Performance of ASR Engines in Capturing Clinical Conversations across different Recording Modalities.

Ramya Sai Swathi Mangu
Donald Bren School of ICS
December 2022

Faculty Advisor: Dr. Kai Zheng
Department of Informatics

Abstract

Poorly transcribed clinical conversations may have serious implications such as incorrect patient information. So, understanding the performance of the current Transcription engines is a vital part for understanding the problem and creating specific goals for improvement. Evaluating the performance of Automatic Speech Recognition (ASR) transcribed clinical conversations across different modalities is also an essential part of analyzing the quality of the engines. This evaluation gives an insight into the capturing capability of medical conversations by each ASR engine through each recording modality through which we will be able to see the feasibility of engines and what happens when limited by modalities in clinics. The ASR engines that have been used for this project are the 4 commercially available engines from established vendors: Amazon general, Amazon medical, Google general and Google medical. The purpose of this research is to evaluate these engine performances across different modalities - Lavalier, Recorder, Cell, TeraMac, Desktop and TeraPhone using metrics such as semantic similarity for evaluating the semantic content of the transcribed data, and BioAnnotator for assessing potentially meaningful information present in patient-clinician conversations. Measuring the semantic similarity includes calculating the Word Mover's Distance (WMD) and Cosine Similarity between the reference and transcribed scripts. We identified the cosine similarity with ~77-89% which is similar across the modalities. The results of the analysis indicate that the performance of the engines across different modalities is roughly similar and that modalities do not have noticeable effects on the performance of the engines. These performance figures of the simulated data are better than the prior analyses with real-word data.

Table of Contents

Introduction

As clinics have adapted to Electronic Health Records (EHRs) the inadequacies of paper medical charts became apparent [1]. EHRs helped clinicians to improve patient care and clinic management [2]. Though they reshaped modern healthcare, EHRs have posed many problems for clinicians. As the documentation of the data and additional requirements on the clinical documentation process for purposes other than direct care of the patient increased, clinicians have found themselves spending more time charting than on clinical care [3]. Added to this are that clinicians and other healthcare professionals' concerns that the quality of the system is inadequate [3].

One proposed solution for this is to use so-called “digital scribes”. This concept technology aims to reduce manual documentation by clinicians or medical scribes [4]. This technology uses voice-to-text software, Artificial Intelligence (AI), and Natural Language Processing (NLP) to transform clinical interactions into meaningful medical notes through ambient listening and subsequent voice-to-text conversion [4 - 5]. If successful, these tools help in shifting the existing documentation burden away from clinicians, and allowing them to better focus on patient care-related activities.

Despite its promising solution and the potential to revolutionize medical charting, many barriers exist to its implementation and are still in development [5]. Some of the key challenges noted are the noises and other environmental conditions in clinics that negatively affect ASR, structuring clinician-patient conversations, and generating clinically meaningful documentation [4]. When applied to conversational clinical speech, the performance of ASR engines was low with Word Error Rates (WERs) of approximately 50% and concept extraction rates of approximately 60% [6].

The main challenge faced in digital scribing is to record the audio of a clinician-patient conversation. Though using high quality audio recorders minimizes errors across the processing pipeline , the position of patient and doctor from the microphone while recording affects the clarity and the volume of the recorded audio as clinician and patient rarely face the microphone during the examination or the patient is sometimes required to shift positions in the exam rooms[4 – 6]. A method that could be considered is to use different audio recorders to know the feasibility of the ASR engines. In 2018, Kodish-Wachs evaluated eight ASR engines using WER and the precision, recall and F1 scores across two different microphones. With ASR performance figures of 35% to 65% WER, they concluded that more focus is needed in improving ASR engine performance before adopting digital scribes for conversational speech in clinics. However, many advancements to medical conversation ASR and speaker recognition have been published. An updated understanding of the performance of these tools across different modalities, as well as the knowledge of whether the recording modalities have adverse effects on ASR performance, may help guide clinician in knowing the feasibility of the engines when limited by modalities and help in thinking about how these technologies could be deployed to better support clinician users.

This study is to report an independent evaluation of the engine's performance across different recording modalities. We obtain a performance snapshot with a corpus of patient-clinician conversations with the engine through each modality and then identify potential development opportunities relevant to the capture, processing of patient-clinician conversations and effects of recording modalities in transcribing the clinical conversations.

Learning Objective

To learn about the current performance of the automatic speech recognition systems across different modalities aimed at capturing the patient-clinician conversations and the potential impact on the development of automated clinical documentation tools.

Methodology

Data for Evaluation

The evaluation was conducted on a total of 6 modalities with each modality having 35 manually transcribed audio recordings of conversations between patients and primary care clinicians. The total number of files for each engine are 210. The patients that were involved in this transcribing were aged between 50 to 80. They were visiting the primary care physician, working in a 26-clinic ambulatory healthcare system in the Midwest United States. This data has been used in analyzing the delivery of preventative services and also to detect emotion and topic in patient-clinical conversation. This data has been created as a part of the Mental Health Discussion Study by Tai-Seale et al [8].

To avoid the external facts such as poor microphone quality, background noise, etc, the audio recordings were re-enacted in a quiet interview studio by two native English speakers. These readers were seated within 1 foot of all the modalities. Readers read the de-identified transcripts into all the modalities. All of the recordings ranged from 12 to 55 minutes in length.

In order to assess the performance of the engines on conversations which may be relevant for clinical documentation , a medical student, repeatedly scanned the transcripts for relevant utterances such as notes on lab values, question answer pairs, information for care coordination

and physical examination findings and highlighted these concepts which were used to segment data for evaluation.

ASR Engine selection

Four ASR engines measured by WER with a subsample of dataset in prior work were selected based on their accessibility and performance. Google Speech-to-Text "medical_conversation" (Google, Mountain View, USA) and Amazon Transcribe Medical "Primary care" & "Conversation" (Amazon, Seattle, USA) were selected as these models were tailored to medical conversations. Hereafter, these engines will be referred to as Google Speech-to-Text Medical Conversation and Amazon Transcribe Medical. I also evaluated the Google Speech-to-Text "Video" and Amazon Transcribe "General" models which are suitable for general audio and multiple speakers to contrast medical models with general-purpose models. Hereafter, these engines will be referred to as Google Speech-to-Text Video and Amazon Transcribe General. The "Default" transcription model was excluded in the Google Speech-to-Text Video for this evaluation due to lower performance compared to the "General" transcription model of Google Speech-to-Text Video.

ASR data generation

The audio recordings were uploaded to cloud storage and transcribed in October 2022. The original transcripts consisted of Protected Health Information (PHI), so as a measure to protect the privacy, PHI was replaced with generic tokens. To remove the local context, while reading the transcripts, readers replaced all identifiers via the "Safe Harbor method" as outlined by HIPAA[7] and manual review to remove local context. In order for the data to be readable, all the irrelevant data was removed and all the text in the transcripts were labeled with the respective speaker and the talk turn. Non -verbal tokens which are the words produced by non-verbal

sounds such as coughing, laughing, yawning, etc, were identified and removed from the text. The text was converted to proper casing and punctuation. All these transcription requests were completed in approximately four hours for each engine.

General Performance Evaluation

As a first step of evaluation, the performance of the engines was evaluated based on the WMD. This metric is useful to get the semantic and syntactic similarity between the two documents. To calculate the WMD between the transcribed text and the reference text, pre-trained word2vec / glove embedding: GoogleNews-vectors-negative300.bin.gz was used. These embeddings were loaded into the Gensim Word2Vec model class and the wmdistance method was used to calculate the semantic similarity between the transcribed text and the reference text. Based on this, the mean WMD of all ASR engines across the modalities was reported.

As a second step, the performance of the engines was evaluated based on their cosine similarity. For this, SentenceTransformers was considered which is widely used for state-of-the-art sentence, text and image embeddings. Hugging Face's SentenceTransformer model to compute embeddings of the text [9]. Using the encode function of the model, the sentence embeddings of the transcribed and reference text was computed. Then, using the cosine similarity function from the util library the cosine similarity between the transcribed text and the reference text was calculated. Based on this, the mean cosine similarity of all ASR engines across the modalities was reported.

Domain Specific Performance Evaluation

As a third step of evaluation, the transcription performance of potentially meaningful information present in patient-clinician conversations was assessed. The capture of medical

concepts was compared as labeled by an automated annotator43 using Logical Observation Identifiers Names and Codes (LOINC) and SNOMED CT ontologies [10]. Based on this, results as precision, recall, and F1 scores relative to reference transcripts were reported.

Results

To get the amount of dissimilarity between the transcribed and the reference texts, that is to calculate the distance between the two texts in a meaningful way, WMD was used. From this evaluation, all evaluated models achieved lower median value of WMD, close to zero (**Figure 1**). In this evaluation, Google Speech-to-Text Medical Conversation has higher WMD average and Google Speech-to-Text Video achieved lower WMD average (**Table 1**). From this we can derive that the dissimilarity between the referenced and the transcribed texts is low and is almost similar across the engines.

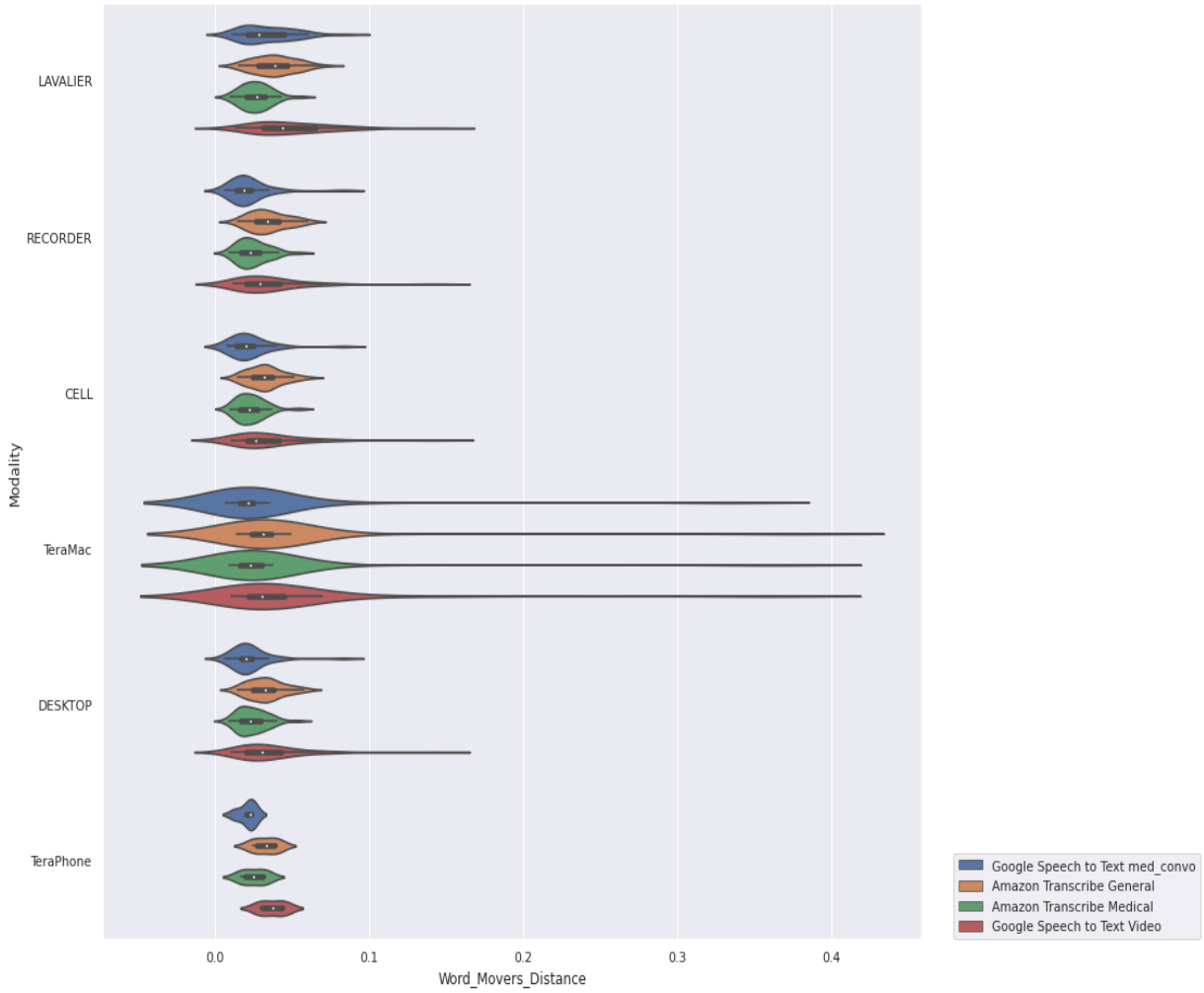

Figure 1: Performance of ASR engines across modalities measured by Word Mover's Distance.

Engine	Modality	Average Word Mover's Distance
Amazon Transcribe General	Cell	0.032
	Desktop	0.033
	Lavalier	0.038
	Recorder	0.034
	TeraMac	0.042
	TeraPhone	0.032
Amazon Transcribe Medical	Cell	0.023
	Desktop	0.024
	Lavalier	0.026
	Recorder	0.024
	TeraMac	0.033
	TeraPhone	0.025
Google Speech-to-Text Medical Conversation	Cell	0.035
	Desktop	0.036
	Lavalier	0.050
	Recorder	0.035
	TeraMac	0.045
	TeraPhone	0.037
Google Speech-to-Text Video	Cell	0.022
	Desktop	0.023
	Lavalier	0.034
	Recorder	0.022
	TeraMac	0.031

	TeraPhone	0.020

Table 1. Performance of four ASR engines measured by the average of Word Mover's Distance.

To get the similarity between the transcribed and the reference texts, cosine similarity was calculated. This is measured by the cosine angle between the two texts. From this evaluation, all evaluated models achieved a high median (**Figure 2**). In this, the cosine similarity varied across different models, with Amazon Transcribe Medical engine having the highest average of Cosine Similarity and Amazon Transcribe general having the lowest (**Table 2**).

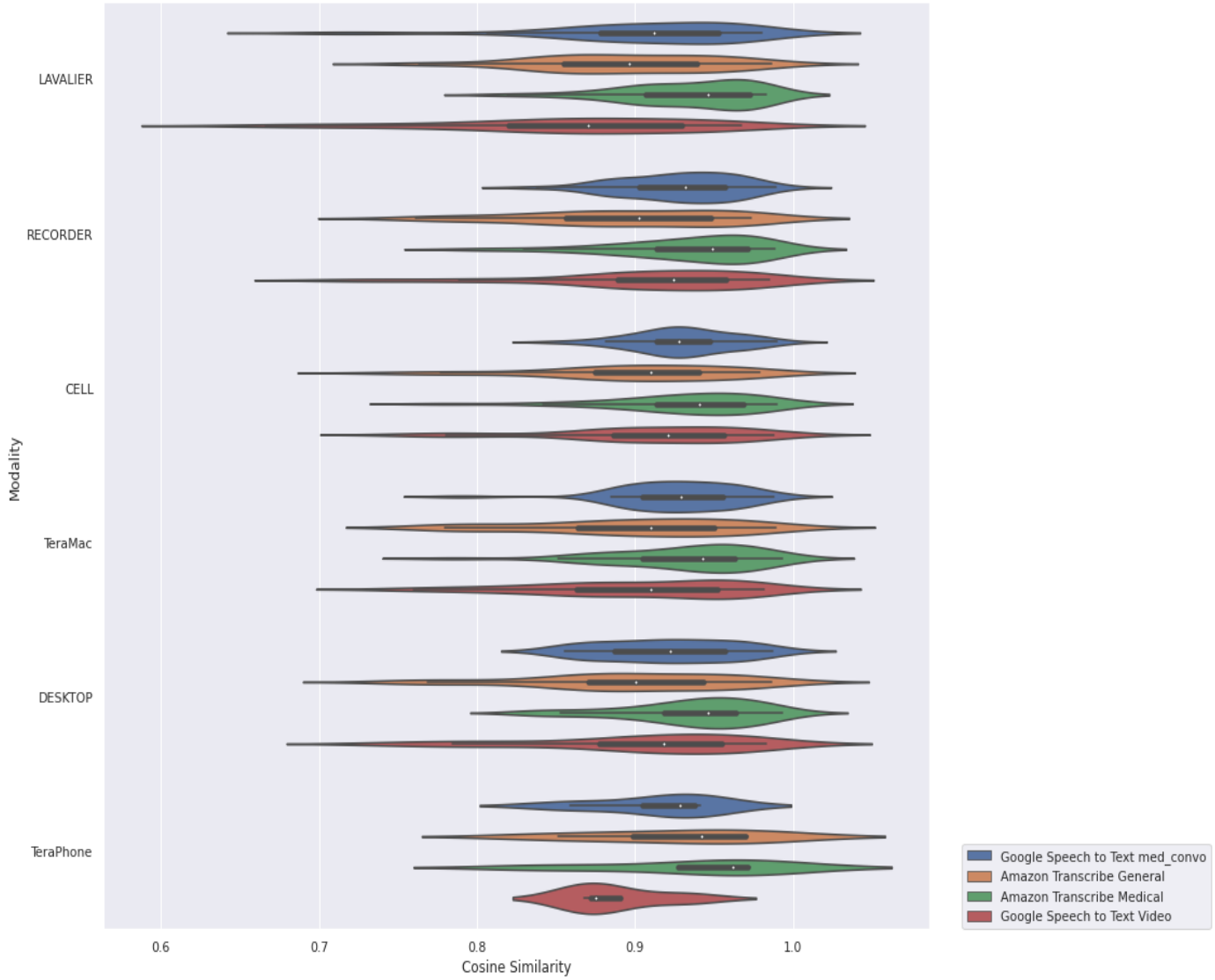

Figure 2: Performance of ASR engines across modalities measured by Cosine Similarity.

Engine	Modality	Average Cosine Similarity
Amazon Transcribe General	Cell	0.90
	Desktop	0.90
	Lavalier	0.89
	Recorder	0.89
	TeraMac	0.90
	TeraPhone	0.92
Amazon Transcribe Medical	Cell	0.93
	Desktop	0.94
	Lavalier	0.93
	Recorder	0.93
	TeraMac	0.93
	TeraPhone	0.94
Google Speech-to-Text Medical Conversation	Cell	0.91
	Desktop	0.90
	Lavalier	0.86
	Recorder	0.91
	TeraMac	0.90
	TeraPhone	0.89
Google Speech-to-Text Video	Cell	0.93
	Desktop	0.92
	Lavalier	0.91
	Recorder	0.93
	TeraMac	0.93

	TeraPhone	0.91

Table 2. Performance of four ASR engines measured by the average of Cosine Similarity.

For the domain-specific evaluation, the transcription fidelity in terms of medically pertinent concepts as labeled by an automatic annotator was evaluated. From this evaluation, all evaluated models achieved low recall and high precision **(Table 3)**. From this we can derive that many of the medically pertinent concepts were not correctly transcribed by ASR, but when they were captured, the concepts were captured correctly.

Engine	Modality	Recall	Precision	F1 Score
Amazon Transcribe General	Cell	0.49	0.96	0.65
	Desktop	0.49	0.96	0.65
	Lavalier	0.49	0.97	0.65
	Recorder	0.49	0.96	0.65
	TeraMac	0.49	0.96	0.65
	TeraPhone	0.49	0.97	0.65
Amazon Transcribe Medical	Cell	0.49	0.97	0.65
	Desktop	0.49	0.97	0.65
	Lavalier	0.49	0.97	0.65
	Recorder	0.49	0.97	0.65
	TeraMac	0.49	0.97	0.65
	TeraPhone	0.49	0.97	0.65
Google Speech-to-Text Medical Conversation	Cell	0.49	0.96	0.65
	Desktop	0.49	0.96	0.65
	Lavalier	0.49	0.96	0.65
	Recorder	0.49	0.96	0.65

	TeraMac	0.49	0.96	0.65
	TeraPhone	0.49	0.96	0.65
Google Speech-to-Text Video	Cell	0.49	0.95	0.65
	Desktop	0.49	0.95	0.65
	Lavalier	0.49	0.95	0.65
	Recorder	0.49	0.95	0.65
	TeraMac	0.49	0.95	0.64
	TeraPhone	0.49	0.95	0.65

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

Discussion

In this study, four general ASR models were compared across six different modalities on the simulated clinical conversations corpus. The semantic similarity of the texts evaluated based on the WMD and cosine similarity is similar across the modalities. The dissimilarity between the texts closer to zero and the cosine similarity closer to 1 shows that the transcribed texts and reference texts are meaningfully similar with variance in the words. From the results, variance from ASR engines have more effect on the performance than the type of the microphone used. This suggests that, under ideal conditions where microphones are evenly separated from the patient and physician, existing ASR engines have enough fidelity across modalities even with cell phone or laptop microphones. If the evaluations are valid, then clinics do not need an expensive or specialized microphone to transcribe the conversations.

The reports generated suggest that the results are similar to the prior work. Based on the Adam Miner study [11], the average semantic similarity achieved between the referenced and the transcribed text is 1.2 which is close to average cosine similarity 0.91 generated in this study.

The results may not be the same under non ideal conditions where the patient or the physician is moving. The study data was generated in sound-studio like conditions, which likely yield better audio quality for transcription than in real-world clinic settings. Therefore this study's results likely represent an upper-bound in performance and can clearly showcase differences between modalities, if present. In practice, performance with the tested modalities may fluctuate more as patients and physicians move closer and further from single microphones (e.g., when the patient moves from a chair to the examination table).

Though certain engines are tuned for the capture of medical conversations, the medical conversation models achieved similar performance to general-purpose models in terms of capturing relevant phrases for documentation and capturing medical concepts. Apart from this, the recall, precision and f1 score were similar across the modalities and did not show any drastic variance in the results. We speculate that LOINC and SNOMED, which was used to guide the automatic annotator, contained concepts which were common in general conversation, such as "I don't know." These concepts appeared frequently in the original patient-physician dialogues, and were therefore similarly annotated across outputs from the specialized and general-purpose ASR engines.

Some of the limitations of this study is that the data is simulated and is a small sample representing 5 providers in primary health care. The content and the length of the conversations from different health providers or different health may drastically differ, which may yield different results. There can be engines that are specifically tuned to certain types of microphones,

which may mitigate low performance based on recording modality. Finally, the evaluation is only based on the ASR performance and the endpoint of evaluating digital scribe output is not directly tested. Once complete digital scribing solutions or NLP capable of transforming transcripts into preliminary notes are available, future work could be conducted to confirm how recording modality may affect the quality of output clinical notes.

Conclusion

In this study, we assessed the feasibility of ASR engines for digital scribe systems for patient -clinical conversations by evaluating the performance of the specialized models. Based on this, we conclude that the choice of engines play a significant role in the performance than the choice of modalities. From the results, variance from ASR engines have more effect on the performance than the type of the microphone used. This suggests that, under ideal conditions where microphones are evenly separated from the patient and physician, existing ASR engines have enough fidelity across modalities even with cell phone or laptop microphones. If the evaluations are valid, then clinics do not need an expensive or specialized microphone to transcribe the conversations.

References

1. Ornstein, Steven, et al. “I He Computer-Based Medical Record: Current Status.” *Family Practice*, vol. 35, no. 5, 1992, cdn.mdedge.com/files/s3fs-public/jfp-archived-issues/1992-volume_35/November%201992/JFP_1992-11_v35_i5_the-computer-based-medical-record-curren.pdf. Accessed 27 Dec. 2022.
2. Nguyen, Lemai, et al. “Electronic Health Records Implementation: An Evaluation of Information System Impact and Contingency Factors.” *International Journal of Medical Informatics*, vol. 83, no. 11, Nov. 2014, pp. 779–796, 10.1016/j.ijmedinf.2014.06.011. Accessed 13 Mar. 2019.
3. Kuhn, Thomson, et al. “Clinical Documentation in the 21st Century: Executive Summary of a Policy Position Paper from the American College of Physicians.” *Annals of Internal Medicine*, vol. 162, no. 4, 17 Feb. 2015, p. 301, 10.7326/m14-2128.
4. Quiroz, Juan C., et al. “Challenges of Developing a Digital Scribe to Reduce Clinical Documentation Burden.” *Npj Digital Medicine*, vol. 2, no. 1, 22 Nov. 2019, pp. 1–6, www.nature.com/articles/s41746-019-0190-1, 10.1038/s41746-019-0190-1. Accessed 22 Oct. 2020.
5. Ghatnekar, Shilpa, et al. “Digital Scribe Utility and Barriers to Implementation in Clinical Practice: A Scoping Review.” *Health and Technology*, 2 June 2021, 10.1007/s12553-021-00568-0.
6. Kodish-Wachs, Jodi, et al. “A Systematic Comparison of Contemporary Automatic Speech Recognition Engines for Conversational Clinical Speech.” *AMIA Annual*

Symposium Proceedings, vol. 2018, 5 Dec. 2018, pp. 683–689, www.ncbi.nlm.nih.gov/pmc/articles/PMC6371385/.

7. Rights (OCR), Office for Civil. "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." *HHS.gov*, 7 Sept. 2012, www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharborguidance.
8. Tai-Seale, Ming, et al. "Evidence-Based Mental Health Discussions during Periodic Health Exams: The Cup Is 1/3 Full." *Journal of Patient-Centered Research and Reviews*, vol. 3, no. 3, 15 Aug. 2016, p. 188, 10.17294/2330-0698.1320. Accessed 17 June 2020.
9. Reimers, Nils. "Sentence-Transformers (Sentence Transformers)." *Huggingface.co*, huggingface.co/sentence-transformers.
10. "About Us | BioPortal." *BioPortal*, www.bioontology.org/about-us/#cite-bioportal.
11. Miner, Adam S., et al. "Assessing the Accuracy of Automatic Speech Recognition for Psychotherapy." Npj Digital Medicine, vol. 3, no. 1, 3 June 2020, 10.1038/s41746-020-0285-8.