Journal of the American Medical Informatics Association, 26(4), 2019, 324–338
doi: 10.1093/jamia/ocy179
Advance Access Publication Date: 11 February 2019
Review

Review

Speech recognition for clinical documentation from 1990 to 2018: a systematic review

Suzanne V. Blackley,[1] Jessica Huynh,[2] Liqin Wang,[2,3] Zfania Korach,[2,3] and Li Zhou[2,3]

[1]Clinical and Quality Analysis, Information Systems, Partners HealthCare, Boston, Massachusetts, USA, [2]General Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts, USA, and [3]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA

Corresponding Author: Suzanne V. Blackley, MA, Partners HealthCare, 399 Revolution Drive, Suite 1315, Somerville, MA 02145, USA (sblackley@partners.org)

Received 20 September 2018; Revised 16 November 2018; Editorial Decision 24 November 2018; Accepted 28 November 2018

ABSTRACT

Objective: The study sought to review recent literature regarding use of speech recognition (SR) technology for clinical documentation and to understand the impact of SR on document accuracy, provider efficiency, institutional cost, and more.

Materials and Methods: 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. We annotated included articles with their research topic(s), medical domain(s), and SR system(s) evaluated and analyzed the results.

Results: One hundred twenty-two articles were included. Forty-eight (39.3%) involved the radiology department exclusively and 10 (8.2%) involved emergency medicine; 10 (8.2%) mentioned multiple departments. Forty-eight (39.3%) articles studied productivity; 20 (16.4%) studied the effect of SR on documentation time, with mixed findings. Decreased turnaround time was reported in all 19 (15.6%) studies in which it was evaluated. Twenty-nine (23.8%) studies conducted error analyses, though various evaluation metrics were used. Reported percentage of documents with errors ranged from 4.8% to 71%; reported word error rates ranged from 7.4% to 38.7%. Seven (5.7%) studies assessed documentation-associated costs; 5 reported decreases and 2 reported increases. Many studies (44.3%) used products by Nuance Communications. Other vendors included IBM (9.0%) and Philips (6.6%); 7 (5.7%) used self-developed systems.

Conclusion: Despite widespread use of SR for clinical documentation, research on this topic remains largely heterogeneous, often using different evaluation metrics with mixed findings. Further, that SR-assisted documentation has become increasingly common in clinical settings beyond radiology warrants further investigation of its use and effectiveness in these settings.

Key words: speech recognition software, clinical documentation, clinical document quality, natural language processing, dictation

INTRODUCTION

Clinician use of speech recognition (SR) technology for clinical documentation has increased in recent years. A recent survey reported that more than 90% of hospitals plan to expand their use of front-end SR systems (ie, direct dictation into free-text fields of the electronic health record [EHR]) in the coming years.[1] As SR-assisted documentation has become more prevalent, there has been a simultaneous increase in research studying the effect of this technology on clinicians' workflow and medical practice. A comprehensive systematic review is needed to analyze and summarize relevant studies, identify knowledge gaps, and shed light on possible future research directions.

BACKGROUND AND SIGNIFICANCE

Several literature reviews about clinicians' SR use have been published over the past decade.[2–6] A 2008 review of the use and evaluation of SR technology in hospital settings found that SR was most commonly used to assist in clinical documentation, although it was also used for interactive voice response systems, controlling medical equipment, and automatic translation systems.[2] It also noted a lack of comprehensive, standardized methods for evaluating SR performance and utility, particularly those capable of considering the diverse clinical environments in which SR is used. This is a continuing problem, demonstrated by a 2015 review of SR in the radiology department, which found substantial heterogeneity across reviewed studies.[4] Nevertheless, existing literature reveals a number of trends regarding the impact of SR on report turnaround time and accuracy. In a 2014 review of SR use in healthcare applications, of the 14 studies included, most evaluated productivity, which typically improved following SR adoption, and report accuracy, which was generally lower with SR than with other documentation methods.[3] Hodgson and Coiera[6] found a similar trend, with mean errors per report tending to be higher for reports created with SR compared with those created with traditional dictation and transcription.

Previous reviews included approximately 15–45 articles and often focused on a specific aspect of SR-assisted documentation, such as effect on productivity or accuracy.[4,7,8] Although there is overlap in the time periods of this review and others, the present review includes over 120 articles from 10 scientific and medical literature databases relating to multiple aspects of SR-assisted clinical documentation over the past 3 decades. This broad scope reflects our aim to identify primary research questions pertaining to clinicians' use of SR for documentation, review major findings related to these questions, determine existing knowledge gaps and challenges, and ultimately propose specific areas we believe present significant opportunities for future research.

MATERIALS AND METHODS

Data sources and searches

This review was conducted in compliance with the 2009 PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement.[9] We conducted systematic database searches to retrieve articles published from database inception through October 15, 2018. Databases searched include PubMed, the Cumulative Index to Nursing and Allied Health Literature, Web of Science, Association for Computing Machinery 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. We subsequently reviewed the references of included articles to identify articles missed by database searching. The final search statements and number of articles yielded are available in Supplementary Appendix A.

Inclusion and exclusion criteria

Inclusion in the review required that all of the following criteria be met: (1) the article was written in English; (2) the article included metadata (authors, title, publication year) and an abstract; (3) the article was published between January 1, 1990, and October 15, 2018 (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 (ie, it was not being used for therapy or interactive voice control systems). Prior literature reviews were excluded from analysis.

Article selection and annotation

Search results from each database were exported, and the title, author(s), journal title or conference name, and publication year were extracted. Duplicate articles were removed. A preliminary screening was conducted to exclude articles failing to meet all inclusion criteria.

Two authors (SB, JH) read the abstracts of all remaining articles, writing a brief summary of each and manually consolidating the summaries into a set of 19 research topics (Table 1).

Each article was then annotated by 2 reviewers with its research topic(s), medical domain, and SR system(s) evaluated, if applicable. Articles could be assigned up to 3 research topics. Disagreements between reviewers were resolved through discussion.

RESULTS

The article selection process is summarized in Figure 1. In total, of the 1343 records retrieved, 122 articles were included in the analysis (Table 1).

Annotators agreed fully on medical domains and SR systems evaluated. They agreed fully on the research topics for 69 of 122 (56.6%) articles. Forty (32.8%) of the partial agreement cases involved 1 annotator selecting an additional research topic or missing a relevant topic. In the remaining 13 (10.7%) cases, the annotators selected completely different research topics.

Research trends over time

Most articles (89.3%) were published in or after 2000 (Figure 2A), with annual count fluctuating and peaking approximately every 7–8 years.

Overall, the largest proportion of studies (48 [39.3%]) were conducted in the radiology department,[13,18,20,22,24,26,28–41,45,46,48,49,52,53,56,64–66,70,72,75,84–90,92,106,113,117–119,127,128,130,131] followed by emergency medicine (10 [8.2%])[27,44,55,57–59,83,91,108,126] and nursing (8 [6.6%]).[67,76,78–80,93,107,124] More recently, 10 (8.2%) studies mentioned multiple departments,[12,16,23,54,77,121–123,125,132] and 15 (12.3%) did not specify the department(s) studied (Figure 2A).[62,82,94,97–100,102,103,110,112,114–116,129] Thirty (24.6%) studies used a version of Dragon NaturallySpeaking or Dragon Medical (Nuance Communications).[17,18,27,31,34,42–44,50,55,57,62,67,73,76,77,82,91,93,94,96,101,104,107,108,118,120,121,126] The second most commonly used system was PowerScribe (Nuance Communications), an SR system designed for radiology reporting, used in 13 (10.7%) studies.[22,28,32,35,39,41,45,52,70,86,87,89,90] Other commonly studied commercial products included Philips Speech Magic (6 [5.0%])[12,21,38,63,66,122] and IBM ViaVoice (6 [5.0%]).[14,15,19,61,95,101] Seven (5.7%) studies used a self-developed SR system rather than one offered by a third-party vendor,[16,74,100,110,115–117] 18 (14.8%) did not mention the system or vendor used,[23,25,29,37,75,80,84,99,102,103,105,111,112,119,123,124,129,133] and 6 (5.0%) studies did not directly evaluate an SR system.[49,60,72,79,98,128]

Most research topics, such as comparison to transcription, error analysis, SR use and impact on clinical workflows, and SR implementation, were studied throughout the review period. Since 2009, more studies have involved user surveys and interviews (Figure 2B).

Table 1. List of research topics among included articles (n = 122)

Research Topic	Description of Relevant Articles	n (%)
Documentation time/cost and productivity analysis[a,b]		48 (39.3)
Documentation time[10–29]	Analysis of time needed for documentation	20 (16.4)
Turnaround time[30–48]	Analysis of the amount of time between dictation completion and report availability	19 (15.6)
Documentation-associated cost[17,24,30,33,34,49–51]	Analysis of how SR effects documentation costs	8 (6.6)
Other[41,52–57]	Analysis of other measures (eg, report completed per time period, mean report length)	7 (5.7)
Usage and workflow[a]		35 (28.7)
Effect of user/environmental characteristics on SR[11,21,56,58–68]	Evaluation of the effect of user characteristics (eg, gender, language) and/or environmental characteristics (eg, location, noise level) on SR accuracy and/or usability	14 (11.5)
With templates/structured reporting[28,42,50,69–74]	Studies involving the use of SR in conjunction with templates or other structured documentation methods, including the use of SR for data entry	9 (7.4)
In the workflow[35,37,53,58,60,75–83]	Evaluation of how/where SR fits into existing clinical workflows and its impact on users	14 (11.5)
Error analysis[a,c] [10,19,20,24,26,33,41,44,57,59,64,66,84–97]	Analysis of the frequency and/or types of errors found in clinical documents created with SR, including comparisons of error rates pre- and postediting	29 (23.8)
Comparison with/in addition to transcription[10,13–18,20,22,24,30,33,36,40,41,43,44,46,48,52,64,66,85,98,99]	Comparison of SR-assisted documentation with traditional dictation and transcription, and/or evaluation of documentation processes that combine these 2 methods	25 (20.5)
Methods[a]		25 (20.5)
Enhancement for clinical documents[100–109]	Enhancement of SR output for use in clinical documentation, either during SR (eg, introducing larger or more specific medical vocabularies) or downstream (eg, automatically performing named entity recognition on SR output)	10 (8.2)
Language modeling and dictionaries[19,94,110–117]	Training and/or testing of language models and/or dictionaries (ie, as part of closed-vocabulary language model) for use in an SR system	10 (8.2)
Acoustic modeling[110,112,115,117]	Training and/or testing of acoustic models for use in an SR system	4 (3.3)
Automatic error detection[93,118–120]	Design, implementation, and/or evaluation of methods of automatically detecting errors in clinical documents created with SR	4 (3.3)
Grammars[71]	Studies involving grammar-based SR systems (ie, SR systems that allow only those utterances that are part of a specific, predefined grammar)	1 (0.8)
User survey/interview[25,47,58,67,68,76–79,83,121–126]	Survey of or interviews with current, future, and/or former SR users	16 (13.1)
Implementation[11,23,32,34,47,50,55,70,75,82,110,127,128]	Studies focused on the time during and immediately after SR implementation, including its impact on the hospital and/or individual clinicians	13 (10.7)
Comparison of commercial SR products[35,69,73,94,95]	Comparison of 2 or more commercially available SR systems (eg, in terms of accuracy, cost, usability)	5 (4.1)
Effect on documentation quality[21,25,53,129]	Evaluation of how SR does or does not affect the quality of the documents produced, optionally with respect to a pre-existing reporting guideline or standard	5 (4.1)
Preparation for SR[12,72,76,79]	Studies focused on the lead-up to SR implementation, including such topics as staff training, SR system selection process, and more	4 (3.3)

SR: speech recognition.

[a]Annotators were instructed not to select "documentation time/cost and productivity analysis," "usage and workflow," or "methods," but instead to select 1 of the more specific subtopics.

[b]Details can be found in Table 2.

[c]Details can be found in Table 3.

Research topics

Documentation time/cost and productivity analysis

The most common research topic was documentation time or cost and productivity analysis, which applied to 48 (39.3%) articles (Table 2).[10–57] Most (27 [56.3%]) evaluated documents created in the radiology department.[13,18,20,22,24,26,28–41,45,46,48,49,52,53,56] The rest involved notes from a variety of medical domains, such as pathology (5 [10.4%])[11,15,42,43,51] and emergency medicine (4 [8.3%])[27,44,55,57] Studies assessing productivity often did not list exact numbers of documents or speakers they evaluated; instead, many stated that all documents created during a certain time period were included the analysis, limiting the ability to compare and summarize results across studies. These studies also demonstrated substantial variation in how productivity was quantified, although certain measures, such as mean documentation time and turnaround time, did emerge as commonly used productivity indicators.

The most frequent measure was time needed for documentation, which was evaluated in 20 of 48 (41.7%) studies. Results were mixed regarding whether incorporating SR technology into the doc-

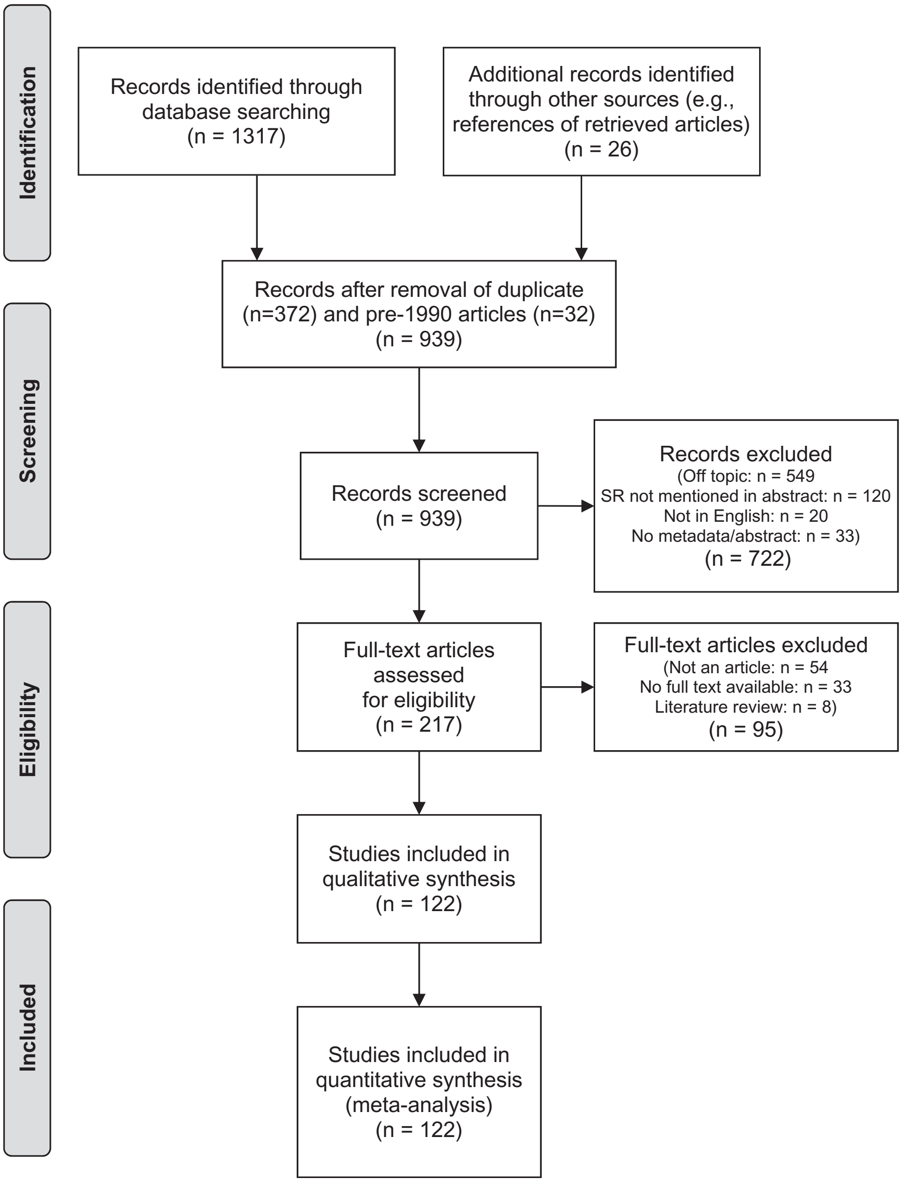
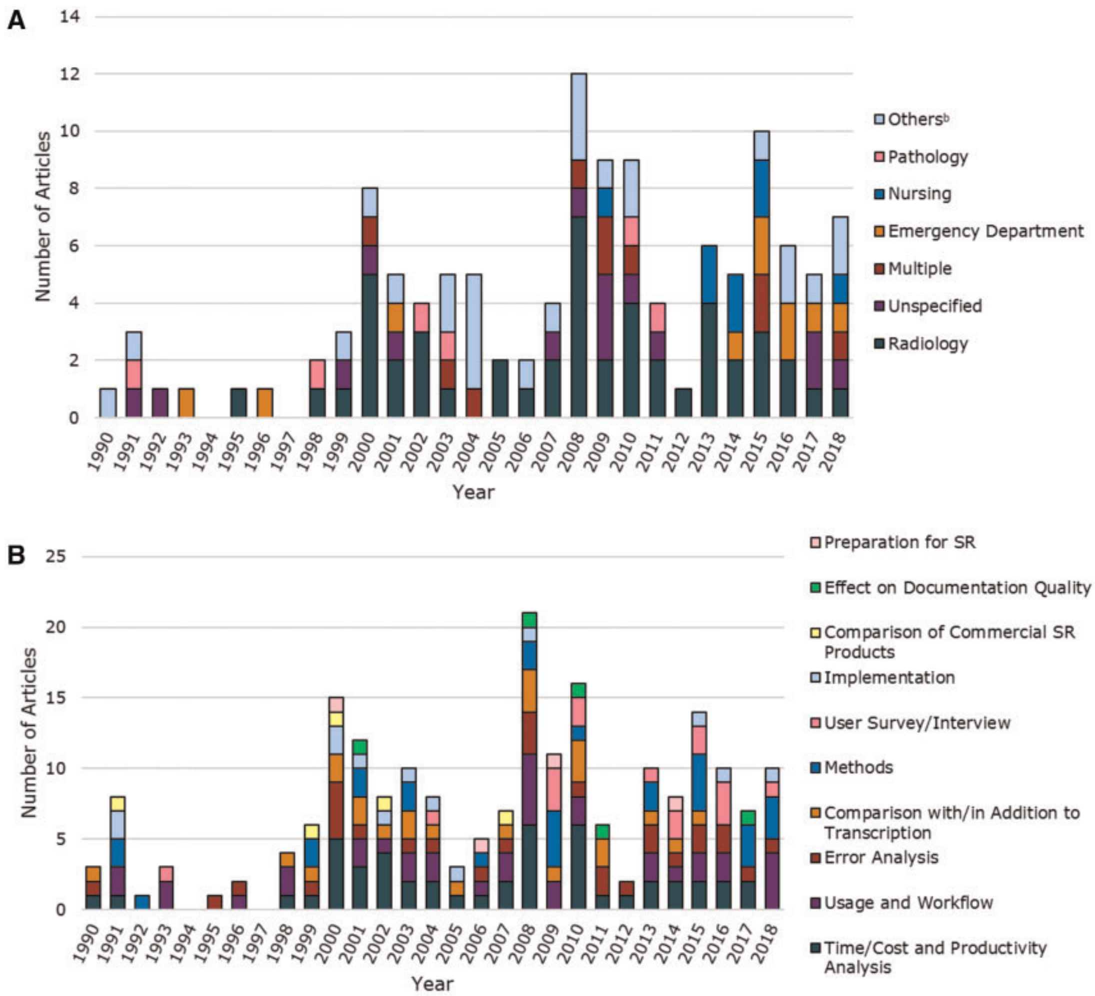


Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2009 flow diagram.[9] SR: speech recognition.

umentation process resulted in an increase or a decrease in overall documentation time compared with other input methods. Five studies reported decreases in documentation time following the introduction of SR technology,[11,13,18,20,21] with observed reductions ranging from 19% with SR-assisted transcription compared with conventional transcription[13] to 92% with SR compared with a keyboard and touchscreen interface.[21] A sixth found that total documentation time began to decrease compared with conventional transcription as the SR system's error rate fell below 16%,[19] and a seventh estimated an 89% decrease in documentation time with SR compared with conventional transcription.[23] However, 9 studies reported increased documentation times[10,12,14–17,22,24,29] and 4 reported no significant change.[25–28] Reported increases ranged from a 13.4% increase in mean document creation time[22] to a 200% increase in mean dictation and correction time with SR compared with conventional dictation and transcription.[17] Studies also varied in which aspect of the documentation process they evaluated. For example, while most studies evaluated SR-assisted documentation in terms of clinicians' productivity, Mohr et al[16] was 1 of few studies to investigate the impact of SR on transcriptionists' productivity, finding that editing SR-generated reports took longer than traditional dictation and transcription.

The second most frequent productivity measure was report turnaround time, which was evaluated in 19 of 48 (39.6%) articles.[30–48] All 19 found that implementing SR technology reduced mean and/or median turnaround times, often by more than 90%.[30,32,37,41,44,46,48] Only 8 (16.7%) articles included cost analyses,[17,24,30,33,34,49–51] the most recent of which was published in 2008.[24] Five (62.5%) of these studies reported decreases in documentation-associated costs following adoption of SR software.[30,33,34,50,51] However, 2 (25.0%) reported increased costs, citing greater software expenses[17] and the fact that highly paid physicians needed to spend more time editing their notes.[24] The remaining paper proposed an econometric model for estimating the impact of SR and transcription on



[a]2018 includes only articles published through October 15, 2018

[b]Others include anesthesia (2), cardiology (1), dentistry (2), EMS/rescue (2), general surgery (1), internal medicine (7), medications (1), oncology (2), orthopedics (1), otorhinolaryngology (1), pediatrics (2), psychiatry (1), pulmonology (1) and sports medicine (1)

[c]Each article could be assigned up to 3 research tpoics; 56 articles were assigned 1 topic, 43 articles were assigned 2 topics and 23 articles were assigned 3 topics

Figure 2. Temporal trends of included articles (n = 122). Part A is the number of topics per year broken down by medical domain. Part B is the number of topics per year broken down by research topic. EMS: emergency medical services; SR: speech recognition.

documentation-associated cost intended to help institutions decide which is most suitible.[49]

Error analysis

The second most prevalent research topic, analysis of errors in clinical documents created with SR technology, applied to 29 (23.8%) articles (Table 3).[10,13,15,19,20,24,26,33,41,44,57,59,64,66,84–97,132] Of these, 15 (51.7%) evaluated errors in SR-generated radiology reports.[20,24,26,33,41,64,66,84–90,92] The rest evaluated notes from various medical domains, most commonly emergency medicine (4 [13.8%]),[44,57,59,91] except 2 (6.9%) that did not specify a medical domain.[94,97] Many error analyses were performed via retrospective analysis of real patient reports,[33,41,64,66,84–91,96,132] although some studies, especially those published before 2008, were conducted in controlled laboratory settings, with study subjects dictating notes about real or fictional patients.[10,19,20,44,59,92–95] The papers typically included error identification and classification frameworks, although these varied widely in scope and granularity. Some had a very narrow focus, such as a study that determined the rate of laterality errors in radiology reports.[87] Others had classification schemas with more than 10 distinct error types.[85,132] There was also substantial variation in the number of speakers and reports evaluated (Table 3).

Studies published after 2008, mostly retrospective analyses, primarily reported the percentage of documents containing errors, which ranged from 4.8%[64] to 71%[91] for finalized (signed) documents.[24,41,64,66,84–91,132] However, earlier studies, mostly controlled laboratory studies, typically reported the percentage of correctly (or incorrectly) recognized words, with accuracies ranging from 92.7%[33] to 98.5%[44] and word error rates ranging from 7.0%[95] to 38.72%[19] with general vocabularies and from 5.21%[19] to 9%[90] with specialized vocabularies.[19,20,33,44,59,92–95,132] Many studies also reported the mean number of errors per document, ranging from 0.6[10] to 4.2.[10,26,44,88,90,91,132]

Table 2. Summary of articles related to documentation/cost and productivity analysis (n = 48)

Subtopic	Measure	Medical Domain	Articles	Summary of Findings
Documentation time (n = 20)	Mean document creation time	Radiology	Vorbeck et al (2000)[13] Rana et al (2005)[18] Bhan et al (2008)[22] Pezzullo et al (2008)[24] Hawkins et al (2012)[26] Hanna et al (2016)[28] Segrelles et al (2017)[29]	5 studies[11,13,18,21,23] reported decreases in mean documentation time after SR adoption, ranging from 19%[13] to 92%[21]; 4 studies[10,22,24,29] recorded increases, ranging from 13.4%[22] to 50%[24], and 3 studies[26–28] reported no statistically significant difference
		Anesthesiology	Alapetite et al (2008)[21]	
		Dentistry	Feldman and Stevens (1990)[10]	
		Emergency department	dela Cruz et al (2014)[27]	
		Pathology	Klatt (1991)[11]	
		Unspecified	Gonzalez Sanchez et al (2008)[23]	
	Mean dictation and/or correction time	Pediatrics	Borowitz (2001)[14] Issenman and Jaffer (2004)[17]	All reported increases after SR adoption, ranging from 13.9%[14] to 200%,[17] although 1[19] reported decreases if the SR error rate was ≤ 16%
		Multiple	Monnich and Wetter (2000)[12]	
		Otorhinolaryngology	Ilgner et al (2006)[19]	
		Pathology	Al-Aynati and Chorneyko (2003)[15]	
	Hours of secretary work per minute of dictation processed	Multiple	Mohr et al (2003)[16]	Secretaries were 55.8%–87.3% less productive with SR vs conventional transcription
	Total documentation time	Radiology	Ichikawa et al (2007)[20]	Decreased by 32.7–71.3% across 4 transcriptionists
	Users' perceptions of SR impact on document creation time	Psychiatry	Derman et al (2010)[25]	No perceived benefit with SR vs other methods
Turnaround time (n = 19)	Mean turnaround time	Radiology	Rosenthal et al (1998)[30] Chapman et al (2000)[31] Lemme and Morin (2000)[32] Ramaswamy et al (2000)[33] Callaway et al (2002)[34] Langer (2002)[35] Langer (2002)[36] Gopakumar et al (2008)[37] Koivikko et al (2008)[38] Hart et al (2010)[39] Krishnaraj et al (2010)[40] Strahan and Schneider-Kolsky (2010)[41]	All reported decreased turnaround times, ranging from 50.3%[33] to nearly 100%[41]
		Pathology	Kang et al (2010)[42] Singh and Pal (2011)[43]	
		Emergency department	Zick and Olsen (2001)[44]	
	Median, 80th percentile, and/or 95th percentile turnaround time	Radiology	Andriole et al (2010)[45] Prevedello et al (2014)[46]	All reported decreases, ranging from 50%[47] to 95.8%[46] (for median turnaround times)
		Pathology	Kang et al (2010)[42]	
		Sports medicine	Ahlgrim et al (2016)[47]	
	Minimum turnaround time	Radiology	Pavlicek et al (1999)[48]	Decreased by 91.7%

(continued)

Table 2. continued

Subtopic	Measure	Medical Domain	Articles	Summary of Findings
Documentation-associated cost(n = 8)	Change in cost over time	Radiology	Rosenthal et al (1998)[30] Ramaswamy et al (2000)[33] Callaway et al (2002)[34] Reinus (2007)[49] Pezzullo et al (2008)[24]	5 studies[30,33,34,50,51] overall decreases in documentation-associated costs following SR introduction; 2[17,24] reported increases; the remaining study[49] involved the development of an econometric model for estimating the impact of SR and transcription on cost
		Orthopedics	Corces et al (2004)[50]	
		Pathology	Henricks et al (2002)[51]	
		Pediatrics	Issenman and Jaffer (2004)[17]	
Other(n = 7)	Reports completed per time period	Radiology	Strahan and Schneider-Kolsky (2010)[41] Williams et al (2013)[52]	Ranged from a 41% increase[52] to a 35% decrease[41]
	Report availability	Radiology	Hayt and Alexander (2001)[53]	Percentage of reports available within 12 h of dictation increased from 3% to 42%
	Mean characters per minute	Multiple	Vogel et al (2015)[54]	Increased from 173 to 217 with SR vs with typing
	Mean length of stay	Emergency department	Lo et al (2015)[55]	Temporarily increased by 9.3%, then settled to a new baseline of 4.3% longer
	Mean report length	Radiology	Kauppinen et al (2013)[56]	433 characters and 11 character corrections per report for new SR users vs 298 and 6 character corrections characters per report for experienced users
	Mean task completion time	Emergency department	Hodgson et al (2017)[57]	18.11% slower with SR vs keyboard and mouse; 16.95% slower for simple tasks, 18.40% slower for complex tasks

SR: speech recognition.

Comparisons between, or assessments of the combination of, SR-assisted dictation and traditional dictation and transcription

Twenty-five (20.5%) articles conducted comparisons between, or assessed the combination of, SR-assisted dictation and traditional dictation and transcription.[10,13–18,20,22,24,30,33,36,40,41,43,44,46,48,52,64,66,85,98,99] Sixteen (64.0%) of these studies were conducted in the radiology department.[13,18,20,22,24,30,33,36,40,41,46,48,52,64,66,85] Twenty (80.0%) studies compared SR and traditional dictation in terms of productivity,[10,13–18,20,22,24,30,33,36,40,41,43,44,46,48,52] such as documentation time[10,13,14,17,18,20,24] or number of reports completed within a certain time period.[14,40,41,43,46,52] The second most common measure was report accuracy, used in 12 (48.0%) studies.[10,13,15,18,20,24,33,41,44,64,66,85]

Generally, studies comparing SR and transcription in terms of productivity found greater clinician productivity with SR than with traditional dictation and transcription (see Documentation Time/Cost and Productivity Analysis). However, studies comparing accuracy unanimously found more errors in self-edited SR-generated reports compared with those transcribed or edited by professional transcriptionists. For example, 1 study found that 23% of reports created with SR contained errors, compared with only 4% of those created with conventional dictation and transcription.[85] Another found that 25.6% of SR reports contained errors, compared with 9.3% of those that were dictated and transcribed.[66]

Impact on clinical workflow

Thirty-five (28.7%) studies evaluated the impact of SR use on clinical workflow.[11,21,28,35,37,42,50,53,56,58–79,83] Of these, approximately half (16 [45.7%]) were conducted in a controlled laboratory or simulation setting,[21,37,59–61,63,65,67,68,71,73,76,78–80,83] while another 14 (40.0%) involved real patient records, either via in vivo observation or retrospective audit.[11,28,35,42,50,53,56,58,62,64,66,75,81,82] Fourteen (40.0%) studies examined the effect of various user and/or environmental characteristics on SR usability and accuracy.[11,21,56,58–68] Findings were mixed regarding whether user characteristics (eg, gender, native language, experience level) impacted SR performance. For example, while some studies[62,67] reported significant differences in recognition rates between male and female speakers, others did not.[61,63] Similarly, some studies[56,62] found that experience level significantly impacted error rates, while others found no difference.[64,66] All 3 studies investigating the impact of native language and/or accent found significant differences in recognition rates between native and non-native speakers.[64,66,67] Of studies evaluating

Table 3. Summary of articles including error analyses (n = 29)

Measure	Medical Domain	Articles	Summary of Study Designs and Findings
Percentage of documents with errors (n = 13)	Radiology	McGurk et al (2008)[64] Pezzullo et al (2008)[24] Quint et al (2008)[84] Strahan and Schneider-Kolsky (2010)[41] Basma et al (2011)[85] Chang et al (2011)[86] Luetmer et al (2013)[87] Hawkins et al (2014)[88] du Toit et al (2015)[66] Ringler et al (2015)[89] Motyer et al (2016)[90]	**Study design** Retrospective, cross-sectional by input method, with real reports[24,41,64,66,85,87] Retrospective, cross-sectional by report type, with real reports[24,85,86,90,132] Retrospective study with real reports[84,88,89,91] Prospective, cross-sectional study with real reports[24] **Number of speakers**a Median: 19[86] Range: 2[41] to 147[89] **Number of documents evaluated** Median: 308[85] Range: 100[24,41,91] to 584 878[87] **Percentage of finalized documents with errors** Median: 26.9%[86] Range: 4.8%[64] to 71%[91]
	Emergency Department	Goss et al (2016)[91]	
	Multiple	Zhou et al (2018)[132]	
Mean errors per document (n = 7)	Radiology	Hawkins et al (2012)[26] Hawkins et al (2014)[88] Motyer et al (2016)[90]	**Study design** Retrospective study of real reports[88,90,91] Retrospective, cross-sectional by report type, with real reports[132] Prospective, cross-sectional by report type, with real reports[26] Controlled lab setting, cross-sectional by input method, with real reports[44] Observational study, cross-sectional by input method, with real reports[10] **Number of speakers**b Median: 12[91] Range: 2[44] to 144[132] **Number of documents** Median: 217[132] Range: 20[10] to 1173[26] **Mean errors per document** Median: 1.3[91] Range: 0.24[90] to 2.5[44]
	Emergency Department	Zick and Olsen (2001)[44] Goss et al (2016)[91]	
	Dentistry	Feldman and Stevens (1990)[10]	
	Multiple	Zhou et al (2018)[132]	
Accuracyc (n = 6)	Radiology	Herman (1995)[92] Ramaswamy et al (2000)[33] Ichikawa et al (2007)[20]	**Study design** Controlled lab setting, cross-sectional by input method, with real reports[20,44,94] Controlled lab setting with real reports[92] Controlled lab setting with fictional patient scenarios[93] Retrospective, cross-sectional study with real reports[33] **Number of speakers**d Median: 1.5 Range: 1[20,93] to 5[33] **Number of words evaluated**[e] Median: 6019 Range: 7277[93] to 18 721[92] **Accuracy** Median: 96.4% Range: 73%[93] to 98.5%[44], but often varied within studies based on the configuration of the SR system(s) evaluated
	Emergency Department	Zick and Olsen (2001)[44]	
	Nursing	Suominen and Ferraro (2013)[93]	
	Unspecified	Zafar et al (1999)[94]	

(continued)

Table 3. continued

Measure	Medical Domain	Articles	Summary of Study Designs and Findings
Word error rate[f] (n = 4)	Emergency Department Internal Medicine Otorhinolaryngology Multiple	Zemmel et al (1996)[59] Devine et al (2000)[95] Ilgner et al (2006)[19] Zhou et al (2018)[132]	**Study design** Controlled lab setting, cross-sectional by report type, with real reports[19,95] Controlled lab setting, cross-sectional by SR system configuration[59] Retrospective, cross-sectional by report type, with real reports[132] **Number of speakers**[g] Median: 12[95] Range: 7[59] to 144[132] **Number of documents** Median: 46 Range: 7[59] to 217[132] **Number of words**[h] Median: 60 874 Range: 11 568[95] to 110 180[132] **Word error rate** Median: 14.5% with general vocabularies, 11% with specialized vocabularies Range: 7.4%[132] to 38.72%[19] with general vocabularies; 5.21%[19] to 9%[59] with specialized vocabularies
Other (n = 3)	Emergency Department	Hodgson et al (2017)[57]	Controlled lab setting, cross-sectional by input method; 35 participants were randomly allocated simple and complex clinical tasks 138 total errors with minor, moderate, or major potential for patient harm with SR across simple and complex tasks, vs 32 with keyboard and mouse
	Internal Medicine	Zafar et al (2004)[96]	Retrospective analysis of 148 real reports (104 created by 1 speaker with SR, 44 human transcribed with multiple speakers) 9 identified categories of SR errors, including enunciation, dictionary, suffix, added words, deleted words, homonym, spelling, nonsense, and critical errors
	Unspecified	McKoskey and Boley (2000)[97]	Unsupervised clustering of 1200 completed dictations from 6 speakers aligned with their original SR output Identified error clusters: short and function words; vowel destressing and cliticization; vowel syncope; words with sounds affected by telephony interference (eg, fricatives)

SR: speech recognition.
[a]4 studies did not report the number of speakers.[66,87,88,90]
[b]4 studies did not report the number of speakers.[10,26,88,90]
[c]Accuracy = number of correctly recognized words/total number of words dictated.
[d]2 studies did not report the number of speakers.[92,94]
[e]2 studies did not report the number of words evaluated.[44,94]
[f]Word error rate = (number of substitutions + number of insertions + number of deletions)/total number of words dictated.
[g]1 study did not report the number of speakers.[19]
[h]2 studies did not report the number of words evaluated.[19,59]

the effect of environmental characteristics (eg, ambient noise level), most found that background noise significantly impacted recognition,[59,63–65] although 1 found differences only with certain types of noise (eg, a ringing telephone or paging device, but not a nearby printer or radiator)[11] and another reported successful recognition regardless of background noise.[21]

Fourteen (11.5%) studies investigated the role of SR in the workflow.[35,37,53,58,60,75–79,83] These studies addressed a variety of workflow-related issues, such as when or where providers conduct dictations[35,53,58,60,78] and the ability of SR to coexist with existing workflow habits.[37,53,58] Nine (7.4%) studies specifically investigated SR use in combination with templates or other structured reporting methods.[28,42,50,69,71–74,134] Of these, 2 outlined the features required of a workflow that successfully combines SR and structured reporting systems.[69,72] Five studies found that SR and templates complement each other well, yielding improved efficiency and accuracy and partially offsetting the additional time required to edit transcribed dictations.[28,42,50,70,71,74] Only 1 study reported that they did not work well together; the authors found that verbally navigating templates required too many commands (as opposed to natural language), making them unintuitive and providing no discernible benefit over navigation via mouse.[73]

SR methods

Twenty-five (20.5%) studies involved an aspect of SR methodology and/or architecture.[19,71,93,94,100–116,118,119] Among these were 10 (40.0%) studies about enhancing SR output for clinical documentation,[100–109] 10 (40.0%) about language modeling and dictionaries,[19,94,110–116] 4 (16.0%) about acoustic modeling,[110,112,115] 4 (16.0%) about automatically detecting errors in SR-generated documents,[93,118,119] and 1 (4.0%) about grammar-based SR systems.[71] Nine (36.0%) methodology studies did not specify a setting but involved SR-assisted medical documentation in general.[94,100,102,103,110,112,114,116,117] Only 4 (16.0%) addressed automatic post-SR error detection, of which 3 attempted to implement such a system,[118–120] while the fourth detailed a preliminary study demonstrating the feasibility of the authors' proposed error detection method.[93] The 3 implemented error detection systems varied substantially in scope, with 2 attempting to capture errors of any type,[118,120] while the third focused specifically on laterality and gender errors in radiology reports.[119]

User surveys and interviews

Sixteen (13.1%) studies included surveys of or interviews with current or future SR users.[25,47,58,67,68,76–79,83,121–126] Studies soliciting user feedback have become more prevalent in recent years; 14 (87.5%) of the studies were published within the past decade (2008–2018).[25,47,67,68,76–79,83,122–126] Seven (43.8%) studies asked about the perceived usability, benefits, and drawbacks of SR, making it the most common area of inquirey.[25,47,58,67,83,123,125] Five (31.3%) asked about clinicians' expectations regarding future adoption of an SR system or experiences with a recently adopted system[78,79,121,122,126] The remaining 4 (25.0%) involved trial or pilot implementations of SR systems, in which users' feedback was collected to help inform future SR adoption.[68,76,77,124]

SR implementation and other topics

Thirteen (10.7%) studies addressed issues related to implementing SR in a healthcare setting, in which authors outlined the SR implementation process, often drawing from personal experience and including guidelines or suggestions for other institutions considering adopting an SR-assisted documentation workflow.[11,23,32,34,47,50,55,70,75,110,127,128] Similarly, 5 (4.1%) studies conducted comparisons of commercially available SR products, presenting the benefits and drawbacks of each system to assist potential users in deciding between these systems[35,69,73,94,95] and 4 (3.3%) described how hospitals can effectively prepare for SR adoption.[12,72,76,79]

Finally, 5 (4.1%) studies investigated the effect of SR on documentation quality.[21,25,53,129,133] For example, 1 study found that SR implementation lowered the percentage of progress notes involving copying and pasting from 92.73% to 49.71%, resulting in reduced errors and higher quality notes.[133] However, another found that SR (in combination with a picture archiving and communication system), despite allowing for faster report access, ultimately negatively affected documentation quality by limiting the time available for face-to-face communication between the radiologist and other clinicians before report creation.[53] A third study found substantial differences in the type and frequency of words present in dictated notes versus typed notes, potentially affecting not only note quality but also the performance of downstream systems such as natural language processing–based clinical decision support tools.[129]

DISCUSSION

We systematically reviewed articles retrieved from 10 scientific and medical literature databases spanning nearly 3 decades to assess the state of current research on the use of SR technology for clinical documentation and identify knowledge gaps and areas in need of further study. Overall, we found that existing research has focused largely on 3 topics: (1) the impact of SR on documentation time/cost and productivity, (2) the accuracy of SR-generated clinical documents and analysis of errors produced by SR systems, and (3) the relationship between SR and traditional dictation and transcription, including comparisons between the 2 documentation modes and analyses of how they can be used together.

In general, there has been a relative lack of studies conducted in nonradiology settings, although the magnitude of this inequity has lessened in recent years as SR use has become more widespread. However, assessing the accuracy and utility of SR on a large scale remains difficult due to continued inconsistencies in how these factors are evaluated. For example, although many studies reported differences in documentation time with SR compared with other input methods, some reported pre- and post-SR documentation times, while others only reported the actual time difference, making it difficult to compare time savings across studies. Similar heterogeneity exists in other commonly reported metrics, such as accuracy and cost.

Fewer articles involved SR methodology compared with other research topics, and most methodology articles were published before 2008. This may be because many hospitals now use SR systems provided by third-party vendors who manage and maintain the actual SR architecture (eg, language and acoustic models). Methods for automatically detecting errors in clinical documents created with SR technology have received particularly little attention. As SR-assisted documentation has become more prevalent, clinicians have expressed concerns about its accuracy and potential impact on document quality.

Clinical notes are a significant source of interprovider communication, and questions about the potential impact of SR technology on note accuracy, clarity, and completeness warrant careful study. Previous studies have shown that incorrect information in the EHR is a contributing factor in up to 20% of EHR-related malpractice cases and that copy-and-paste in particular contributes to 8%–10%.[135,136] Among studies investigating how using SR to create notes affects the medical or linguistic quality of the document produced, results were mixed regarding whether SR technology was a help or a hindrance. Studies evaluating use of SR with templates or structured reporting were similarly mixed, suggesting that SR may not function well with current structured documentation methods.

Future directions

Based on the trends described previously, we have identified multiple aspects of SR-assisted documentation in need of further study, including, but not limited to, the following.

Impact on document quality and patient safety

Previous studies evaluating how SR affects documentation quality, particularly when used with structured reporting, indicate a need for additional research. Further investigation involving a broad range of input methods and documentation scenarios is needed to understand where and how SR can most appropriately and successfully be integrated with and used in the EHR. Studies have demonstrated that clinicians often underestimate errors generated by SR and do not have sufficient time to review their dictated documents.[84] Education and training about SR-associated errors emphasizing the importance of manual revision and editing is needed, as is investigation into the effect of SR use on patient safety and outcomes. Differences in notes' linguistic quality may also impact the performance of downstream natural language processing tasks.

SR usability and clinical workflow

Choice of documentation method plays a key role in clinicians' satisfaction and their ability to perform their work efficiently and effectively. Therefore, studies focusing on understanding clinicians' practices, preferences, and potential concerns before, throughout, and after the SR implementation process remain necessary. Many well-studied and proven tools exist for measuring the usability of EHR and other software applications.[137] Usability of EHR systems in general is a widely studied topic.[138,139] However, only in recent years has usability of EHR systems integrated with SR software become its own area of study.[25,79,83] SR technology allows physicians to dictate and edit their notes directly to the EHR without further assistance from traditional transcription or scribe services. While such a solution may reduce transcription costs, it may increase clerical burden to physicians already experiencing burnout.[140] The development of robust and standardized scales, questionnaires, and other tools tailored for evaluating SR usability and clinical workflow may help identify specific problems and possible solutions.

Standardization of evaluation methods and metrics

Given the rising prevalence of SR in medical settings, not only for documentation but also in other aspects of health care and delivery (eg, voice-enabled care), the need for standardized methods and measures for evaluating its accuracy and effectiveness is greater than ever. Many of the reviewed studies offered imprecise or overlapping definitions of similar, but distinct, productivity measures. While some measures (eg, time to report availability) may directly impact patient care, others may impact clinician workflow (eg, dictation time) or reimbursement process (eg, turnaround time); as such, these measures may be worth studying and reporting independently. In addition, the phrase "error rate" has been applied at both word and document levels, and many studies only report 1 or the other of these metrics, despite the fact that both are useful measures of SR accuracy and should be reported. A recent review identified a similar pattern in studies about radiology report accuracy.[4] Our findings suggest this trend is widespread, as it held true for articles related to emergency medicine, internal medicine, nursing, pathology, and more. This issue also exists in EHR usability analyses more broadly. For example, a 2017 review of literature related to EHR navigation found wide variation in the vocabulary used to discuss the same navigation actions and concepts.[141]

The economic impact of SR adoption has also been inconsistently reported, despite increased financial pressures faced by many health care institutions. Therefore, systematic means of assessing the financial impact of different documentation methods, both prospectively (eg, by developing an econometric model[49]) and retrospectively, are also needed.

Automatic error detection

Error detection systems intended for use with medical text have primarily been designed for written (typed) text.[142,143] Typing errors frequently involve misspellings, while SR errors involve words which are spelled correctly (as an SR engine will only propose words that exist in its dictionary) but are incorrect given the context. Many studies included analyses of errors in documents created with SR; however, comparatively little work has been done toward developing automated methods of detecting and/or correcting these errors. While this may be partially due to researchers' lack of access to the inner workings of the "black box" of vendor SR systems, previously attempted post-SR error detection tools have shown promise in their ability to identify, and thereby ultimately reduce, errors.[118–120] The development of automatic error detection methods for SR-generated medical text to improve document quality and patient safety therefore represents a significant research opportunity.

Limitations

Although we took steps to reduce the likelihood of having missed articles, including iteratively developing the search statements and screening the references of retrieved articles for additional papers, the database searches may not have yielded all relevant articles published during the time period of interest. Additionally, the included papers remain subject to reporting bias, and the heterogeneity of the included studies limited the ability to conduct a robust quantitative synthesis of their findings, even within individual research topics. Finally, the research topics we defined are subjective and based on the authors' prior knowledge and understanding of the field.

CONCLUSION

SR technology is increasingly used for clinical documentation. Research has been done to examine the effects of this technology on report accuracy and clinician productivity, largely focusing on a few clinical domains such as radiology and emergency medicine. However, a need remains for research to better understand SR usability when integrated with the EHR or other platforms, its impact on documentation quality, efficiency and cost, and user satisfaction over time and across different clinical settings. Standardized, comprehensive evaluation methods are also needed to help identify challenges and solutions for continued improvement.

FUNDING

This work was supported by Agency for Healthcare Research and Quality grant R01HS024264.

CONTRIBUTORS

The following authors contributed to the tasks of idea conception: LZ; development of database queries: SB, JH; development of annotation schema: SB, JH, LZ; article collection: SB, JH; article annotation: SB, JH, LW, ZK, LZ; development of analysis methods: SB, LZ; data analysis: SB; manuscript writing: SB, LZ; manuscript revision: SB, JH, LW, ZK, LZ.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank Kenneth H. Lai, MA, Leigh T. Kowalski, MS, and Clay Riley, MA for their annotation work.

Conflict of interest statement: None declared.

REFERENCES

1. Stewart B. *Front-End Speech 2014: Functionality Doesn't Trump Physician Resistance*. Orem, UT: KLAS; 2014.
2. Durling S, Lumsden J. Speech recognition use in healthcare applications. In: proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia; November 24–26, 2008; Linz, Austria.
3. 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.
4. Hammana I, Lepanto L, Poder T, Bellemare C, Ly MS. Speech recognition in the radiology department: a systematic review. *Health Inf Manag* 2015; 44 (2): 4–10.
5. Ajami S. Use of speech-to-text technology for documentation by healthcare providers. *Natl Med J India* 2016; 29 (3): 148–52.
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): e169–79.
7. 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 (3): 541–52.
8. Poder TG, Fisette J-F, Déry V. Speech recognition for medical dictation: overview in Quebec and systematic review. *J Med Syst* 2018; 42 (5): 89.
9. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Ann Intern Med* 2009; 151 (4): 264–9.
10. Feldman CA, Stevens D. Pilot study on the feasibility of a computerized speech recognition charting system. *Community Dent Oral Epidemiol* 1990; 18 (4): 213–5.
11. Klatt EC. Voice-activated dictation for autopsy pathology. *Comput Biol Med* 1991; 21 (6): 429–33.
12. Monnich G, Wetter T. Requirements for speech recognition to support medical documentation. *Methods Inf Med* 2000; 39 (1): 63–9.
13. Vorbeck F, Ba-Ssalamah A, Kettenbach J, Huebsch P. Report generation using digital speech recognition in radiology. *Eur Radiol* 2000; 10 (12): 1976–82.
14. Borowitz SM. Computer-based speech recognition as an alternative to medical transcription. *J Am Med Inform Assoc* 2001; 8 (1): 101–2.
15. Al-Aynati MM, Chorneyko KA. Comparison of voice-automated transcription and human transcription in generating pathology reports. *Arch Pathol Lab Med* 2003; 127 (6): 721–5.
16. Mohr DN, 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 (1): 85–93.
17. Issenman RM, Jaffer IH. Use of voice recognition software in an outpatient pediatric specialty practice. *Pediatrics* 2004; 114 (3): e290–3.
18. Rana DS, Hurst G, Shepstone L, Pilling J, Cockburn J, Crawford M. Voice recognition for radiology reporting: is it good enough? *Clin Radiol* 2005; 60 (11): 1205–12.
19. Ilgner J, Duwel P, Westhofen M. Free-text data entry by speech recognition software and its impact on clinical routine. *Ear Nose Throat J* 2006; 85 (8): 523–7.
20. Ichikawa T, Kitanosono T, Koizumi J, *et al*. Radiological reporting that combine continuous speech recognition with error correction by transcriptionists. *Tokai J Exp Clin Med* 2007; 32 (4): 144–7.
21. Alapetite A. Speech recognition for the anaesthesia record during crisis scenarios. *Int J Med Inform* 2008; 77 (7): 448–60.
22. Bhan SN, Coblentz CL, Norman GR, Ali SH. Effect of voice recognition on radiologist reporting time. *Can Assoc Radiol J* 2008; 59 (4): 203–9.
23. Gonzalez Sanchez MJ, Framinan Torres JM, Parra Calderon CL, Del Rio Ortega JA, Vigil Martin E, Nieto Cervera J. Application of Business Process Management to drive the deployment of a speech recognition system in a healthcare organization. *Stud Health Technol Inform* 2008; 136: 511–6.
24. Pezzullo JA, Tung GA, Rogg JM, Davis LM, Brody JM, Mayo-Smith WW. Voice recognition dictation: radiologist as transcriptionist. *J Digit Imaging* 2008; 21 (4): 384–9.
25. Derman YD, Arenovich T, Strauss J. Speech recognition software and electronic psychiatric progress notes: physicians' ratings and preferences. *BMC Med Inform Decis Mak* 2010; 10: 44.
26. Hawkins CM, Hall S, Hardin J, Salisbury S, Towbin AJ. Prepopulated radiology report templates: a prospective analysis of error rate and turnaround time. *J Digit Imaging* 2012; 25 (4): 504–11.
27. Dela Cruz JE, Shabosky JC, Albrecht M, *et al*. Typed versus voice recognition for data entry in electronic health records: emergency physician time use and interruptions. *West J Emerg Med* 2014; 15 (4): 541–7.
28. Hanna TN, Shekhani H, Maddu K, Zhang C, Chen Z, Johnson JO. Structured report compliance: effect on audio dictation time, report length, and total radiologist study time. *Emerg Radiol* 2016; 23 (5): 449–53.
29. Segrelles JD, Medina R, Blanquer I, Marti-Bonmati L. Increasing the efficiency on producing radiology reports for breast cancer diagnosis by means of structured reports. A comparative study. *Methods Inf Med* 2017; 56 (3): 248–60.
30. Rosenthal DI, Chew FS, Dupuy DE, *et al*. Computer-based speech recognition as a replacement for medical transcription. *AJR Am J Roentgenol* 1998; 170 (1): 23–5.
31. Chapman WW, Aronsky D, Fiszman M, Haug PJ. Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department. *Proc AMIA Symp* 2000: 131–5.
32. Lemme PJ, Morin RL. The implementation of speech recognition in an electronic radiology practice. *J Digit Imaging* 2000; 13 (S1): 153–4.
33. Ramaswamy MR, Chaljub G, Esch O, Fanning DD, vanSonnenberg E. Continuous speech recognition in MR imaging reporting: advantages, disadvantages, and impact. *AJR Am J Roentgenol* 2000; 174 (3): 617–22.
34. Callaway EC, Sweet CF, Siegel E, Reiser JM, Beall DP. Speech recognition interface to a hospital information system using a self-designed visual basic program: initial experience. *J Digit Imaging* 2002; 15 (1): 43–53.
35. Langer S. Radiology speech recognition: workflow, integration, and productivity issues. *Curr Probl Diagn Radiol* 2002; 31 (3): 95–104.
36. Langer SG. Impact of speech recognition on radiologist productivity. *J Digit Imaging* 2002; 15 (4): 203–9.
37. Gopakumar B, Wang S, Khasawneh MT, Cummings D, Srihari K. Reengineering radiology transcription process through Voice Recognition. Paper presented at: 2008 IEEE International Conference on Industrial Engineering and Engineering Management; December 8–11, 2008; Singapore.
38. Koivikko MP, Kauppinen T, Ahovuo J. Improvement of report workflow and productivity using speech recognition—a follow-up study. *J Digit Imaging* 2008; 21 (4): 378–82.
39. Hart JL, McBride A, Blunt D, Gishen P, Strickland N. Immediate and sustained benefits of a "total" implementation of speech recognition reporting. *Br J Radiol* 2010; 83 (989): 424–7.
40. Krishnaraj A, Lee JK, Laws SA, Crawford TJ. Voice recognition software: effect on radiology report turnaround time at an academic medical center. *AJR Am J Roentgenol* 2010; 195 (1): 194–7.
41. Strahan RH, Schneider-Kolsky ME. Voice recognition versus transcriptionist: error rates and productivity in MRI reporting. *J Med Imaging Radiat Oncol* 2010; 54 (5): 411–4.
42. Kang HP, Sirintrapun SJ, Nestler RJ, Parwani AV. Experience with voice recognition in surgical pathology at a large academic multi-institutional center. *Am J Clin Pathol* 2010; 133 (1): 156–9.
43. Singh M, Pal TR. Voice recognition technology implementation in surgical pathology: advantages and limitations. *Arch Pathol Lab Med* 2011; 135 (11): 1476–81.
44. Zick RG, Olsen J. Voice recognition software versus a traditional transcription service for physician charting in the ED. *Am J Emerg Med* 2001; 19 (4): 295–8.

45. Andriole KP, Prevedello LM, Dufault A, *et al*. Augmenting the impact of technology adoption with financial incentive to improve radiology report signature times. *J Am Coll Radiol* 2010; 7 (3): 198–204.
46. Prevedello LM, Ledbetter S, Farkas C, Khorasani R. Implementation of speech recognition in a community-based radiology practice: effect on report turnaround times. *J Am Coll Radiol* 2014; 11 (4): 402–6.
47. Ahlgrim C, Maenner O, Baumstark MW. Introduction of digital speech recognition in a specialised outpatient department: a case study. *BMC Med Inform Decis Mak* 2016; 16 (1): 132.
48. Pavlicek W, Muhm JR, Collins JM, Zavalkovskiy B, Peter BS, Hindal MD. Quality-of-service improvements from coupling a digital chest unit with integrated speech recognition, information, and Picture Archiving and Communications Systems. *J Digit Imaging* 1999; 12 (4): 191–7.
49. Reinus WR. Economics of radiology report editing using voice recognition technology. *J Am Coll Radiol* 2007; 4 (12): 890–4.
50. Corces A, Garcia M, Gonzalez V. Word recognition software use in a busy orthopaedic practice. *Clin Orthop Relat Res* 2004; 421: 87–90.
51. Henricks WH, Roumina K, Skilton BE, Ozan DJ, Goss GR. The utility and cost effectiveness of voice recognition technology in surgical pathology. *Mod Pathol* 2002; 15 (5): 565–71.
52. Williams DR, Kori SK, Williams B, *et al*. Journal Club: voice recognition dictation: analysis of report volume and use of the send-to-editor function. *AJR Am J Roentgenol* 2013; 201 (5): 1069–74.
53. Hayt DB, Alexander S. The pros and cons of implementing PACS and speech recognition systems. *J Digit Imaging* 2001; 14 (3): 149–57.
54. 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.
55. Lo MD, Rutman LE, Migita RT, Woodward GA. Rapid electronic provider documentation design and implementation in an academic pediatric emergency department. *Pediatr Emerg Care* 2015; 31 (11): 798–804.
56. Kauppinen TA, Kaipio J, Koivikko MP. Learning curve of speech recognition. *J Digit Imaging* 2013; 26 (6): 1020–4.
57. Hodgson T, Magrabi F, Coiera E. Efficiency and safety of speech recognition for documentation in the electronic health record. *J Am Med Informat Assoc* 2017; 24 (6): 1127–33.
58. Gosbee JW, Clay M. Human factors problem analysis of a voice-recognition computer-based medical record. Paper presented at: [1993] Computer-Based Medical Systems-Proceedings of the Sixth Annual IEEE Symposium; June 13–16, 1993; Ann Arbor, MI.
59. Zemmel NJ, Park SM, Schweitzer J, O'Keefe JS, Laughon MM, Edlich RF. Status of voicetype dictation for windows for the emergency physician. *J Emerg Med* 1996; 14 (4): 511–5.
60. Verheijen EJA, De Bruijn LM, Van Nes FL, Hasman A, Arends JW. The influence of time on error-detection. *Behav Inf Technol* 1998; 17 (1): 52–8.
61. Groschel J, Philipp F, Skonetzki S, Genzwurker H, Wetter T, Ellinger K. Automated speech recognition for time recording in out-of-hospital emergency medicine-an experimental approach. *Resuscitation* 2004; 60 (2): 205–12.
62. Rodger JA, Pendharkar PC. A field study of database communication issues peculiar to users of a voice activated medical tracking application. *Decis Support Syst* 2007; 43 (1): 168–80.
63. Alapetite A. Impact of noise and other factors on speech recognition in anaesthesia. *Int J Med Inform* 2008; 77 (1): 68–77.
64. McGurk S, Brauer K, Macfarlane TV, Duncan KA. The effect of voice recognition software on comparative error rates in radiology reports. *Br J Radiol* 2008; 81 (970): 767–70.
65. Zwemer J, Lenhart A, Kim W, Siddiqui KM, Siegel EL. Effect of ambient sound masking on the accuracy of computerized speech recognition. *Radiology* 2009; 252 (3): 691–5.
66. Du Toit J, Hattingh R, Pitcher R. The accuracy of radiology speech recognition reports in a multilingual South African teaching hospital. *BMC Med Imaging* 2015; 15: 8.
67. Suominen H, Johnson M, Zhou L, *et al*. Capturing patient information at nursing shift changes: methodological evaluation of speech recognition and information extraction. *J Am Med Inform Assoc* 2015; 22 (e1): e48–66.
68. Darcy R, Gallagher P, Moore S, Varghese D. The potential of using voice recognition in patient assessment documentation in mountain rescue. Paper presented at: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS); June 20–24, 2016; Dublin, Ireland, and Belfast, Northern Ireland.
69. Petroni M, Collet C, Fumai N, *et al*. An automatic speech recognition system for bedside data entry in an intensive care unit. Paper presented at: [1991] Computer-Based Medical Systems@m_Proceedings of the Fourth Annual IEEE Symposium; May 12–14, 1991; Baltimore, MD.
70. Sistrom CL, Honeyman JC, Mancuso A, Quisling RG. Managing predefined templates and macros for a departmental speech recognition system using common software. *J Digit Imaging* 2001; 14 (3): 131–41.
71. von Berg J. A grammar-based speech user interface generator for structured reporting. *Int Congr Ser* 2003; 1256: 887–92.
72. Liu D, Zucherman M, Tulloss WB Jr. Six characteristics of effective structured reporting and the inevitable integration with speech recognition. *J Digit Imaging* 2006; 19 (1): 98–104.
73. Yuhaniak Irwin J, Fernando S, Schleyer T, Spallek H. Speech recognition in dental software systems: features and functionality. *Stud Health Technol Inform* 2007; 129 (Pt 2): 1127–31.
74. Nagy M, Hanzlicek P, Zvarova J, *et al*. Voice-controlled data entry in dental electronic health record. *Stud Health Technol Inform* 2008; 136: 529–34.
75. Mehta A, McLoud TC. Voice recognition. *J Thorac Imaging* 2003; 18 (3): 178–82.
76. Carter-Wesley J. Voice recognition dictation for nurses. *J Nurs Adm* 2009; 39 (7–8): 310–2.
77. Hoyt R, Yoshihashi A. Lessons learned from implementation of voice recognition for documentation in the military electronic health record system. *Perspect Health Inf Manag* 2010; 7 (Winter): 1e.
78. Keskinen T, Melto A, Hakulinen J, *et al*. Mobile dictation for healthcare professionals. In: proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia; December 2–5, 2013; Luleå, Sweden.
79. Dawson L, Johnson M, Suominen H, *et al*. A usability framework for speech recognition technologies in clinical handover: a pre-implementation study. *J Med Syst* 2014; 38 (6): 56.
80. Awan S, Dunoyer E, Genuario K, *et al*. Using voice recognition enabled smartwatches to improve nurse documentation. Paper presented at: Systems and Information Engineering Design Symposium (SIEDS); April 27, 2018; Charlottesville, VA.
81. Mørck P, Langhoff TO, Christophersen M, Møller AK, Bjørn P. Variations in oncology consultations: how dictation allows variations to be documented in standardized ways. *Comput Support Coop Work* 2018; 27 (3–6): 539–68.
82. 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: 91–6.
83. Hodgson T, Magrabi F, Coiera E. Evaluating the usability of speech recognition to create clinical documentation using a commercial electronic health record. *Int J Med Inform* 2018; 113: 38–42.
84. 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): 1196–9.
85. 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* 2011; 197 (4): 923–7.
86. Chang CA, Strahan R, Jolley D. Non-clinical errors using voice recognition dictation software for radiology reports: a retrospective audit. *J Digit Imaging* 2011; 24 (4): 724–8.
87. Luetmer MT, Hunt CH, McDonald RJ, Bartholmai BJ, Kallmes DF. Laterality errors in radiology reports generated with and without voice recognition software: frequency and clinical significance. *J Am Coll Radiol* 2013; 10 (7): 538–43.

88. Hawkins CM, Hall S, Zhang B, Towbin AJ. Creation and implementation of department-wide structured reports: an analysis of the impact on error rate in radiology reports. *J Digit Imaging* 2014; 27 (5): 581–7.
89. Ringler MD, Goss BC, Bartholmai BJ. Syntactic and semantic errors in radiology reports associated with speech recognition software. *Stud Health Technol Inform* 2015; 216: 922.
90. Motyer RE, Liddy S, Torreggiani WC, Buckley O. Frequency and analysis of non-clinical errors made in radiology reports using the National Integrated Medical Imaging System voice recognition dictation software. *Ir J Med Sci* 2016; 185 (4): 921–7.
91. Goss FR, Zhou L, Weiner SG. Incidence of speech recognition errors in the emergency department. *Int J Med Inform* 2016; 93: 70–3.
92. Herman SJ. Accuracy of a voice-to-text personal dictation system in the generation of radiology reports. *AJR Am J Roentgenol* 1995; 165 (1): 177–80.
93. Suominen H, Ferraro G. Noise in speech-to-text voice: analysis of errors and feasibility of phonetic similarity for their correction. Paper presented at: proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013); December 4–6, 2013; Brisbane, Australia.
94. Zafar A, Overhage JM, McDonald CJ. Continuous speech recognition for clinicians. *J Am Med Inform Assoc* 1999; 6 (3): 195–204.
95. Devine EG, Gaehde SA, Curtis AC. Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. *J Am Med Inform Assoc* 2000; 7 (5): 462–8.
96. 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 Inform* 2004; 73 (9–10): 719–30.
97. McKoskey D, Boley D. *Error Analysis of Automatic Speech Recognition Using Principal Direction Divisive Partitioning*. Berlin: Springer; 2000.
98. David GC, Garcia AC, Rawls AW, Chand D. Listening to what is said–transcribing what is heard: the impact of speech recognition technology (SRT) on the practice of medical transcription (MT). *Sociol Health Illn* 2009; 31 (6): 924–38.
99. Garcia AC, David GC, Chand D. Understanding the work of medical transcriptionists in the production of medical records. *Health Informatics J* 2010; 16 (2): 87–100.
100. Nugues P, ElGuedj PO, Cazenave F, Ferrière B. Issues in the design of a voice man machine dialogue system generating written medical reports. Paper presented at: 1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; October 29–November 1, 1992; Paris, France.
101. Happe A, Pouliquen B, Burgun A, Cuggia M, Le Beux P. Automatic concept extraction from spoken medical reports. *Int J Med Inform* 2003; 70 (2–3): 255–63.
102. Jancsary J, Matiasek J, Trost H. Revealing the structure of medical dictations with conditional random fields. In: proceedings of the Conference on Empirical Methods in Natural Language Processing; October 25–27, 2008; Honolulu, HI.
103. Matiasek J, Jancsary J, Klein A, Trost H. Identifying segment topics in medical dictations. In: proceedings of the 2nd Workshop on Semantic Representation of Spoken Language; March 30, 2009; Athens, Greece.
104. Klann JG, Szolovits P. An intelligent listening framework for capturing encounter notes from a doctor-patient dialog. *BMC Med Inform Decis Mak* 2009; 9 (Suppl 1): S3.
105. Schreitter S, Klein A, Matiasek J, Trost H. Using domain knowledge about medications to correct recognition errors in medical report creation. In: proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents; June 5, 2010; Los Angeles, CA.
106. Do BH, Wu AS, Maley J, Biswal S. Automatic retrieval of bone fracture knowledge using natural language processing. *J Digit Imaging* 2013; 26 (4): 709–13.
107. Zhou L, Suominen H, Hanlen L. Evaluation Data and Benchmarks for Cascaded Speech Recognition and Entity Extraction. In: proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia; October 26–30, 2015; Brisbane, Australia.
108. Suominen H, Zhou L, Hanlen L, Ferraro G. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR Med Inform* 2015; 3 (2): e19.
109. Uddin MM, Huynh N, Vidal JM, Taaffe KM, Fredendall LD, Greenstein JS. Evaluation of Google's voice recognition and sentence classification for health care applications. *Eng Manag J* 2015; 27 (3): 152–62.
110. Nugues P, Bellegarde E, Coci H, Pontier R, Trognon P. A voice man machine dialogue system generating written medical reports. Paper presented at: proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Volume 13; 31 October–3 November, 1991.
111. Pakhomov SV. Modeling filled pauses in medical dictations. In: proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics; June 20–26, 1999; College Park, MD.
112. Pakhomov S, Schonwetter M, Bachenko J. *Generating Training Data for Medical Dictations*. 2002. http://aclweb.org/anthology/N01-1017. Accessed July 10, 2017.
113. Paulett JM, Langlotz CP. Improving language models for radiology speech recognition. *J Biomed Inform* 2009; 42 (1): 53–8.
114. Sethy A, Georgiou PG, Ramabhadran B, Narayanan S. An iterative relative entropy minimization-based data selection approach for n-gram model adaptation. *IEEE Trans Audio Speech Lang Process* 2009; 17 (1): 13–23.
115. Edwards E, Salloum W, Finley GP, *et al*. Medical speech recognition: reaching parity with humans. In: Karpov A, Potapava R, Mporas I, eds. *SPECOM 2017. LNCS (LNAI)*. Vol. 10458. Cham, Switzerland: Springer; 2017: 512–24.
116. Salloum WE, E, Ghaffarzadegan S, Suendermann-Oeft, D, Miller, M. Crowdsourced continuous improvement of medical speech recognition. Paper presented at: the AAAI-17 Workshop on Crowdsourcing, Deep Learning, and Artificial Intelligence Agents; February 5–7, 2017, San Francisco, CA.
117. Paats A, Alumae T, Meister E, Fridolin I. Retrospective analysis of clinical performance of an Estonian speech recognition system for radiology: effects of different acoustic and language models. *J Digit Imaging* 2018; 31 (5): 615–21.
118. Voll K, Atkins S, Forster B. Improving the utility of speech recognition through error detection. *J Digit Imaging* 2008; 21 (4): 371–7.
119. Minn MJ, Zandieh AR, Filice RW. Improving radiology report quality by rapidly notifying radiologist of report errors. *J Digit Imaging* 2015; 28 (4): 492–8.
120. Lybarger K, Ostendorf M, Yetisgen M. Automatically detecting likely edits in clinical notes created using automatic speech recognition. *AMIA Annu Symp Proc* 2017; 2017: 1186–95.
121. Parente R, Kock N, Sonsini J. An analysis of the implementation and impact of speech-recognition technology in the healthcare sector. *Perspect Health Inf Manag* 2004; 1: 5.
122. Alapetite A, Boje Andersen H, Hertzum M. Acceptance of speech recognition by physicians: a survey of expectations, experiences, and social influence. *Int J Hum Comput Stud* 2009; 67 (1): 36–49.
123. Viitanen J. Redesigning digital dictation for physicians: a user-centred approach. *Health Informatics J* 2009; 15 (3): 179–90.
124. Fratzke J, Tucker S, Shedenhelm H, Arnold J, Belda T, Petera M. Enhancing nursing practice by utilizing voice recognition for direct documentation. *J Nurs Adm* 2014; 44 (2): 79–86.
125. Clarke MA, King JL, Kim MS. Toward successful implementation of speech recognition technology: a survey of SRT utilization issues in healthcare settings. *South Med J* 2015; 108 (7): 445–51.
126. Lyons JP, Sanders SA, Fredrick Cesene D, Palmer C, Mihalik VL, Weigel T. Speech recognition acceptance by physicians: a temporal replication of a survey of expectations and experiences. *Health Informatics J* 2016; 22 (3): 768–78.
127. McEnery KW, Suitor CT, Hildebrand S, Downs R. RadStation: client-based digital dictation system and integrated clinical information display with an embedded Web-browser. *Proc AMIA Symp* 2000: 561–4.

128. White KS. Speech recognition implementation in radiology. *Pediatr Radiol* 2005; 35 (9): 841–6.
129. Zheng K, Mei Q, Yang L, Manion FJ, Balis UJ, Hanauer DA. Voice-dictated versus typed-in clinician notes: linguistic properties and the potential implications on natural language processing. *AMIA Annu Symp Proc* 2011; 2011: 1630–8.
130. Sistrom CL. Conceptual approach for the design of radiology reporting interfaces: the talking template. *J Digit Imaging* 2005; 18 (3): 176–87.
131. Boland GW, Guimaraes AS, Mueller PR. Radiology report turnaround: expectations and solutions. *Eur Radiol* 2008; 18 (7): 1326–8.
132. Zhou L, Blackley SV, Kowalski L, *et al*. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. *JAMA Network Open* 2018; 1 (3): e180530.
133. Al Hadidi S, Upadhaya S, Shastri R, Alamarat Z. Use of dictation as a tool to decrease documentation errors in electronic health records. *J Community Hosp Intern Med Perspect* 2017; 7 (5): 282–6.
134. Green HD. Adding user-friendliness and ease of implementation to continuous speech recognition technology with speech macros: case studies. *J Healthc Inf Manag* 2004; 18 (4): 40–8.
135. Ruder D. Malpractice claims analysis confirms risks in EHRs. *Patient Saf Qual Healthc* 2014; 11 (1): 20–3.
136. Graber ML, Siegal D, Riah H, Johnston D, Kenyon K. Electronic health record-related events in medical malpractice claims. *J Patient Saf* 2015 Nov 6 [E-pub ahead of print].
137. Kirakowski J, Corbett M. SUMI: the software usability measurement inventory. *Br J Educ Technol* 1993; 24 (3): 210–2.
138. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. *J Biomed Inform* 2011; 44 (6): 1056–67.
139. Walji MF, Kalenderian E, Piotrowski M, *et al*. Are three methods better than one? A comparative assessment of usability evaluation methods in an EHR. *Int J Med Inform* 2014; 83 (5): 361–7.
140. Wright AA, Katz IT. Beyond burnout—redesigning care to restore meaning and sanity for physicians. *N Engl J Med* 2018; 378 (4): 309–11.
141. Roman LC, Ancker JS, Johnson SB, Senathirajah Y. Navigation in the electronic health record: a review of the safety and usability literature. *J Biomed Inform* 2017; 67: 69–79.
142. Tolentino HD, Matters MD, Walop W, *et al*. A UMLS-based spell checker for natural language processing in vaccine safety. *BMC Med Inform Decis Mak* 2007; 7: 3.
143. Lai KH, Topaz M, Goss FR, Zhou L. Automated misspelling detection and correction in clinical free-text records. *J Biomed Inform* 2015; 55: 188–95.