD, Goss BC, Bartholmai BJ. Syntactic and semantic errors in radiology reports associated with speech recognition software. Health Informatics J 2017;23(01):3–13
21 Payne TH, tenBroek AE, Fletcher GS, Labuguen MC. Transition from paper to electronic inpatient physician notes. J Am Med Inform Assoc 2010;17(01):108–111
22 Tai-Seale M, Olson CW, Li J, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff 2017;36(04):655–662
23 Payne TH, Alonso WD, Markiel JA, et al. Using voice to create inpatient progress notes: effects on note timeliness, quality, and physician satisfaction. JAMIA Open 2018. Doi: 10.1093/jamiaopen/ooy036
24 Payne TH, Alonso WD, Markiel JA, Lybarger K, White AA. Using voice to create hospital progress notes: Description of a mobile application and supporting system integrated with a commercial electronic health record. J Biomed Inform 2018;77:91–96
25 Payne TH. Improving accuracy of electronic notes using a faster, simpler approach. Available at: https://healthit.ahrq.gov/ahrq-funded-projects/improving-accuracy-electronic-notes-using-faster-simpler-approach?nav=publications. Accessed June 14, 2018
26 Payne TH, Hirschmann JV, Helbig S. The elements of electronic note style. J AHIMA 2003;74(02):68–70
27 Ratclif JW, Metzener DE. Pattern-matching the gestalt approach. Dr Dobb's J 1988;13:46
28 Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3(02):164–174
29 Nuance. Dragon Medical Practice Edition. Available at: https://www.nuance.com/healthcare/physician-and-clinical-speech/dragon-medical-practice-edition.html. Accessed June 1, 2017
30 Kalgaonkar K, Liu C, Gong Y. Estimating confidence scores on ASR results using recurrent neural networks. Proc IEEE Int Conf Acoust Speech Signal Process 2015;2015:4999–5003
31 Ogawa A, Hori T. ASR error detection and recognition rate estimation using deep bidirectional recurrent neural networks. Proc IEEE Int Conf Acoust Speech Signal Process 2015;2015:4370–4374
32 Angel Del-Agua M, Piqueras S, Gimenez A, Sanchis A, Civera J, Juan A. ASR confidence estimation with speaker-adapted recurrent neural networks. Proc InterSpeech 2016;2016:3464–3468
33 Yoshikawa M, Shindo H, Matsumoto Y. Joint transition-based dependency parsing and disfluency detection for automatic speech recognition texts. Proc Conf Empir Methods Nat Lang Process 2016;2016:1036–1041
34 Cho E, Niehues J, Waibel A. NMT-based segmentation and punctuation insertion for real-time spoken language translation. Proc InterSpeech 2017;2017:2645–2649
35 Susanto RH, Chieu HL, Lu W. Learning to capitalize with character-level recurrent neural networks: an empirical study. Proc Conf Empir Methods Nat Lang Process 2016;2016:2090–2095
36 Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. Proc NAACL-HLT 2016;2016:260–270
37 Lybarger K, Ostendorf M, Yetisgen M. Automatically detecting likely edits in clinical notes created using automatic speech recognition. AMIA Annu Symp Proc 2017;2017:1186–1195
38 Association for Healthcare Documentation Integrity [Internet]. Healthcare Documentation Quality Assessment and Management Best Practices (updated July 2017). Available at: https://www.ahdionline.org/general/custom.asp?page=qa. Accessed April 23, 2018