Journal of the American Medical Informatics Association, 30(4), 2023, 703–711
https://doi.org/10.1093/jamia/ocad001
Advance Access Publication Date: 23 January 2023
Research and Applications

Research and Applications

"Mm-hm," "Uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology?

Brian D. Tran [1,2], Kareem Latif [3], Tera L. Reynolds[4], Jihyun Park[5], Jennifer Elston Lafata[6,7], Ming Tai-Seale[8], and Kai Zheng[1]

[1]Department of Informatics, Donald Bren School of Informatics and Computer Science, University of California, Irvine, Irvine, California, USA, [2]School of Medicine, University of California, Irvine, Irvine, California, USA, [3]School of Medicine, California University of Science and Medicine, Colton, California, USA, [4]Department of Information Systems, University of Maryland, Baltimore County, Baltimore, Maryland, USA, [5]Department of Computer Science, Donald Bren School of Informatics and Computer Science, University of California, Irvine, Irvine, California, USA, [6]Division of Pharmaceutical Outcomes and Policy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [7]Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, Michigan, USA and [8]Department of Family Medicine and Public Health, School of Medicine, University of California, San Diego, La Jolla, California, USA

Corresponding Author: Kai Zheng, PhD, Department of Informatics, Donald Bren School of Informatics and Computer Science, University of California, Irvine, 6095 Donald Bren Hall, Irvine, CA 92697, USA; kai.zheng@uci.edu

Received 29 August 2022; Revised 13 December 2022; Editorial Decision 2 January 2023; Accepted 12 January 2023

ABSTRACT

Objectives: Ambient clinical documentation technology uses automatic speech recognition (ASR) and natural language processing (NLP) to turn patient–clinician conversations into clinical documentation. It is a promising approach to reducing clinician burden and improving documentation quality. However, the performance of current-generation ASR remains inadequately validated. In this study, we investigated the impact of non-lexical conversational sounds (NLCS) on ASR performance. NLCS, such as *Mm-hm* and *Uh-uh*, are commonly used to convey important information in clinical conversations, for example, *Mm-hm* as a "yes" response from the patient to the clinician question "are you allergic to antibiotics?"

Materials and Methods: In this study, we evaluated 2 contemporary ASR engines, Google Speech-to-Text Clinical Conversation ("Google ASR"), and Amazon Transcribe Medical ("Amazon ASR"), both of which have their language models specifically tailored to clinical conversations. The empirical data used were from 36 primary care encounters. We conducted a series of quantitative and qualitative analyses to examine the word error rate (WER) and the potential impact of misrecognized NLCS on the quality of clinical documentation.

Results: Out of a total of 135 647 spoken words contained in the evaluation data, 3284 (2.4%) were NLCS. Among these NLCS, 76 (0.06% of total words, 2.3% of all NLCS) were used to convey clinically relevant information. The overall WER, of all spoken words, was 11.8% for Google ASR and 12.8% for Amazon ASR. However, both ASR engines demonstrated poor performance in recognizing NLCS: the WERs across frequently used NLCS were 40.8% (Google) and 57.2% (Amazon), respectively; and among the NLCS that conveyed clinically relevant information, 94.7% and 98.7%, respectively.

Discussion and Conclusion: Current ASR solutions are not capable of properly recognizing NLCS, particularly those that convey clinically relevant information. Although the volume of NLCS in our evaluation data was very small (2.4% of the total corpus; and for NLCS that conveyed clinically relevant information: 0.06%),

incorrect recognition of them could result in inaccuracies in clinical documentation and introduce new patient safety risks.

Key words: medical scribe, digital scribe, information technology [L01.479], electronic health records [E05.318.308.940.968.625.500], speech recognition software [L01.224.900.889], workflow [L01.906.893], documentation [L01.453.245]

INTRODUCTION

Clinician burnout has been recently declared as a public health crisis.[1] Clinical documentation, defined as "the capturing and recording of clinical information, often in real time whilst the patient is present—for example, during consultation, assessment, imaging, and treatment,"[2] is a tedious process known to be a major contributor to clinician burnout.[3] The widespread adoption of electronic health records (EHR) in the United States has exacerbated the situation. Recent work has found that clinicians could spend over half of their workday performing desktop medicine tasks using the HER,[4,5] often at the expense of time that could be otherwise spent on direct patient care. This issue has prompted a growing interest in developing technological solutions to automate, at least in part, the clinical documentation process.

One promising approach is the use of ambient audio-recording devices installed in the exam room, along with an automatic speech recognition (ASR) system, to capture and transcribe the patient–clinician conversation followed by using natural language processing (NLP) to generate clinical documentation based on the resultant transcript.[6–8] This ambient clinical documentation technology, also referred to as "digital scribes," provides great potential to alleviate the documentation burden in addition to producing potentially higher quality clinical data.[9]

In the past decade, substantial investments have been made by both the industry (eg, Google,[10] Microsoft/Nuance[11,12]), 3M M*Modal Fluency Align,[13] as well as the academic research community[6,14,15] in developing the ambient clinical documentation technology. For example, Microsoft launched Project EmpowerMD in 2018 that aimed to create artificial intelligence-based tools to automate clinical documentation; and the company recently acquired Nuance Communications, which developed the commonly used medical speech recognition software Dragon Medical One, to further this ambition.[16,17] Some of these efforts have now led to commercially available digital scribe solutions. For example, at the time of writing, both Google[18] and Amazon,[19] among others, have released the commercially available version of their speech recognition engines specifically trained for clinical conversation.

While significant progress has been made, the complexity of generating clinical documentation based on patient–clinician conversation remains inadequately explored.[6,9] Despite the fact that there has been abundant research on the interactions between patients and clinicians in the exam room through the lenses of sociolinguistic perspectives (eg, conversation analysis, discourse analysis),[20–24] few studies have taken such dynamics into account in building digital scribe systems.[25,26] One particularly salient issue originates from expressions such as *Mm-hm*, *Mm*, and *Uh-huh*, which are commonly used by both patients and clinicians to communicate acknowledgement, positive or negative affirmations, or questioning.[27,28] For example, when a clinician asks, "are you allergic to penicillin," a patient may respond with *Hm* as a conversation filler, *Uh-huh* or *Mm-hm* as yes, or *Uh-uh* as no. Such expressions are generally called continuers,[29] turn-initial particles,[30] backchannels,[31] or non-lexical conversational sounds (NLCS),[32,33] depending on the context. For simplicity, we refer to all of them as NLCS in this article.

While frequently used in clinical conversation, NLCS can be incorrectly and/or inconsistently recognized by ASR engines[34,35] because they are often acoustically similar (eg, *Uh-huh* vs *Uh-uh;* see pronunciation examples at Supplementary File—Video), and their interpretation may be modulated by non-lexical features of speech, for example, changes in the intonation and rhythm may alter the meaning of *Mm* from answering positively to a question, or simply to note to the other party of the conversation that they are listening.[32,36,37] For the ambient clinical documentation technology to work, the recognition of such sounds and proper interpretation of their meaning are important to generate high-quality transcripts for use in later processes (eg, using NLP) to produce preliminary clinical documentation. Failure to do so could result in incomplete clinical documentation at the minimum, and in certain cases could lead to severe patient safety consequences, for example, when a "yes" response to an allergy-related question is not properly captured due to the use of NLCS.

In this article, we report an empirical study that assessed how commercially available ASR engines perform in recognizing NLCS sounds from clinical conversation. The objectives were 4-fold: (1) to quantify the prevalence of NCLS from a sample set of ambulatory primary care physician office visits, (2) to classify the semantic type and the meaning of NLCS into distinct categories (eg, serving as a conversational filler vs providing explicit answer to a clinical probe, positive affirmation vs negative affirmation), (3) to evaluate if contemporary ASR engines can accurately recognize and differentiate NLCS sounds, and (4) to analyze recognition errors identified in the performance evaluation.

MATERIALS AND METHODS

ASR engines evaluated

As this study focuses on the ambient clinical documentation technology, we selected ASR engines specifically designed to transcribe clinical conversations that involved at least 2 parties and generally took place in an exam room, as opposed to ASR systems intended for use as a one-on-one dictation tool. To the best of our knowledge, Google Cloud Speech-to-Text (Google, Mountain View, USA) with the "medical_conversations" model, and Amazon Transcribe Medical (Amazon, Seattle, USA) with the combined "primarycare" and "conversation" model, are the only 2 such ASR engines that are currently commercially available. Although we were aware of other efforts in developing ASR engines for clinical conversation, notably Dragon Medical SpeechKit SDK (Nuance, Burlington, USA) and DeepScribe.ai (DeepScribe, San Francisco, USA), despite repeated requests, we were unable to obtain their software for this evaluation. Therefore, this study was based on the language models of Google Cloud Speech-to-Text and Amazon Transcribe Medical tailored for clinical conversation; hereafter referred to as "Google ASR" and "Amazon ASR," respectively.

Evaluation data

The evaluation data were based on anonymized transcripts of 36 patients' in-person primary care encounters with 5 primary care providers. The original dataset was collected between 2007 to 2009 in southeast Michigan and were transcribed by professional transcriptionists, for an NIH-funded project that aimed to study patient–clinician interactions in the exam room.[38–40] All patients were between 50 and 80 years of age. This dataset has been extensively used in prior research on topics such as detection of discussion topics, clinician adherence to best practice guidelines, and analysis of shared decision making.[38–51]

To eliminate the potential undesired effects of recording-related factors (eg, varied volume and background noise levels) and speaker-related factors (eg, non-native English speaker vs native, speakers with strong accent), and to achieve the highest quality of audio recording possible, we re-enacted these encounters based on the transcripts in a sound studio using a professional-grade microphone (Blue Yeti, Logitech International S.A., Lausanne, Switzerland). Two native-English speaking graduate student research assistants read aloud off the transcripts. Neither of them had any prior knowledge that the recognition accuracy of NLCS was to be evaluated. The microphone was placed within 3 feet between them. Additional information on our recording setup and the re-enacting procedures is provided in Supplementary Data File S1. The Institutional Review Board reviewed the protocol of this study and deemed it to be nonhuman subjects' research.

Qualitative analysis for determining the semantic types of NLCS

To study how NLCS may influence ASR performance, we first conducted a qualitative analysis to delineate the conversational functions of the NLCS utterances found in the evaluation data. This analysis was informed by prior sociolinguistic work, for example, Ward's inventory of NLCS that provides a list of commonly used NLCS in American English.[32] The results are henceforth referred to as "semantic types," defined in detail in Table 1.

A video file illustrating the differences between these semantic types can be found at Supplementary File—Video. For example, *Mm-hm* could be used to denote the receipt of information or to note that the person was listening ("acknowledgement or backchannel"), as part of a disfluency ("filler words or speech disfluency"), or to raise a question ("question"). It could also be used to express agreement or disagreement ("positive or negative affirmation in response to declarative questions or statements," hereafter referred to as "affirmation-declarative"), in lieu of "yes" or "no" to answer a probe ("positive or negative affirmation in response to non-declarative questions," hereafter referred to as "affirmation-non-declarative"). Among these semantic types, "acknowledgement or backchannel" and "filler words or speech disfluency" generally do not convey any real meaning but are instead used to facilitate the flow of the conversation.

Qualitative analysis for determining the clinical relevance of NLCS

In addition to distinguishing between NLCS semantic types, we also analyzed the clinical relevance of NLCS. For each of the NLCS utterances found in the evaluation data, we examined each utterance to determine if it were incorrectly captured or were omitted from clinical documentation, whether it might result in loss or change of clinically relevant information; for example, omitting *Uh-huh* that represents a "yes" response to the question "are you allergic to aspirin?"

In this analysis, the clinical relevance determination was guided by a primary care physician task list developed by Wetterneck et al[54] and a generic annual history checklist for primary care.[55] The former is a comprehensive taxonomy of common tasks that primary care physicians perform. The latter provides a list of information that physicians would solicit from the patient during a typical primary care encounter, such as medication history, preventative screening information, and social history. An NLCS utterance was marked as conveying clinically relevant information if it pertained to a specified primary care task or annual history checklist item. In this study, we did not assess the clinical significance of an NLCS, or the likelihood or magnitude/severity of incorrect recognition of it on affecting clinical decision making or patient safety.

Coding process for classifying NLCS semantic type and clinical relevance

We used a 2-step process to code for NLCS utterances by semantic type and clinical relevance. First, one of the authors (BDT) conducted the initial screening of all NLCS utterances identified in the data ($N = 3284$) to classify them as (1) containing potentially clinically relevant information (eg, positive or negative affirmation in response to a nondeclarative question) or (2) being used to facilitate conversation (eg, filler word or speech disfluency). Then, 2 coders (BDT and KL) independently coded the utterances by semantic type and clinical relevance among the NLCS found to potentially contain clinically relevant information ($N = 767$). In this second step, the inter-rater agreement (Cohen's Kappa) achieved between the 2 coders: 0.80 when determining the semantic type and 0.83 when determining if an NLCS conveyed clinically relevant information. All disagreements were resolved in consensus development meetings.

ASR performance evaluation

The audio files were uploaded to Google ASR and Amazon ASR and processed using the following parameters: (1) language: English, (2) model: "medical_conversations" for Google ASR and "primarycare" + "conversation" for Amazon, (3) automatic punctuation: yes, and (4) speaker diarization: 2 speakers. Each hour of audio recording took approximately 15 min for either of the ASR engines to process. The cost for processing a total of 975 min of recordings was $33 for Google ASR and $73 for Amazon ASR, at the time when this study was conducted (November 2021).

RESULTS

Descriptive statistics

The anonymized transcripts of the 36 primary care encounters each contained 1527 to 13 203 words (mean: 3692, SD: 1994); 26.3% to 67.7% (mean: 43.4%, SD: 9.7%) of them were spoken by patients. The duration of the reenacted audio recordings was between 12.6 and 55.3 min (mean: 27.1, SD: 13.8). In total, the transcripts contained 135 647 spoken words. Among them, there were 3284 NLCS utterances, ranging 21 to 245 instances (mean: 91, SD: 55) per encounter. The most common NLCS were *Mm-hm*, *Oh*, *Um*, *Uh*, *Ah*, *Uh-huh*, *Huh*, *Mm*, *Hm*, *Eh*, and *Uh-uh*. Some other NCLS, such as *Aw, Hum, Oops, Hooray, Geez, Uh-oh,* and *Woo-hoo*, also appeared in the data but infrequently. They were not analyzed in this article and are instead listed in Supplemental Data File S1.

Table 1. NLCS semantic types

Semantic type	Definition	Example
Acknowledgement or backchannel	A short verbal response to acknowledge the receipt of information, or to maintain the continuity of a conversation.[52,53]	Patient: *We talked about getting my eyes checked last month, and—* Doctor: *Mm-hm. And did they check your eye pressures at your last visit?*
Filler word or speech disfluency	Stuttering, repetitions, and revisions as part of a conversational dialog.	Doctor: *How's your mother?* Patient: *Just turned 90. She doesn't really have any major health problems, Uh-huh.*
Positive or negative affirmation in response to a declarative question or statement	A short verbal response to indicate agreement or disagreement to a declarative question or statement.[32]	*Example of agreement:* Doctor: *Looking at my records here, you had a complete hysterectomy.* Patient: *Mm-hm.* *Example of disagreement:* Doctor: *The chart shows that you had your ultrasound on the 22nd.* Patient: *Uh-uh.*
Positive or negative affirmation in response to a nondeclarative question	A short verbal response to provide "yes" or "no" answer to a nondeclarative question.[32]	*Example of a "yes" response:* Doctor: *Nicotine gum for your smoking. Do you want to try it?* Patient: *Mm-hm.* *Example of a "no" response:* Doctor: *Do you have time to see a therapist today?* Patient: *Uh-uh.*
Question	A short verbal response to specify questioning by the speaker.[32]	Doctor: *Still need a refill on that medication, huh?* Patient: *Yes.*
Unsure	The conversational function of the NLCS cannot be readily ascertained from the context.	Patient: *Do you see that, right there?* Doctor: *Hmm. Uh-huh. Okay.*

NLCS: non-lexical conversational sounds.

NLCS semantic types

Table 2 reports how often the frequently uttered NLCS (ie, *Mm-hm*, *Oh*, *Um*, *Uh*, *Ah*, *Uh-huh*, *Huh*, *Mm*, *Hm*, *Eh*, and *Uh-uh*) appeared under different semantic types.

As shown in Table 2, a majority of these NLCS were used for filler word and disfluencies ($N = 2047$), followed by acknowledgement or backchannel ($N = 887$), affirmation-declarative ($N = 111$), and affirmation-nondeclarative ($N = 43$). The NLCS that often conveyed clinically relevant information included *Mm* ($N = 7$), *Mm-hm* ($N = 57$), *Uh-huh* ($N = 6$), and *Uh-uh* ($N = 6$). All of them fell into the affirmation-declarative semantic type or the affirmation-nondeclarative type. For example, the doctor stated, "Do you want to see a therapist next door?," and the patient uttered *Mm-hm* as a "yes" response (affirmative-declarative); or the doctor asked "Now, um, apart from the hysterectomy, uh, any other surgery you had?," and the patient responded negatively using *Uh-uh* (affirmative-nondeclarative). It is noteworthy that almost all of the NCLS that conveyed clinically relevant information were uttered by the patient speaker (75 of 76).

Table 2. NLCS counts and their usage in conveying clinically relevant information

Semantic type	Total number of instances	Number of instances that conveyed clinically relevant information
Acknowledgement or backchannel	887	0
Affirmation-declarative	111	44
Affirmation-nondeclarative	44	32
Filler word and disfluencies	2047	0
Question	68	0
Unsure	5	0

NLCS: non-lexical conversational sounds.

ASR performance with NLCS

The performance of the 2 ASR engines in correctly recognizing NLCS is reported in Table 3. The upper portion shows the results for the NLCS instances that conveyed clinically relevant information. The error rate was 94.7% for Google ASR and 98.7% for Amazon ASR. Of the NLCS conveying clinically relevant information that were incorrectly recognized, the meaning of 72 of them (100% of total instances) was lost for Google ASR and 67 (89.3%) was lost for Amazon ASR, either because the original NLCS were deleted, or were replaced with some other NLCS whose meaning could no longer be ascertained (eg, *Mm-hm* became *Hum*). We did not find any clinically relevant NLCS whose meaning was reversed (eg, *Mm-hm* "yes" became *Uh-uh* "no," or vice versa). Using the total number of NLCS as denominator (ie, 3284), the error rate in recognizing NLCS that conveyed clinically relevant information was 2.2% for Google ASR and 2.0% for Amazon ASR. The lower portion of Table 3 shows the results pertaining to all frequently used NLCS regardless of whether they conveyed clinically relevant information. The error rate was 40.8% for Google ASR and 57.2% for Amazon ASR.

Table 3 also shows a breakdown by error type, that is, substitution (an NLCS utterance was substituted with another NLCS, or with an irrelevant word), deletion (an NLCS utterance was omitted

Table 3. Word error rate (WER) for NLCS

NLCS type	Google ASR					Amazon ASR				
	Substitution rate (%)	Deletion rate (%)	Insertion rate (%)	WER (%)	Total instances as transcribed[a]	Substitution rate (%)	Deletion rate (%)	Insertion rate (%)	WER (%)	Total instances as transcribed[a]
NLCS that conveyed clinically relevant information										
Mm (conveying "yes")	100.0	0.0	0.0	100.0	7	28.6	71.4	0.0	100.0	7
Mm-hm (conveying "yes")	93.0	7.0	0.0	100.0	57	24.6	75.4	0.0	100.0	57
Uh-huh (conveying "yes")	33.3	0.0	0.0	33.3	6	83.3	0.0	0.0	83.3	6
Uh-uh (conveying "no")	66.7	33.3	0.0	100.0	6	83.3	16.7	0.0	100.0	6
Average	86.8	7.9	0.0	94.7	76	34.2	64.5	0.0	98.7	76
All frequently used NLCS										
Average	37.8	2.1	1.0	40.8	3179	20.3	36.9	0.0	57.2	3186

NLCS: non-lexical conversational sounds.

[a]These 2 columns are not identical because of insertion and deletion errors generated by ASR.

from the transcribed text), and insertion (NLCS that were not uttered in the spoken conversation were added to the transcribed text). In general, Google ASR tended to produce a higher rate of substitution errors (37.8%), and Amazon ASR tended to produce a higher rate of deletion errors (36.9%). In other words, Google ASR captured most of the NLCS utterances but failed to transcribe many of them correctly, whereas Amazon ASR omitted many of the NLCS utterances.

Table 4 exhibits some representative examples of the deletion, substitution, and insertion errors. The first row shows that *Mm-hm*, commonly used by patients to convey a "yes" response, was omitted. This deletion error could result in loss of information in the eventual clinical documentation. The second row shows 3 examples of the substitution error. In the first example, the NLCS utterance *Mm-hm* was replaced with "is it," a phrase that does not have any interpretable meaning in the context of the conversation. In the second and third example, the original conversions contained *Mm* and *Uh-uh*, which respectively conveyed the patient's "yes" and "no" answer response to the questions asked by the clinician. In both cases, they were substituted with *Hum-um*, which again does not have any interpretable meaning in the context of the conversation and the substitution could thus lead to loss of information. The last row shows an insertion error wherein *Uh-huh* was added when no such NLCS was uttered by either the patient or the clinician during the spoken conversation.

Common substitution words that the NLCS were changed to, of those that could convey clinically relevant information, are reported in Table 5. As shown in the table, there is no consistent pattern as to with what these NCLS would be replaced by the ASR engines. Supplementary Data File S1 lists common substitution errors for frequently used NLCS that we identified from the empirical data.

ASR performance with non-NLCS

In addition to NLCS, we also performed a separate analysis on the recognition accuracy of non-NLCS words of the 2 ASR engines. When only non-NLCS words were included ($N = 132\ 363$), the word error rate (WER) was 11.8% for Google ASR and 12.8% for Amazon ASR. These error rates are much lower than the results obtained with NLCS. These error rates are also lower than what has been published in the literature (14–65%[6,56]), but higher than the results of studies that evaluated ASR performance in one-on-one dictation (ranging from 5% to 9% or 7% to 40% depending on whether a specialized vocabulary was used[57]). This may be a reflection of the higher quality audio recordings that we produced through re-enacting the clinical encounters in a professional sound studio. Such recordings were, however, still more challenging to process when compared to audio data from one-on-one dictation scenarios.

DISCUSSION

There has been an extensive body of literature studying patient–clinician interactions in the exam room. These studies have examined topics such as how clinicians solicit patient concerns, how patients provide narratives on their symptomatology, and how clinicians conduct shared decision-making processes with patients.[28,58–61] Some of these studies have also looked into non-lexical sounds and the roles that they play in conveying meaning or facilitating clinical conversation.[20] For example, Stivers and Heritage[28] showcased how *Mm-mm* could be used by the patient to answer a clinical question, or how *Mm-hm* could be used as a backchannel by the physician. To the best of our knowledge, the present study is the first to examine the implications of properly handling NLCS in the context of generating clinical documentation using the emerging ambient clinical documentation technology, or "digital scribes." As previously mentioned, this technology holds great promise for reducing documentation burden, mitigating clinician burnout, and improving the comprehensiveness and accuracy of clinical data.

Our evaluation results however show that the 2 contemporary ASR engines, Google Speech-to-Text and Amazon Transcribe Medical, performed poorly in recognizing the NLCS. Many NCLS that conveyed clinically relevant information were omitted, and many were substituted with other NLCS or irrelevant words. Such findings are important because, as documented in the literature[20] and demonstrated by our empirical data, NLCS were frequently used by both patients and clinicians to convey important meaning, for example, "yes" or "no" answers to questions such as "are you allergic to aspirin?" Our analysis of the 36 primary care clinical encounters shows that, on average, NLCS were used more than 30 times per encounter. Some of these NLCS were used to communicate clinically relevant information that, if not properly captured, could result in inaccuracies in clinical documentation and possibly adverse patient safety events. Because the quality of the transcripts generated by ASR engines is critical for the success of downstream tasks, for example, using NLP to transform verbatim transcripts into clinical documentation to facilitate patient care, compliance, and billing for

Table 4. Sample ASR errors

Error type	Spoken conversation	Transcribed text
Deletion	*And your dad had lung cancer?*	*And your dad had lung cancer?*
	Mm-hm.	*[deleted]*
Substitution	*Okay. Your vision is good?*	*Okay. Your vision is good?*
	Mm-hm.	*Is it.*
	And your dentures fit fine?	*And your dentures fit fine?*
	Yep.	*Yep.*
	No problems with them?	*No problems with them?*
	Mm.	*Hum-um.*
	All right. No frequency, no burning?	*All right. No frequency? No burning?*
	Uh-uh.	*Hum-um.*
Insertion	*I leave at 5:45, Monday through Friday, out my front door and walk for 45 minutes.*	*I leave at 5:45 Monday through Friday. Uh-huh. Out in my front door and walk for 45 minutes.*

ASR: automatic speech recognition; NLCS: non-lexical conversational sounds.

Table 5. Common substituted words for each NLCS type that could convey clinically relevant information

	Google ASR		Amazon ASR	
NLCS Type	Substituted to	Number of instances	Substituted to	Number of instances
Mm	*Hum*	26	*Uh-huh*	5
	Um	23	*Um*	5
	Um-hum	12		
Mm-hm	*Um-hum*	733	*Uh-huh*	47
	Uh-huh	6	*Um*	16
	Hum	1	*Mm*	11
Uh-huh	*Huh-uh*	4	*Uh*	64
	Uh	2	*Huh*	11
	Um-hum	1	*Oh*	8
Uh-uh	*Huh-uh*	4		
	Hum-um	1		

ASR: automatic speech recognition; NLCS: non-lexical conversational sounds.

services, our findings provide insights into further developing the ambient clinical documentation technology to improve recognition accuracy and minimize potential patient safety risks.

This finding is also thought-provoking for 3 additional reasons. First, the 2 ASR engines evaluated in this study are specifically tailored for clinical conversations, which have demonstrated superior performance in processing conversational data when compared to other generic ASR engines (eg, Nuance Dragon Medical which is optimized for one-on-one dictation).[62,63] Second, our re-enacted recordings were produced in a professional audio studio setting and thus are likely of much better quality than recordings from realistic clinical settings. Third, the re-enaction was performed by 2 American English native speakers, eliminating the complication that may be caused by speakers' accents. These 3 reasons suggest that our results likely represent an upper bound of potential ASR performance in the context of transcribing NLCS contained in patient–clinician conversations. Real-world conditions will likely result in reduced performance.

The findings of this study suggest that the ability of the contemporary ASR engines in correctly recognizing NLCS has much room for improvement—improvement that we believe is within reach. In listening to the re-enacted recordings, we are confident that the semantic type and the meaning of a majority of the NLCS can be reliably determined by a human listener. This means that different non-lexical sounds exhibit distinct characteristics, and training machines to differentiate them should be technically possible. We thus urge developers of the ambient clinical documentation technology to use such non-lexical sounds, including attributes such as tone, rhythm, and the surrounding context, to ensure that NLCS are not only correctly transcribed, but that their semantic type and conveyed meaning are properly captured. Our findings illustrate how such abilities are essential prerequisites for the success of the subsequent clinical documentation generation processes based on the resultant transcripts. Additionally, we also encourage clinicians to adopt new communication strategies, for example, by verbally confirming the intended meaning of patient answers conveyed by NLCS, to avoid ambiguities and to help ASR achieve best possible performance.

It should be noted that in our study, the WER of Google ASR and Amazon ASR for non-NLCS words—11.8% and 12.8%, respectively—are better than what has been reported in other studies evaluating the performance of conversational ASR engines (ranging from 14% to 65%[6,56]). This difference may be accounted for by the superior quality of the audio recordings that we produced in a professional studio setting for this study. It should also be noted that the volume of NLCS, compared to the number of non-NLCS words exchanged in clinical conversations, is very small (2.4% vs 97.6% according to our empirical data). However, the WERs for NLCS are much higher. As reported earlier, in our study, over 40% of the frequently used NLCS (2.4% of total spoken words), and over 94% of the NLCS that conveyed clinically relevant information (0.06% of total spoken words, 2.3% of all NLCS), were not correctly recognized by Google ASR; and these rates were 57% and 98% for Amazon ASR, respectively. Although misrecognized clinically relevant NLCS comprised less than 0.06% of all spoken words, we believe that such errors could lead to patient safety risks. Lastly, since we did not assess the clinical significance of these errors or the likelihood that clinicians would blindly trust the ASR-recognized results and act on the erroneous information without verification, these may be worthwhile subjects for future studies. For example, Zhou et al[64] identified a 20-fold decrease in clinically relevant errors after physician review of dictated medical reports. That said, we believe that improving ASR recognition accuracy is still of vital importance both to minimize the chance of errors and to reduce clinicians' burden on recognizing and correcting them during their busy clinical work.

This study has several limitations. First, we used the transcripts from 36 primary care encounters in evaluating the ASR engines. The relatively small sample size, coupled with the fact that only the primary care setting was included, may limit the generalizability of our results. Second, we re-enacted the clinical conversations based on the original transcripts in a professional audio studio. Thus, our findings may not reflect the true ASR performance when applied to audio recordings obtained from realistic clinical environments which are susceptible to multiple dimensions of complications such as background noises, interruptions (eg, the clinician answering a phone call), different styles of enunciation and intonation used by patients and clinicians, and the possibility that there may be more than 2 speakers in the room. This does not invalidate our findings,

though, as the real-world conditions will be more challenging and will likely result in poorer ASR performance. Third, our study only evaluated 2 ASR engines. We are aware that there have been other efforts by academic institutions and start-up companies (eg, DeepScribe.ai) to develop next-generation ASR engines specifically to enable the ambient clinical documentation technology. However, we were unable to access these ASR engines for this study. Further, there could be other efforts in developing similar ASR products that we were not aware of when this study was conducted (eg, M*Modal Fluency Align from 3M,[13] suki.ai[65]). Lastly, our evaluation focused on NLCS that conveyed clinically relevant meaning. There could be nonclinical information exchanged during casual conversation with NLCS that is equally critical to patient wellbeing, such as social determinants of health information (eg, food insecurity, transportation access, or financial stability). That said, it would seem reasonable to assume that our findings on ASR performance with respect to the exchange of clinically relevant information would similarly apply to the exchange of other nonclinical information in the exam room.

CONCLUSION

We evaluated the performance of 2 contemporary ASR engines in recognizing NLCS commonly used in the exam room, in the context of assessing the feasibility of the ambient clinical documentation technology. The failure to correctly capture such sounds, many of which could convey critical clinical information (eg, *Mm-hm* as "yes" and *Uh-huh* as "no" to the question "are you allergic to antibiotics?"), may cause inaccuracies in clinical documentation and introduce new patient safety risks. The results show that while both ASR engines yielded better results with non-NLCS words, their performance on correctly recognizing NLCS was suboptimal, with a majority of the NLCS being either omitted or substituted with irrelevant words. Although the clinically relevant NLCS error rate is under 0.06%, such errors could result in missing information or incorrect interpretation of what was expressed by the patient or clinician during a clinical encounter. Future work is therefore needed to improve the performance of ASR, including correct recognition of NLCS, in order to minimize recognition errors and patient safety risks of the ambient clinical documentation technology.

FUNDING

Transcripts used in this work were obtained under NCI R01-CA112379 (JEL) and NIMH R01-MH081098 (MTS). This work was also supported in part by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR001414 (KZ); and the National Institute of General Medical Sciences, National Institutes of Health, through Grant T32-GM008620 (BDT). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

BDT and TLR re-enacted the clinical conversations. BDT and KL analyzed the data. JP helped with developing tools for part of the data analysis. JEL and MTS provided the original transcripts and feedback on the final manuscript. KZ and MTS contributed to the concept development of the study. BDT and KZ drafted the manuscript. All authors contributed to the revision of the manuscript.

SUPPLEMENTARY MATERIAL

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

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

DATA AVAILABILITY

The data analyzed in this study will be shared upon request with proper IRB oversight.

REFERENCES

1. 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 April 23, 2022.
2. Clinical documentation. Thefreedictionary.com. https://medical-dictionary.thefreedictionary.com/clinical+documentation. Accessed April 23, 2022.
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, DC: National Academies Press; 2020. doi:10.17226/25521.
4. Tai-Seale M, Olson CW, Li J, *et al*. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Aff* 2017; 36 (4): 655–62.
5. Sinsky C, Tutty M, Colligan L. Allocation of physician time in ambulatory practice. *Ann Intern Med* 2017; 166 (9): 683–4.
6. 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 (1): 57.
7. 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.
8. Bossen C, Chen Y, Pine KH. The emergence of new data work occupations in healthcare: the case of medical scribes. *Int J Med Inform* 2019; 123: 76–83.
9. Coiera E, Kocaballi B, Halamka J, *et al*. The digital scribe. *NPJ Digit Med* 2018; 1: 58.
10. 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. Accessed October 23, 2022.
11. 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.
12. Ambient Clinical Intelligence: The Exam of the Future has Arrived. Nuance Communications. https://www.nuance.com/healthcare/ambient-clinical-intelligence.html. Accessed April 20, 2021.
13. Ambient clinical documentation and virtual assistant solutions. 3M. https://www.3m.com/3M/en_US/health-information-systems-us/create-time-to-care/clinician-solutions/virtual-assistant-solutions/. Accessed December 9, 2022.

14. *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.
15. 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(5): 808–17.
16. Project EmpowerMD: medical Conversations to Medical Intelligence; 2018. Microsoft. https://www.microsoft.com/en-us/research/project/empowermd/. Accessed April 20, 2021.
17. Gardizy PKRC. The Future of Voice Tech in Medicine is Here. Can it Live Up to the Promise? Boston Globe; April 4, 2020. https://www.statnews.com/2021/04/20/microsoft-nuance-voice-technology-medicine/. Accessed April 21, 2021.
18. Select a Transcription Model | Cloud Speech-to-Text Documentation | Google Cloud. Google Cloud. https://cloud.google.com/speech-to-text/docs/transcription-model. Accessed April 15, 2022.
19. Amazon Transcribe Medical. Amazon Web Serices. https://aws.amazon.com/transcribe/medical/. Accessed April 15, 2022.
20. Heritage J, Maynard DW, eds. *Studies in Interactional Sociolinguistics: Communication in Medical Care: Interaction between Primary Care Physicians and Patients Series Number 20*. Cambridge, England: Cambridge University Press; 2009. doi:10.1017/cbo9780511607172.
21. Drew P, Chatwin J, Collins S. Conversation analysis: a method for research into interactions between patients and health-care professionals. *Health Expect* 2001; 4(1): 58–70.
22. Byrne PS, Long BEL, eds. *Doctors Talking to Patients*. London, England: Royal College of General Practitioners; 1984.
23. Hodges BD, Kuper A, Reeves S. Discourse analysis. *BMJ* 2008; 337: a879.
24. Gunnarsson BL, Linell P, Nordberg B. *The Construction of Professional Discourse*. London, England: Longman; 1997.
25. Wang N, Song Y, Xia F. Studying challenges in medical conversation with structured annotation. 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-1.3.
26. Quiroz JC, Laranjo L, Kocaballi AB, *et al.* Identifying relevant information in medical conversations to summarize a clinician–patient encounter. *Health Informatics J* 2020; 26(4): 2906–14.
27. Drew P, Heritage J. *Studies in Interactional Sociolinguistics: Talk at Work: Interaction in Institutional Settings Series Number 8*. Cambridge, England: Cambridge University Press; 1993.
28. Stivers T, Heritage J. Breaking the sequential mold: answering 'more than the question' during comprehensive history taking. *Text Interdiscip J Study Discourse* 2001; 21: 151–85.
29. Moore RJ, Arar R. *Conversational UX Design*. New York: ACM Books; 2019. doi:10.1145/3304087.
30. Heritage J, Sorjonen M-L, eds. *Between Turn and Sequence*. Amsterdam: John Benjamins Publishing; 2018. doi:10.1075/slsi.31.
31. Benus S, Gravano A, Hirschberg J. The prosody of backchannels in American English. In: Trouvain J, Barry WJ, eds. *Proceedings of the 16th International Congress of Phonetic Sciences ICPhS XVI*. August 6–10, 2007; Saarbrücken, Germany.
32. Ward N. Non-lexical conversational sounds in American English. *P&C* 2006; 14(1): 129–82.
33. Tolba H, O'Shaughnessy D. Towards recognizing "non-lexical" words in spontaneous conversational speech. In: European Speech Communication Association, ed. *Proceedings of 6th European Conference on Speech Communication and Technology (EUROSPEECH'99)*. September 5–9, 1999; Budapest, Hungary.
34. 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
35. Xiong W, Wu L, Alleva F, *et al.* The Microsoft 2017 conversational speech recognition system. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; April 15–18, 2018; Calgary, Canada. doi:10.1109/icassp.2018.8461870.
36. Heldner M, Edlund J, Hirschberg J. Pitch similarity in the vicinity of backchannels. In: *Interspeech 2010*. ISCA; September 26–30, 2010; Chiba, Japan. doi:10.21437/interspeech.2010-58.
37. Cole J. Prosody in context: a review. *Lang Cogn Neurosci* 2015; 30(1–2): 1–31.
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(5): 480–6.
39. 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(10): 2168–73.
40. 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(3): 358–63.
41. 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(12): 1493–504.
42. 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(10): e350–7.
43. Johnson Shen M, Elston Lafata J, D'Agostino TA, *et al.* Lower adherence: a description of colorectal cancer screening barrier talk. *J Health Commun* 2020; 25 (1): 43–53.
44. Lafata JE, Wunderlich T, Flocke SA, *et al.* Physician use of persuasion and colorectal cancer screening. *Transl Behav Med* 2015; 5 (1): 87–93.
45. Park J, Jindal A, Kuo P, *et al.* Automated rating of patient and physician emotion in primary care visits. *Patient Educ Couns* 2021; 104 (8): 2098–105.
46. 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(2): 728–45.
47. Shay LA, Dumenci L, Siminoff LA, *et al.* Factors associated with patient reports of positive physician relational communication. *Patient Educ Couns* 2012; 89(1): 96–101.
48. 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(4): 901–8.
49. 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(2): 164–73.
50. 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. *J Ambul Care Manage* 2017; 40(3): 246–56.
51. 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(1): 76–82.
52. Kawahara T, Yamaguchi T, Inoue K, *et al.* Prediction and generation of backchannel form for attentive listening systems. In: *Interspeech 2016*. ISCA; September 8–12, 2016; San Francisco, CA. doi:10.21437/interspeech.2016-118.
53. Jefferson G. Notes on a systematic deployment of the acknowledgement tokens "Yeah"; and "Mm Hm". *Pap Linguist* 1984; 17(2): 197–216.
54. 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(1): 47–53.
55. Krogsbøll LT, Jørgensen KJ, Gøtzsche PC. General health checks in adults for reducing morbidity and mortality from disease. *Cochrane Database Syst Rev* 2019; 1(1): CD009009.
56. Weninger F, Gaudesi M, Leibold R, *et al.* Dual-Encoder Architecture With Encoder Selection for Joint Close-Talk and Far-Talk Speech Recognition. *arXiv [eess.AS]*; 2021. http://arxiv.org/abs/2109.08744
57. Blackley SV, Huynh J, Wang L, *et al.* Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *J Am Med Inform Assoc* 2019; 26(4): 324–38.
58. Halkowski T. Realizing the illness: patients' narratives of symptom discovery. In: Heritage J, Maynard DW, eds. *Communication in Medical*

Care. Cambridge: Cambridge University Press; 2009: 86–114. doi:10.1017/cbo9780511607172.006.

59. Robinson JD. Soliciting patients' presenting concerns. In: Heritage J, Maynard DW, eds. *Communication in Medical Care*. Cambridge: Cambridge University Press; 2009: 22–47. doi:10.1017/cbo9780511607172.004.
60. Boyd E, Heritage J. Taking the history: questioning during comprehensive history-taking In: Heritage J, Maynard DW, eds. *Communication in Medical Care*. Cambridge: Cambridge University Press; 2009: 151–84. doi:10.1017/cbo9780511607172.008.
61. Stivers T. Treatment decisions: negotiations between doctors and patients in acute care encounters. In: Heritage J, Maynard DW, eds. *Communication in Medical Care*. Cambridge: Cambridge University Press; 2009: 279–312. doi:10.1017/cbo9780511607172.012.
62. Chiu C-C, Tripathi A, Chou K, *et al.* Speech recognition for medical conversations. In: *Interspeech 2018*. ISCA; September 2–6, 2018; Hyderabad, India.
63. 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.
64. Zhou L, Blackley SV, Kowalski L, *et al.* Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. *JAMA Netw Open* 2018; 1(3): e180530.
65. Home. Suki AI; 2022. https://www.suki.ai. Accessed December 12, 2022.