Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician conversations

Brian D. Tran[1], Ramya Mangu[1], Ming Tai-Seale[2], Jennifer Elston Lafata[3,4], Kai Zheng[1]
[1]University of California Irvine, Irvine, CA, USA; [2]University of California San Diego, La Jolla, USA; [3]University of North Carolina at Chapel Hill, Chapel Hill, USA, [4]Henry Ford Health System, Detroit, USA

Abstract

One promising solution to address physician data entry needs is through the development of so-called "digital scribes," or tools which aim to automate clinical documentation via automatic speech recognition (ASR) of patient-clinician conversations. Evaluation of specialized ASR models in this domain, useful for understanding feasibility and development opportunities, has been difficult because most models have been under development. Following the commercial release of such models, we report an independent evaluation of four models, two general-purpose, and two for medical conversation with a corpus of 36 primary care conversations. We identify word error rates (WER) of 8.8%-10.5% and word-level diarization error rates (WDER) ranging from 1.8%-13.9%, which are generally lower than previous reports. The findings indicate that, while there is room for improvement, the performance of these specialized models, at least under ideal recording conditions, may be amenable to the development of downstream applications which rely on ASR of patient-clinician conversations.

Introduction

Reliance on clinicians as a key data entry source remains a fundamental healthcare informatics challenge[1]. In addition to capturing data necessary for the direct provision of care, clinician users are often obligated to capture data on behalf of a wide variety of secondary users, such as public health researchers or quality improvement specialists[2,3]. Complicating these circumstances are reports of poor EHR usability and the need to fulfill myriad billing and administrative documentation requirements[3,4]. Pressures from these sources contribute to a high degree of documentation burden faced by currently practicing clinicians.

One promising solution to address this problem is through the development of so-called "digital scribes." These tools seek to automate, at least in part, the process of clinical documentation through the use of automatic speech recognition (ASR) and natural language processing (NLP) technologies to capture patient-clinician conversations in the examination room[5]. If successful, these tools may shift existing documentation burden away from clinicians, allowing them to better focus on patient care-related activities.

Interest in digital scribes has resulted in efforts from academic as well as industry groups. Prototype systems[6,7], empirical studies and commentaries related to the concept[8–10], and development of NLP techniques[11,12] have been published by academic researchers. Companies such as Microsoft[13], Google[14–16], and Amazon[17], among others[18,19] have also begun to publish work relevant to digital scribe development. Further, perhaps in recognition of the value of working with natural language in the clinical setting, Microsoft recently acquired Nuance Communications[20], a vendor for clinical dictation software. These efforts indicate significant investment into the development of tools and techniques which can be used to create a digital scribe.

Despite these efforts, the digital scribe concept and its associated technologies are still considered nascent and rapidly evolving[21]. Two currently outstanding issues important to the development of these systems are: (1) the performance of speech recognition and (2) speaker recognition technologies, which are likely necessary for machine understanding of natural language in patient-clinician conversations[21,22]. Here, speech recognition performance references the accuracy and usefulness of words transcribed from conversation, while speaker recognition performance focuses on differentiating between speakers (also known as speaker diarization) and understanding which speaker is a clinician or patient (role attribution). While there are a few works that have published performance metrics for both speech recognition[15] and speaker recognition[19,23], most models have been under development and have not been made readily available for independent evaluation by researchers. As a result, it has been difficult to independently ascertain performance of current state-of-the-art technologies in this domain.

Following exploratory work in conversational speech recognition for medical conversations[6], a 2018 evaluation by Kodish-Wachs investigated the feasibility of commercialized general-purpose ASR models for the transcription of simulated patient-clinician conversations[24]. They concluded that, given modest performance figures of 35% to 65%

word error rate (WER), improvements to ASR approaches are needed before further use in the clinical setting. Since the publication of this study, however, advancements to medical conversation ASR and patient/clinician speaker recognition have been published[13,19,21,23,25,26]. Companies, such as Google and Amazon, have since released ASR solutions which have been marketed as capable of transcribing patient-clinician conversations and conducting speaker diarization[27,28], the latter of which references the task of differentiating speakers in audio. An updated understanding of the general performance of these tools, as well as a confirmation of whether these specialized models improve ASR performance and speaker recognition, may help guide clinician and developer thinking about how these technologies could be deployed or improved to better support clinician users or improve data quality.

In this study, we report an independent evaluation of four commercially available solutions for ASR: two marketed for patient-clinician conversations, and two marketed for general multi-speaker audio. We do this by obtaining a performance snapshot with a corpus of simulated patient-clinician conversations and then identify potential performance and design implications relevant to the capture and processing of patient-clinician conversations.

Learning Objective

Learn about the current performance of state-of-the-art automatic speech recognition systems which focus on capturing patient-clinician conversations and the understand how their performance may impact the development of speech-based clinical documentation tools.

Methods

Generation of reference audio and transcripts

We conducted our evaluation on 36 professionally transcribed audio recordings of conversations between patients and primary care clinicians. Patients were aged 50-80 and were visiting a primary care physician in a 26-clinic ambulatory healthcare system in the Midwest United States. The data was created as part of the Mental Health Discussions Study by Tai-Seale et al.[29], and has been used in prior work such as analyzing the delivery of preventative services[30–39] as well as for creating topic detection[40] and emotion detection[41] models using patient-clinician conversations. For this study, the subsampled conversations were manually de-identified per HIPAA safe harbor standards.

To reduce the impact of recording-related factors (e.g., poor microphone quality, high background noise), and speaker-related factors (e.g., non-native English speakers), we re-enacted audio recordings in a quiet interview studio. Two native English speakers, one male and one female, read de-identified transcripts into a high-quality desktop microphone. Speakers completed the reenactment while within three feet of a Blue Yeti microphone (Logitech, Lausanne, Switzerland). Recordings were captured at 44100hz, in single channel audio, and were encoded in a lossless format. Recordings ranged from 12 to 55 minutes in length.

To assess performance on conversation which may be relevant for clinical documentation, BDT, a medical student, scanned transcripts for utterances potentially useful for generating documentation related to topics within a primary care physician task list[42]. Relevant utterances, which included information such as question answer pairs, notes on lab values, physical examination findings as well as information for care coordination were highlighted and were used to segment data for evaluation. Examples of utterances marked as relevant included statements such as, “Yes, my father had heart disease,” and “I exercise twice a week.”

ASR Engine selection and ASR data generation

For this study, we selected four ASR engines based on their accessibility and prior performance as measured by word error rate (WER) with a subsample of the dataset in early 2022. For models tailored to medical conversation, we selected Google Speech-to-Text “medical_conversation” (Google, Mountain View, USA) and Amazon Transcribe Medical “Primary care” & “Conversation” (Amazon, Seattle, USA). To contrast with general-purpose models suitable for general audio and multiple speakers, we also evaluated Google Speech-to-Text “Video” and Amazon Transcribe “General.” Hereafter, we refer to these systems as Google Medical Conversation ASR, Amazon Medical Conversation ASR, Google General ASR, and Amazon General ASR, respectively. We excluded the available Google Speech-to-Text “Default” transcription model due lower performance relative to Google General ASR. We also attempted to obtain access to Nuance Dragon Medical SpeechKit (Nuance Communications, Burlington, MA) and Deepscribe.ai (Deepscribe, San Francisco, CA) by contacting the respective institutions for research purposes. However, despite requests, we were unable to obtain their software for evaluation.

To generate the ASR data, the de-identified audio recordings were uploaded to cloud storage and transcribed in March 2022. All transcription requests were standardized to include specification of the transcription to use an English language model, to perform speaker recognition with two speakers, and to provide an output with proper casing and punctuation.

Performance evaluation

We conducted our evaluation in two phases: (1) assessment of general ASR performance and (2) assessment of ASR performance on potentially meaningful information expressed in patient-clinician conversations. In the first phase, we assessed the general overlap of words between reference and ASR transcripts by WER. Because the selected ASR engines also had a speaker diarization feature, which attempts to separate speakers based on recorded vocal qualities, we also chose to assess performance on the task of recognizing different speakers within each conversation. We evaluated these systems using word-level diarization error rate (WDER), a metric which has been used previously by Shafey et al. for use in evaluating joint ASR and speaker diarization systems[23]. WDER characterizes the overlap of relative speaker labels between reference transcripts and ASR transcripts. As speaker labels for Amazon Medical Conversation ASR, Google General ASR, and Amazon General ASR were provided as "0" or "1," we assigned speaker labels in a way that assumed the best potential performance in this aspect for each system (i.e., minimizing incorrect role attribution). To do this, we mapped the provided speaker labels on a per-conversation and per-engine basis to either "Doctor" or "Patient" in a way that minimized diarization errors. For Google Medical Conversation ASR, which provided embedded speaker labels of "spkr:patient" or "spkr:provider" with each utterance, we also evaluated the system using the embedded speaker labels.

In the second phase of our evaluation, we assessed the transcription performance of potentially meaningful information present in patient-clinician conversations. First, we stratified errors based on human annotated documentation relevance generated from the procedure noted above. Second, we compared the capture of medical concepts as labeled by an automated annotator[43] using Logical Observation Identifiers Names and Codes (LOINC) and SNOMED CT ontologies. Here, we report results as precision, recall, and F1 scores relative to reference transcripts. Third, we report common words in reference transcripts affected by deletion and substitution errors. For all analyses, we confirmed the veracity of the results with a manual review of a subset of errors within reference transcripts.

Results

Automated transcription performance as measured by WER across evaluated models differed by less than 2%, ranging from 8.8% by Google General ASR to 10.5% by Amazon Medical Conversation ASR (**Table 1**). When WER was broken down into constituent substitution, deletion, and insertion rates, substitution errors were the most common type of error across all models. General purpose models achieved slightly smaller substitution error rates relative to their medical conversation counterparts. The inverse was true for deletion errors, where general purpose models obtained slightly higher error rates.

Engine Type	WER (%)	Substitution Rate (%)	Deletion Rate (%)	Insertion Rate (%)	WDER (%)	Total Words as Transcribed
Amazon General ASR	9.4	4.3	3.6	1.4	1.8	134,649
Amazon Medical Conversation ASR	10.5	6.5	2.4	1.6	5.1	134,910
Google General ASR	8.8	5.2	2.5	1.1	6.3	134,240
Google Medical Conversation ASR	9.1	5.7	1.0	2.4	13.9*	135,909

Table 1. Word error rate (WER), WER Components, and word diarization error rate (WDER) for four automatic speech recognition (ASR) engines with speaker diarization on patient-clinician conversations. *With embedded "patient" and "doctor" speaker labels.

Performance of speaker labels at the word level ranged from 1.8% by Amazon General ASR to 13.9% by Google Medical Conversation ASR (**Table 1**). Evaluated general-purpose models achieved a lower WDER than their specialized counterparts. Amazon General ASR achieved the lowest WDER of the four models. Google Medical Conversation ASR, despite providing patient/clinician speaker labels embedded in its output, generated the highest WDER of 13.9%, well above its corresponding general-purpose model of 6.3%. Ignoring embedded role labels and applying role labels which would maximize diarization performance, Google Medical Conversation ASR achieved 14.0% WDER.

The next stage of our analysis focused on transcription performance of potentially important information categories. When error rates were segmented by speakers, the clinician speaker, in general, had a slightly higher WER relative to the patient speaker across all models (**Table 2**). WDER typically differed between speakers by less than 1%, except for Amazon Medical Conversation ASR, which had a 2.8% increase in clinician WDER. When focused on utterances potentially relevant for clinical documentation, WER differed by 1% or less across all models. WDER, on the other hand, appeared to be lower across all models for phrases relevant to documentation and ranged from a 1% difference with Amazon General ASR and a 4.5% difference with Google Medical Conversation ASR.

Engine Type	Category	Segment	WER (%)	WDER (%)	Total Words as Transcribed
Amazon General ASR	Speaker	Patient	8.0	1.8	61,803
		Clinician	10.6	1.9	72,846
	Documentation relevance	Yes	10.2	1.0	28,440
		No	9.2	2.1	106,209
Amazon Medical Conversation ASR	Speaker	Patient	8.7	3.6	61,930
		Clinician	12.1	6.4	72980
	Documentation relevance	Yes	11.1	2.6	28,372
		No	10.4	5.7	106,538
Google General ASR	Speaker	Patient	7.3	6.4	61,653
		Clinician	10.1	6.2	72,587
	Documentation relevance	Yes	9.2	4.4	28,208
		No	8.7	6.8	106,032
Google Medical Conversation ASR	Speaker	Patient	7.9	13.4	62,300
		Clinician	10.1	14.6	73,609
	Documentation relevance	Yes	9.3	10.5	28,492
		No	9.1	15.0	107,417

Table 2. Performance of four ASR engines with speaker diarization on patient-clinician conversations, stratified by additional categories. WER: Word error rate, WDER: Word diarization error rate.

To supplement the prior analysis, we also assessed the transcription fidelity in terms of medically pertinent concepts as labeled by an automatic annotator. From this assessment, all evaluated models achieved low recall while retaining high precision **(Table 3)**. In other words, many of the medically pertinent concepts in the reference were not correctly transcribed by ASR; but when concepts were captured, the concepts tended to be captured correctly. Qualitatively, concepts that were commonly not annotated relative to reference included answers to physician questions, such as “Don’t Know,” “Can’t Do,” or “I don’t know,” as well as concepts that were transcribed in a way that was not detected by the annotator (e.g., “blood pressure” in the reference → “bp” after ASR was not recognized).

Engine Type	Recall	Precision	F1 Score
Amazon General ASR	0.49	0.96	0.65
Amazon Medical Conversation ASR	0.48	0.95	0.64
Google General ASR	0.49	0.96	0.65
Google Medical Conversation ASR	0.49	0.95	0.64

Table 3. Performance of four ASR engines by recall, precision, and F1 score of medically pertinent concepts relative to a standardized reference.

Perhaps not surprisingly, words frequently affected by deletion and substitution errors appeared to be common words which are often useful for interpretation of phrases in conversation, rather than specific medical terminology (which may be more common in clinician dictations). We identified 10 of the most common words affected by deletion and substitution errors for each engine, excluding substitution errors that were a result of different spelling standards across engine types (e.g., “mhm” in Amazon Medical Conversation ASR vs. “um-hum” in Google Medical Conversation ASR or “blood pressure” in Amazon General ASR vs. “bp” in Amazon Medical Conversation ASR) (**Table 4**). From this procedure, we identified that common words such as “I,” “you,” “yeah,” or “okay,” were often deleted or substituted across all engine types.

Common Deletion Errors							
Amazon General ASR		**Amazon Medical Conversation ASR**		**Google General ASR**		**Google Medical Conversation ASR**	
Word	Count	Word	Count	Word	Count	Word	Count
okay	492	i	151	i	159	i	60
yeah	411	and	109	a	88	a	42
i	166	a	95	you	66	you	36
no	154	you	75	the	59	and	35
you	133	the	62	and	52	the	33
a	92	it	62	it	51	that	26
it	66	that	40	yeah	32	in	16
know	65	okay	36	okay	25	yeah	15
yes	63	yeah	36	that	20	okay	14
the	55	of	28	in	20	of	14

Common Substitution Errors											
Amazon General ASR			**Amazon Medical Conversation ASR**			**Google General ASR**			**Google Medical Conversation ASR**		
Word	Count	Common substitutions	Word	Count	Common substitutions	Word	Count	Common substitutions	Word	Count	Common substitutions
too	69	to, two	and	185	on, in	too	119	to, two	a	107	uh, the
okay	67	all, right	i	128	they, you	no	93	know, yeah	and	82	in, uh
yeah	61	all, so	a	116	the, uh	i	74	it, a	i	79	they, uh
you	56	you're, you've	no	77	know, now	and	69	in, right	that	67	the, uh
i	55	they, it	too	76	to, two	in	59	and, an	is	54	it's, here's
no	47	know, all	is	73	it's, here's	is	58	it's, here's	the	52	a, uh
a	46	the, an	are	70	you're, they're	it	54	it's, a	doctor	51	doctor's, doctors
and	45	in, an	the	67	that, a	you	54	he, me	you	48	you're, you'd
the	39	a, this	you	66	you're, he	a	52	the, an	it	47	it's, you
are	36	you're, we're	it	64	a, it's	that	48	the, but	in	42	and, it

Table 4. Top 10 common words affected by deletion and substitution errors following ASR across four engines, excluding words with different spelling standards across engines. Substitution errors panel showcases two most common substitutions.

Discussion

This study compared the performance of four contemporary ASR transcription models, two for general-purpose use, and two for medical conversations, on a corpus of simulated patient-physician conversations. From the evaluation, WER across models were similar and ranged from 8.8% and 10.5%. We also report WDER range from 1.8% to 13.9%. In our error analysis, phrases which were marked as potentially relevant for documentation were transcribed by ASR with similar fidelity as non-relevant phrases. Words in the conversations most frequently affected by ASR errors appeared to be useful for the interpretation of statements, rather than medical terminology. These findings suggest that, despite progress in the performance of ASR technology for the transcription of patient-clinician conversations, which may enable new opportunities for applications, there is room for additional improvement. The nature and degree of improvement that is needed is likely contingent on the performance needs of technologies which will process the automatically generated transcripts.

We report that performance as measured by WER and WDER has improved over prior evaluations. First our obtained WER was lower than the 35% WER figure previously reported by Kodish-Wachs[24] and the 34% WER achieved by Schloss and Konam with Google ASR Video[19]. Our results also indicate improvement over the 18% WER reported by Chiu et al. in the development of Google's medical conversation ASR model[15]. Similarly, WDER achieved by all engines were below 14%, which was lower than Shafey et al.'s WDER of 15.8% using a baseline system and a corpus of patient-clinician conversations[23]. Because this study's evaluation dataset was generated under ideal recording conditions, we believe that our reported figures likely represent an upper bound in potential performance. More realistic clinic conditions with noise, non-native speakers, and lower-quality recording devices would likely increase ASR error rates.

Interestingly, Google Medical Conversation ASR, which provided embedded speaker labels as part of its output, obtained the highest WDER of all evaluated engines. This result was particularly striking when compared the 2.2% reported by Shafey et al. of Google for their joint ASR and speaker diarization model[23]. Upon scanning of errors in speaker tags, we observed that such tags were incorrectly flipped for large portions of the conversation, which was comparable to an observed issue that Shafey et al. reported in their study. This finding suggests that, if the evaluated model and the model referenced by Shafey et al. are similar, the observed issue may not have been resolved.

From our analysis of medical concepts and words frequently affected by ASR errors, we identified an area that may require additional improvement: the handling of pronouns and agreement tokens. These words, if modified, have the potential to drastically alter the meaning of utterances in conversation, yet were commonly affected by deletion and substitution errors. We believe that this could potentially affect the ability of automated approaches for natural language understanding of statements from patient-clinician conversation. Indeed, at least one study has attempted to address a consequence of this issue: in their symptom recognition model, Rajkomar et al. also focused on identifying whether a patient experienced a symptom given a conversation, as compared to labeling symptoms alone[14]. Future evaluation work could more closely attempt to understand how errors on words such as pronouns and agreement tokens may affect the accuracy of information extracted from patient-clinician conversation.

Despite being advertised as tuned for the capture of medical conversation, the medical conversation models achieved similar performance to general-purpose models in terms of capturing relevant phrases for documentation, capturing medical concepts, and generating common substitution and deletion errors. We speculate that these results were achieved due the fact that the primary care conversations used for evaluation largely contained general conversational English language instead of dense medical jargon (e.g., as in a dictated clinical note). Phrases which were deemed relevant for documentation were often stated in plain language (e.g., "I exercise twice a week."). The non-specialized language would likely be similarly captured across both general and specialized models. We believe that corpora with conversations that include dense and complex medical terminology (e.g., readings of medical reports) may yield different results. One future direction for evaluating performance between these general purpose and specialized models can focus on the capture of specific elements of information in conversation outside of medical terminology that are pertinent to clinical documentation needs.

It is currently unclear whether the reported performance figures suggest that existing ASR and speaker diarization technologies are now ready for digital scribing systems. The error rates of medical conversation ASR are now comparable to human error rates on non-medical standardized conversation corpora, such as 5.9% WER for *Switchboard* and 11.3% WER for *CallHome*[44]. While specific conclusions cannot be drawn from this comparison because of differences in content and intended use of the conversation data, the general similarity in WER suggests that the performance of the evaluated models may now be amenable to new opportunities for useful applications which process patient-clinician conversations captured by ASR. Despite this possibility, the technical approaches to modeling scribing-related tasks are still nascent in maturity and greatly vary in task scope and data used to generate results[11,12,21]. As a result, the fidelity of ASR, particularly as measured by the general performance metrics used in this study, may not map to overall system performance in a straightforward manner. This is further complicated by the potentially impactful effects of poorly performing ASR for digital scribing systems: any mis-captured information may result in a reduction in time and effort savings for the clinician user or may even introduce new sources of medical error. Additional development and standardization in ASR, speaker recognition, and NLP technologies for digital scribing systems as well as approaches to evaluate those technologies in appropriate contexts will be needed in order to answer whether ASR and speaker diarization technologies for patient-clinician conversations has the appropriate fidelity for real-world clinic use.

There are several limitations to this study. First, the original conversational transcripts were collected from a study which focused on conversations with patients that had the possibility of discussing mental health concerns. This may reduce generalizability. Upon manual review, however, we found that the majority of the subsampled conversations did not discuss mental health as a chief complaint. Second, the original transcripts were collected from a single healthcare system in a primary care setting. Conversations between providers at a different health system or different specialty may drastically differ in content and length. Third, while both medical conversation models are from vendors with an established publication record in state-of-the-art ASR, it is possible that there are other systems still in development, were not considered, or were not accessible to the researchers, which may have better performance with the evaluation data. Fourth, as stated above, the evaluation data were created under different conditions relative to real-world clinical conditions (i.e., native-English speakers, good-quality recording equipment, quiet recording settings). While this step was taken to limit audio quality and speaker characteristics as a bottleneck to potential performance, the results likely limit generalizability to real-world settings.

Conclusion

In this study, we assessed the current feasibility of ASR of patient-clinician conversations for digital scribe systems by evaluating the performance of specialized commercial speech recognition systems on ASR and speaker diarization. We report that, while there is room for improvement, error rates, at least in ideal conditions, have improved relative to prior evaluations, potentially creating opportunities for the development of downstream applications which rely on ASR of patient-clinician conversations.

Acknowledgements

The authors would like to thank Tera Reynolds for their assistance in generating the recordings used in this study.

References

1. Hripcsak G, Bloomrosen M, FlatelyBrennan P, *et al.* Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting. *J Am Med Inform Assoc* 2014;21:204–11.
2. Shortliffe EH. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. 3rd ed. New York, NY: : Springer 2006.
3. National Academies of Sciences, Engineering, and Medicine, National Academy of Medicine, Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being. *Taking action against clinician burnout*. Washington, D.C., DC: : National Academies Press 2020.
4. Massachusetts medical society: A crisis in health care: A call to action on physician burnout. https://www.massmed.org/Publications/Research,-Studies,-and-Reports/A-Crisis-in-Health-Care--A-Call-to-Action-on--Physician-Burnout/ (accessed 29 Jul 2022).
5. Coiera E, Kocaballi B, Halamka J, *et al.* The digital scribe. *NPJ Digit Med* 2018;1:58.
6. 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.
7. Jeblee S, Khan Khattak F, Crampton N, *et al.* Extracting relevant information from physician-patient dialogues for automated clinical note taking. In: *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2019. doi:10.18653/v1/d19-6209
8. Quiroz JC, Laranjo L, Kocaballi AB, *et al.* Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digit Med* 2019;2:114.
9. Tran BD, Chen Y, Liu S, *et al.* How does medical scribes' work inform development of speech-based clinical documentation technologies? A systematic review. *J Am Med Inform Assoc* 2020;27:808–17.

10. Kocaballi AB, Coiera E, Tong HL, *et al.* A network model of activities in primary care consultations. *J Am Med Inform Assoc* 2019;26:1074–82.
11. Bhatia P, Lin S, Gangadharaiah R, *et al.*, editors. *Proceedings of the first workshop on natural language processing for medical conversations*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2020.
12. *Proceedings of the second workshop on natural language processing for medical conversations*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2021.
13. Enarvi S, Amoia M, Del-Agua Teba M, *et al.* Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In: *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2020. doi:10.18653/v1/2020.nlpmc-1.4
14. Rajkomar A, Kannan A, Chen K, *et al.* Automatically charting symptoms from patient-physician conversations using machine learning. *JAMA Intern Med* 2019;179:836–8.
15. Chiu C-C, Tripathi A, Chou K, *et al.* Speech Recognition for Medical Conversations. In: *Interspeech 2018*. ISCA: : ISCA 2018. doi:10.21437/interspeech.2018-40
16. Shafran I, Du N, Tran L, *et al.* The medical scribe: Corpus development and model performance analyses. arXiv [cs.CL]. 2020.http://arxiv.org/abs/2003.11531
17. Sunkara M, Ronanki S, Dixit K, *et al.* Robust prediction of punctuation and truecasing for medical ASR. In: *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2020. doi:10.18653/v1/2020.nlpmc-1.8
18. Mani A, Palaskar S, Konam S. Towards understanding ASR error correction for medical conversations. In: *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2020. doi:10.18653/v1/2020.nlpmc-1.2
19. Schloss B, Konam S. Towards an automated SOAP note: Classifying utterances from medical conversations. arXiv [cs.CL]. 2020.http://arxiv.org/abs/2007.08749
20. Ross C, Feuerstein A, Garde D, Cohrs R, Florko N, Molteni M. Microsoft closes $16B acquisition of Nuance, launching a bold experiment in health AI. STAT. 2022.https://www.statnews.com/2022/03/04/microsoft-closes-16-billion-acquisition-nuance-communications/ (accessed 8 Mar 2022).
21. van Buchem MM, Boosman H, Bauer MP, *et al.* The digital scribe in clinical practice: a scoping review and research agenda. *NPJ Digit Med* 2021;4:57.
22. Park TJ, Kanda N, Dimitriadis D, *et al.* A review of speaker diarization: Recent advances with deep learning. *Comput Speech Lang* 2022;72:101317.
23. Shafey LE, Soltau H, Shafran I. Joint Speech Recognition and Speaker Diarization via Sequence Transduction. In: *Interspeech 2019*. ISCA: : ISCA 2019. doi:10.21437/interspeech.2019-1943
24. Kodish-Wachs J, Agassi E, Kenny P 3rd, *et al.* A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. *AMIA Annu Symp Proc* 2018;2018:683–9.
25. Miner AS, Haque A, Fries JA, *et al.* Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digit Med* 2020;3:82.
26. Shafey LE, Soltau H, Shafran I. Joint Speech Recognition and Speaker Diarization via Sequence Transduction. arXiv [cs.CL]. 2019.http://arxiv.org/abs/1907.05337
27. Amazon Transcribe Medical. Amazon Web Services, Inc. https://aws.amazon.com/transcribe/medical/ (accessed 8 Mar 2022).
28. Select a transcription model: Cloud Text-to-Speech. Google Cloud. https://cloud.google.com/speech-to-text/docs/transcription-model (accessed 8 Mar 2022).
29. Tai-Seale M, Hatfield LA, Wilson CJ, *et al.* Periodic health examinations and missed opportunities among patients likely needing mental health care. *Am J Manag Care* 2016;22:e350–7.
30. Lafata JE, Wunderlich T, Flocke SA, *et al.* Physician use of persuasion and colorectal cancer screening. *Transl Behav Med* 2015;5:87–93.
31. Lafata JE, Shay LA, Brown R, *et al.* Office-based tools and primary care visit communication, length, and preventive service delivery. *Health Serv Res* 2016;51:728–45.
32. Shay LA, Dumenci L, Siminoff LA, *et al.* Factors associated with patient reports of positive physician relational communication. *Patient Educ Couns* 2012;89:96–101.
33. Ports KA, Barnack-Tavlaris JL, Syme ML, *et al.* Sexual health discussions with older adult patients during periodic health exams. *J Sex Med* 2014;11:901–8.

34. Shires DA, Stange KC, Divine G, *et al.* Prioritization of evidence-based preventive health services during periodic health examinations. *Am J Prev Med* 2012;42:164–73.
35. Foo PK, Frankel RM, McGuire TG, *et al.* Patient and physician race and the allocation of time and patient engagement efforts to mental health discussions in primary care: An observational study of audiorecorded periodic health examinations. *J Ambul Care Manage* 2017;40:246–56.
36. Lafata JE, Cooper G, Divine G, *et al.* Patient-physician colorectal cancer screening discussion content and patients' use of colorectal cancer screening. *Patient Educ Couns* 2014;94:76–82.
37. Flocke SA, Stange KC, Cooper GS, *et al.* Patient-rated importance and receipt of information for colorectal cancer screening. *Cancer Epidemiol Biomarkers Prev* 2011;20:2168–73.
38. Lafata JE, Cooper GS, Divine G, *et al.* Patient-physician colorectal cancer screening discussions: delivery of the 5A's in practice. *Am J Prev Med* 2011;41:480–6.
39. Wunderlich T, Cooper G, Divine G, *et al.* Inconsistencies in patient perceptions and observer ratings of shared decision making: the case of colorectal cancer screening. *Patient Educ Couns* 2010;80:358–63.
40. Park J, Kotzias D, Kuo P, *et al.* Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. *J Am Med Inform Assoc* 2019;26:1493–504.
41. Park J, Jindal A, Kuo P, *et al.* Automated rating of patient and physician emotion in primary care visits. *Patient Educ Couns* 2021;104:2098–105.
42. Wetterneck TB, Lapin JA, Krueger DJ, *et al.* Development of a primary care physician task list to evaluate clinic visit workflow. *BMJ Qual Saf* 2012;21:47–53.
43. Musen MA, Noy NF, Shah NH, *et al.* The National Center for Biomedical Ontology. *J Am Med Inform Assoc* 2012;19:190–5.
44. Xiong W, Droppo J, Huang X, *et al.* Achieving human parity in conversational speech recognition. arXiv [cs.CL]. 2016.http://arxiv.org/abs/1610.05256