HHS Public Access

Author manuscript

Int J Med Inform. Author manuscript; available in PMC 2020 June 01.

Published in final edited form as:

Int J Med Inform. 2016 September ; 93: 70–73. doi:10.1016/j.ijmedinf.2016.05.005.

Incidence of speech recognition errors in the emergency department

Foster R. Goss[a,b,*], **Li Zhou**[c,d,e,g], **Scott G. Weiner**[f]

[a]University of Colorado, Department of Emergency Medicine, Aurora, CO, United States

[b]Tufts Medical Center, Department of Emergency Medicine and Clinical Decision Making, Boston, MA, United States

[c]Division of General Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, United States

[d]Clinical & Quality Analysis, Partners HealthCare System, Boston, MA, United States

[e]Clinical Informatics, Partners eCare, Partners HealthCare System, Boston, MA, United States

[f]Division of Health Policy Translation, Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, United States

[g]Harvard Medical School, Boston, MA, United States

Abstract

Background: Physician use of computerized speech recognition (SR) technology has risen in recent years due to its ease of use and efficiency at the point of care. However, error rates between 10 and 23% have been observed, raising concern about the number of errors being entered into the permanent medical record, their impact on quality of care and medical liability that may arise. Our aim was to determine the incidence and types of SR errors introduced by this technology in the emergency department (ED). Setting:

Level 1 emergency department with 42,000 visits/year in a tertiary academic teaching hospital. Methods:

A random sample of 100 notes dictated by attending emergency physicians (EPs) using SR software was collected from the ED electronic health record between January and June 2012. Two board-certified EPs annotated the notes and conducted error analysis independently. An existing classification schema was adopted to classify errors into eight errors types. Critical errors deemed to potentially impact patient care were identified.

[*]Corresponding author at: Department of Emergency Medicine, University of Colorado, United States. Foster.Goss@ucdenver.edu (F.R. Goss).

Author contributions

FRG and SW conceived of the study, collected the data and wrote the paper. All authors participated in the statistical analysis and its revisions. FRG and SW take responsibility for the paper as a whole.

Disclaimer

The authors have no disclosures or conflicts to report.

Results: There were 128 errors in total or 1.3 errors per note, and 14.8% (n = 19) errors were judged to be critical. 71% of notes contained errors, and 15% contained one or more critical errors. Annunciation errors were the highest at 53.9% (n = 69), followed by deletions at 18.0% (n = 23) and added words at 11.7% (n = 15). Nonsense errors, homonyms and spelling errors were present in 10.9% (n = 14), 4.7% (n = 6), and 0.8% (n = 1) of notes, respectively. There were no suffix or dictionary errors. Inter-annotator agreement was 97.8%.

Conclusions: This is the first estimate at classifying speech recognition errors in dictated emergency department notes. Speech recognition errors occur commonly with annunciation errors being the most frequent. Error rates were comparable if not lower than previous studies. 15% of errors were deemed critical, potentially leading to miscommunication that could affect patient care.

Keywords

Speech recognition; Emergency medicine; Patient safety

1. Introduction

1.1. Background

Physician use of computerized speech recognition (SR) technology has risen in recent years due to its ease of use and efficiency at the point of care. Nearly half of all licensed U.S. physicians use SR to enter information into the electronic health record (EHR) in their practice via a variety of methods [1]. Traditionally, a voice dictation generated by the physician is sent to a medical transcriptionist who manually transcribes the document and sends it back to the physician for review. Front-end SR (or SR-generated documentation) occurs when a physician dictates into a text field in the EHR or text document using SR software and edits the dictation in real-time before saving it. Back-end SR (or SR-assisted transcription) occurs when a physician dictates and the recorded transcription is automatically processed by SR software that sends it to a human transcriptionist to review and finally to the physician for review. While front-end SR is the most likely method used in emergency department (ED) and what was used in this study, other methods do exist and often depend on the type of SR systems supported within their respective institution.

1.2. Errors rates using speech recognition

Despite the advantages of SR technology, high error rates ranging from 10 to 23% have been observed in clinical documents generated by this technology [2], raising concern about the number of errors being entered into the permanent medical record, their impact on quality of care and the medical liability that may arise. To date, there have been few studies published on the use SR in ED [3–5]. A recent study by Zick et al. evaluated the accuracy and cost savings of traditional voice dictation as compared to a real-time SR software and observed high accuracies of 99.7% and 98.5% respectively [5]. Turnaround time was faster using the SR software as compared to traditional transcription and SR generated notes were less costly. While accuracy was reported, the types of errors that occurred were not systematically classified. In this pilot study, we sought to systematically classify and identify the incidence of SR errors in ED using a predefined classification schema by Zafar et al [6].

To the best our knowledge, this has never been studied before in the ED. This work attempts to add to a much wider discussion on the use of technology its impact on patient care and safety.

1.3. Case report

A 25-year old female presented to the emergency department with an abscess on her arm. On questioning, the patient mentioned that she had missed her period. The patient was evaluated by the attending physician and a note was dictated. The physician commented in their note that the patient had missed her period. The software interpreted the physician's reference to "period" as a punctuation mark ".". She returned the following day with a worsening cellulitis on her arm and a colleague of the first doctor prescribed an antibiotic that was contraindicated during pregnancy.

2. Methods

2.1. Data collection and sampling

This study was conducted in an urban academic emergency department located in Boston with 42,000 patient visits per year. A random sample of 360 attending emergency physician (EP) notes recorded from January to June 2012 (60 notes/month) was collected from the ED EHR system. Notes could be either hand-typed or dictated using Dragon Medical Software 10.0 or 10.1 (Nuance Communications Inc.). Dictations were performed using the Nuance PowerMic II and primarily dictated in the Emergency Department, an environment with higher ambient noise than the office setting. Only dictated notes were analyzed. To ensure each sample contained a representative number of words to analyze, we excluded sentences with less than 50 words (e.g. fragments or partial/incomplete dictations). Sample size was calculated using pilot data [7] and the software PASS (Power Analysis and Sample Size Software, version 11) [8]. We determined a sample size of 100 notes yielded an acceptable 95% confidence interval for notes containing critical errors of 14.3–31.4%, respectively. IRB approval was obtained for this study and determined it to be exempt as it posed no more than minimal risk to patient and all information was de-identified.

2.2. Error analysis

Error analysis was conducted independently by the two reviewers. Notes were reviewed and annotated using Knowtator [9], a text annotation tool built upon Protégé [10], an open-source ontology editor from Stanford University. We created a classification schema in Protégé based on Zafar et al. [6] (Table 1). Errors were deemed to be "critical" if they were believed by the reviewing physicians to potentially impact patient care. Identified errors were then reviewed and then jointly classified by the two reviewers and inter-annotator agreement calculated using Knowtator. Summary statistics were generated.

3. Results

Two board-certified EPs reviewed the notes and excluded those that were not dictated (n = 55), partially dictated (n = 7) or less than 50 words (n = 198). In total, 100 notes were included, dictated by 12 providers with a mean of 8.3 (SD 4.3) notes per provider. The

number of words in the notes ranged from 50 to 500 with the mean being 140.0 (SD 74.9). Inter-annotator agreement on the jointly classified errors was 97.8%. Overall, 71% of the notes contained errors. There were 128 errors in total or 1.3 errors per note. Annunciation errors were the highest (53.9%) followed by deletions (18.0%), added words (11.7%), and nonsense errors (10.9%). Homonyms and spelling errors were lower at 4.7% and 0.8%, respectively. There were no suffix or dictionary errors. Of these errors, 14.8% were judged to be critical errors. 15% of notes contained one or more critical errors. A summary of these types of errors are shown in Table 2 and examples of critical errors in Table 3.

4. Discussion

Errors in medical documentation are a critical problem that can adversely impact quality of care and patient safety [11–15]. We found nearly 71% of notes contained errors with an average of 1.3 errors per note. Annunciation errors were the most common followed by deletions and added words. In comparison to Zick et al., the number of errors we found were actually less, with 1.3 errors per note as compared 2.5 errors per note. The lack of spelling errors was expected, as words entered by SR systems are all based on words typically contained within a dictionary. Of concern, nearly 15% of errors were deemed to be critical errors, which were found in 15% of ED notes. Those errors could potentially impact patient care.

Multiple prior studies on SR accuracy and correction have been conducted in the field of radiology [5,16–24]. Quint et al. found that nearly 22% of SR-generated radiology reports contained significant errors [22]. Basma et al. found nearly 23% of imaging reports generated by SR contained at least 1 major error compared to 4% of conventional dictation transcriptions and were nearly 8 times more likely to contain major errors as compared to traditional dictation transcriptions. In the Medspeak IBM SR system, Kanel et al. found an error rate of 10.3% [25]. In a corpus of psychiatric SR generated notes, Derman et al. noted frequent word errors and sub-stitutions, making them contextually incorrect [26] even though all the words were spelled correctly. Automated correction of SR documents has been attempted by several studies [27–29], including Voll et al. who developed a statistical error detection method to detect post-transcription errors with a 96% detection rate [30].

Wong et al. developed an automatic system to process noisy clinical notes in real-time and achieved an accuracy of 88.7% [31]. The system was implemented within their clinical information system to drive decision support and further analytical functions. While promising, automated methods for correcting SR dictations errors continue to suffer from low accuracy [32,33] likely attributed to their algorithms for correcting the errors. These error rates are consistent with our findings if not slightly higher (i.e. Basma and Quint et al.).

Physicians use different means to document clinical information and the optimal form of documentation is still debatable. In the out-patient setting, Bates et al. [34] conducted an EHR documentation study which showed among 1088 physicians, 49% used templates, 22% used dictation (Back-end SR), 13% used free-form (typing) and 16% used a hybrid approach (template, dictation, free-form). Dictation was preferred by those who were hospital-based, practiced in an academic medical center and had used the EHR longer. Many clinicians feel

that templates make it inherently difficult to document the patient story and for this reason, prefer dictation. While the best means of documentation remains uncertain, it is believed that dictation will continue to be widely used, likely with the assistance of advanced SR techniques and natural language processing solutions that understand speech-captured data [21,26,34] and correct documentation errors.

At present, clinical documentation is the single most time-consuming part of using an EHR and can consume up to a third of a clinician's time [35]. While there are potential cost savings and productivity advantages by using SR technology [36,37] the cost of medical errors or patient harm associated with miscommunication of information or liability associated with errors being introduced into permanent medical record, could be far greater. These questions raise concern for the widespread adoption of SR technology and its implications on patient safety. Further studies will be valuable to better understand the clinical impact of its use.

5. Limitations

This was a single site and only used front-end speech recognition. We did not take into consideration the experience of our users (e.g., how long they had use the system, how well the SR software was trained). While dictations could have occurred in the ED, they also may have occurred in the office setting after a shift, which may have affected the accuracy of the dictations due to differences in ambient noise and physician attentiveness to accuracy of dictation. All dictations were by physicians whose native language was English. Accuracy may be affected by those for whom English was not their first language and we are conducting further research on the impact of users' accents on error rates in dictated medical documents. As this was a pilot study, we did not evaluate whether the errors were associated with actual adverse events.

6. Conclusions

Speech recognition errors occur commonly with annunciation errors being the most frequent. Error rates were comparable if not lower than previous studies. 15% of notes contained at least one critical error, potentially leading to miscommunication that could affect patient care.

Acknowledgments

This work was supported in part by AHRQ R01HS024264.

References

[1]. Zhou L, Parsons S, Hripcsak G, The evaluation of a temporal reasoning system in processing clinical discharge summaries, J. Am. Med. Inform. Assoc 15 (January–Febraury(1)) (2008) 99–106. [PubMed: 17947618]

[2]. Basma S, Lord B, Jacks LM, Rizk M, Scaranelo AM, Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription, Am. J. Roentgenol 197 (October(4)) (2011) 923–927. [PubMed: 21940580]

[3]. Spacone AB, Microcomputer voice-recognition program in a hospital emergency department, J. Soc. Health Syst 1 (November(2)) (1989) 111–118. [PubMed: 2519102]

[4]. Zemmel NJ, Park SM, Schweitzer J, O&apos, Keefe JS, Laughon MM, Edlich RF, Status of voicetype dictation for windows for the emergency physician, J. Emerg. Med 14 (July(4)) (1996) 511–515. [PubMed: 8842928]

[5]. Zick RG, Olsen J, Voice recognition software versus a traditional transcription service for physician charting in the ED, Am. J. Emerg. Med 19 (July(4)) (2001) 295–298. [PubMed: 11447517]

[6]. Zafar A, Mamlin B, Perkins S, Belsito AM, Overhage JM, McDonald CJ, A simple error classification system for understanding sources of error in automatic speech recognition and human transcription, Int. J. Med. Inf 73 (September(9–10)) (2004) 719–730.

[7]. Goss FR, Weiner SG, Incidence of speech recognition errors in the emergency department, Ann. Emergency Med 62 (4) (2013) S95–S96.

[8]. Power Analysis and Sample Size Software, (cited 08.04.14). Available from: http://www.ncss.com/software/pass/.

[9]. Ogren PV, Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. New York: Association for computational linguistics (2006).

[10]. Protege, (cited 02.08.14). Available from: http://protege.stanford.edu.

[11]. Motamedi SM, Posadas-Calleja J, Straus S, et al., The efficacy of computer-enabled discharge communication interventions: a systematic review, BMJ Qual. Saf 20 (May(5)) (2011) 403–415.

[12]. American Medical Association, Clinical Documentation, http://www.ama-assn.org/ama/pub/physician-resources/practice-management-center/claims-revenue-cycle/clinical-documentation.page? (last accessed 01.17.13.).

[13]. Davidson SJ, Zwemer FL Jr., Nathanson LA, Sable KN, Khan AN, Where's the beef? The promise and the reality of clinical documentation, Acad. Emerg. Med 11 (November(11)) (2004) 1127–1134. [PubMed: 15528575]

[14]. Cowan J, Clinical governance and clinical documentation: still a long way to go? Clin. Perform. Qual. Health Care 8 (3) (2000) 179–182. [PubMed: 11185832]

[15]. Schiff GD, Bates DW, Can electronic clinical documentation help prevent diagnostic errors? New Engl. J. Med 362 (12) (2010) 1066–1069. [PubMed: 20335582]

[16]. Leeming BW, Porter D, Jackson JD, Bleich HL, Simon M, Computerized radiologic reporting with voice data-entry, Radiology 138 (March(3)) (1981) 585–588. [PubMed: 7465833]

[17]. Robbins AH, Horowitz DM, Srinivasan MK, et al., Speech-controlled generation of radiology reports, Radiology 164 (August(2)) (1987) 569–573. [PubMed: 3602404]

[18]. Herman SJ, Accuracy of a voice-to-text personal dictation system in the generation of radiology reports, AJR Am. J. Roentgenol 165 (July(1)) (1995) 177–180. [PubMed: 7785581]

[19]. Mehta A, Dreyer KJ, Schweitzer A, Couris J, Rosenthal D, Voice recognition–an emerging necessity within radiology: experiences of the Massachusetts general hospital, J. Digit. Imaging 11 (4 (Suppl. 2)) (1998) 20–23. [PubMed: 9848057]

[20]. Basma S, Lord B, Jacks LM, Rizk M, Scaranelo AM, Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription, AJR Am. J. Roentgenol (October(4)) (2011) 923–927. [PubMed: 21940580]

[21]. Quint DJ, Voice recognition: ready for prime time? J. Am. Coll. Radiol 4 (October(10)) (2007) 667–679, discussion 70–1. [PubMed: 17903745]

[22]. 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 5 (December(12)) (2008) 1196–1199. [PubMed: 19027683]

[23]. Koester HH, Usage, performance, and satisfaction outcomes for experienced users of automatic speech recognition, J. Rehabil. Res. Dev 41 (September(5)) (2004) 739–754. [PubMed: 15558404]

[24]. Schiller NO, Horemans I, Ganushchak L, Koester D, Event-related brain potentials during the monitoring of speech errors, Neuroimage 44 (January(2)) (2009) 520–530. [PubMed: 18930155]

[25]. Kanal KM, Hangiandreou NJ, Sykes AM, et al., Initial evaluation of a continuous speech recognition program for radiology, J. Digit. Imaging 14 (March(1)) (2001) 30–37. [PubMed: 11310913]

[26]. Derman YD, Arenovich T, Strauss J, Speech recognition software and electronic psychiatric progress notes: physicians' ratings and preferences, BMC Med. Inform. Decis. Mak 10 (2010) 44. [PubMed: 20738875]

[27]. Bassil Y, Alwani M, Post-editing error correction algorithm for speech recognition using bing spelling suggestion, Int. J. Adv. Comput. Sci. Appl 3 (2) (2012).

[28]. Brandow RL, Strzalkowski T, Improving speech recognition through text-based linguistic post-processing, United States Patent 6064957 (2000).

[29]. Ringger EK, Allen JF, Error correction via a post-processor for continuous speech recognition acoustics, speech, and signal processing, 7–10 May, in: 1996 ICASSP-96 Conference Proceedings, 1996 IEEE International Conference, 1, 1996, pp. 427–430.

[30]. Voll K, Atkins S, Forster B, Improving the utility of speech recognition through error detection, J. Digit. Imaging 21 (April(4)) (2007) 371–377.

[31]. Wong W, Glance D, Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes, Artif. Intell. Med 53 (November(3)) (2011) 171–180. [PubMed: 21924593]

[32]. Rabiner LR, Juang BJ, Statistical Methods of Speech Recognition, second edition, Elsevier Encyclopedia of Language and Linguistics, 2005.

[33]. Choularton S, Early Stage Detection of Speech Recognition Errors. PhD Thesis, University of Sydney, 2010.

[34]. Pollard SE, Neri PM, Wilcox AR, et al., How physicians document outpatient visit notes in an electronic health record, Int. J. Med. Inform 82 (January(1)) (2013) 39–46. [PubMed: 22542717]

[35]. Friedberg MW, Chen PG, Van Busum KR, et al., Factors affecting physician professional satisfaction and their implications for patient care, Health Syst. Health Policy Rand Corporation (2013).

[36]. Houston JD, Rupp FW, Experience with implementation of a radiology speech recognition system, J. Digit. Imaging 13 (August(3)) (2000) 124–128. [PubMed: 15359751]

[37]. Speech recognition—accelerating the adoption of EHR, Nuance White Paper, http://www.nuance.com/ucmprod/groups/healthcare/@web-enus/documents/collateral/nd002649.pdf (last accessed 12.31.12.).

Summary Points

What is already known?

- Half of all licensed U.S physicians use speech recognition technology to enter information into the electronic health record.
- High error rates have been observed in clinical documents generated by this technology.

What this study adds?

- The majority of emergency department notes (71%) generated by this technology contain speech recognition errors.
- Errors classified as critical occur at rates similar to prior studies if not slightly lower.
- Clinicians using this technology need to be cognizant of its limitations and carefully proofread their notes before entering them into the permanent electronic health record.

Table 1

SR Error Types, Descriptions and Examples

Error type	Description	Example[a]
Annunciation errors	Occurring due to speaker mispronunciation	*He will see her on reactive* |Pupils equal round and reactive
Dictionary errors	Resulting from missing terms	*solo bricks* | Celebrex
Suffix errors	Caused by misrecognition of appropriate tenses of a word	*markedly* | marked
Added words	Word added	8 year old male *the* history
Deleted words	Word deleted	O2 saturation_percent
Homonym errors	Resulting from substitution of a phonetically identical word	Nares *or* two days | for two days
Spelling errors	Occurred only in human- transcribed notes and not with speech- recognized notes	and *get* her sugars | yet
Nonsense errors	Resulting from words or phrases whose meaning could not be understood by examining the context	*Patient up been admitted for stable gait*
Critical errors	Were deemed to be “critical” if they could potentially impact patient care	pulse 175, respiration 32, *temperature 12.9*, room air O2 *saturation percent.*

[a]Actual examples of SR errors identified within this study with the exception of Dictionary and Suffix errors, of which our sample had none.

Table 2

Frequencies of Different SR Error Types by Error and by Note

Error type	By Error	By Note[a]	
	N=128 n(%)	Notes containing errors N=71 n(%)	All notes N= 100 n(%)
No errors	-	-	29 (29.0)
Annunciation errors	69 (53.9)	48 (67.6)	48 (48.0)
Deleted words	23 (18.0)	19 (26.8)	19 (19.0)
Added words	15 (11.7)	11 (15.5)	11 (11.0)
Nonsense errors	14 (10.9)	11 (15.5)	11 (11.0)
Homonym errors	6 (4.7)	6 (8.5)	6 (6.0)
Spelling errors	1 (0.8)	1 (1.4)	1 (1.0)
Dictionary errors	0 [0]	0 [0]	0 [0]
Suffix errors	0 [0]	0 [0]	0 [0]
Critical errors[b]	19 (14.8)	15 (21.1)	15 (15.0)

[a]Number of notes containing that error type. The total of percentages is greater than 100% because some notes contained more than one type of errors.

[b]Critical errors can also be any other error types as defined above.

Table 3

Examples of Critical Errors

Dictation error	Possible Interpretation
Cardiac Exam is *regular regular*	Irregular irregular
Wet mount moderate amount of *high fever*	Hyphae
Cranial nerves II through XII intact, *he*	5 out of 5 motor strength *is out of* 5 motor strength
Extremities without CC *in no cord* or tenderness	Cords
Suspect pulpitis, and recommend *gentamycin* antibiotics	clindamycin (ordered due to penicillin allergy)
Temperature *12.9,* room air 02 saturation percent	102.9
He will see her on reactive	Pupils equal round and reactive
Pulling of secretions	Pooling of secretions
Exposure was a *pap* was found in room the family house	Bat
But will give *2 p.o* azithromycin for presumed Chlamydia	2g