Journal of Healthcare Informatics Research
https://doi.org/10.1007/s41666-025-00193-w

RESEARCH ARTICLE

Assessing the Effectiveness of Automatic Speech Recognition Technology in Emergency Medicine Settings: a Comparative Study of Four AI-Powered Engines

Xiao Luo[1,2] · Le Zhou[3] · Kathleen Adelgais[4] · Zhan Zhang[3]

Received: 12 July 2024 / Revised: 23 January 2025 / Accepted: 24 February 2025

Abstract

This study investigates the potential of advanced automatic speech recognition (ASR) technology for transcribing and recognizing medical information during patient encounters, with the aim of enabling real-time clinical documentation to alleviate clinicians' workload. While ASR holds promise, its effectiveness in noisy and dynamic medical settings, such as emergency medical services (EMS), remains underexplored. To address this, four ASR engines—Google Speech-to-Text Clinical Conversation, OpenAI Speech-to-Text, Amazon Transcribe Medical, and Azure Speech-to-Text—were evaluated using 40 EMS simulation recordings. Transcriptions were analyzed for accuracy across 23 electronic health record (EHR) categories relevant to EMS. Google Speech-to-Text Clinical Conversation showed the best overall performance, excelling in categories such as "mental state" (F1 = 1.0), "allergies" (F1 = 0.912), and "electrolytes" (F1 = 1.0). However, all engines struggled with critical EMS categories like "airway" (F1 = 0.524) and "pupils" (F1 = 0.542). These findings highlight the limitations of current ASR technologies and the need for further advancements to improve accuracy and usability in time-sensitive and high-pressure medical environments.

Keywords Automatic speech recognition · Clinical documentation · Medical scribe · Emergency medicine · Electronic health records

✉ Zhan Zhang
zzhang@pace.edu

[1] Department of Management Science and Information Systems, Oklahoma State University, Stillwater, OK, USA

[2] School of Medicine, Indiana University, Bloomington, IN, USA

[3] School of Computer Science and Information Systems, Pace University, New York, NY, USA

[4] School of Medicine, University of Colorado, Aurora, CO, USA

Published online: 19 March 2025

1 Introduction

Clinical documentation is a critical but time-consuming and challenging task in healthcare. While electronic health records (EHRs) have been widely adopted to streamline this process, they have contributed to increased clinician burnout [1], reduced time spent on patient care [2], and lower patient satisfaction [3]. These unintended consequences of EHR use are particularly evident in time-critical and dynamic settings, such as emergency medical services (EMS), where clinicians operate in fast-paced, high-pressure, out-of-hospital environments [4–6].

During EMS care, real-time clinical documentation using handheld EHR systems (e.g., tablet devices) is often impractical for various reasons [7, 8]. For instance, EMS clinicians are typically both physically and cognitively occupied with high-acuity patients and hands-on care activities, limiting their capacity to use handheld EHR systems in real time to document all necessary information. Furthermore, the interruptive nature of EMS work, often caused by patients and family members, further complicates the documentation task. As such, EMS clinicians often delay documentation until after the patient has been handed off to the receiving hospital team [8]. However, this practice is a primary cause of incomplete, delayed, or erroneous documentation in EMS [9]; for example, nearly 40% of fields on EMS records have been reported to contain missing or incorrect information [10]. These findings reveal that the current design of handheld EHR devices is not suited well to the highly dynamic, hands-busy EMS workflow, highlighting the need for novel approaches to address longstanding issues in clinical documentation in EMS and other time-critical medical settings.

Researchers have proposed using automatic speech recognition (ASR) technologies, powered by natural language processing (NLP) techniques, to automate, at least in part, clinical documentation [11, 12]. These technologies enable clinicians to dictate patient information into the free-text fields of the EHR. They can also record and process conversations during patient-provider encounters to extract and summarize relevant clinical information, which could, for instance, be used to populate specific EHR fields, create billing codes, and generate decision and diagnostic support.

Earlier research on ASR technologies has reported mixed findings regarding their effectiveness for clinical documentation [12]. For example, several studies reported reductions in documentation time (e.g., by 19 to 92%) [13, 14] and turnaround time (often by over 90%) [15], while other studies noted increases in documentation time, ranging from 13.4 to 200% [16, 17]. It is worth noting that these studies evaluated ASR technologies developed prior to the exponential advancements in artificial intelligence (AI). Today's advanced AI-powered ASR tools, offered by companies such as Google [18] and Amazon [19], are expected to significantly improve clinical documentation. However, most studies evaluating these tools have been conducted in quieter medical settings, such as patient exam rooms or doctors' offices, and have primarily focused on one-on-one interactions, such as those between a physician and a patient during clinical encounters [11,

20, 21]. As a result, their performance in dynamic, noisy, and multi-speaker environments, like EMS settings, remains largely unexplored.

In this study, we report an empirical assessment of how four commercially available AI-powered ASR engines perform in transcribing and recognizing medical information from 40 audio recordings of high-fidelity EMS simulations. We aim to address the following research questions through this work: (1) How effectively do AI-powered ASR engines transcribe and recognize clinically relevant information in dynamic and noisy emergency care environments? (2) What are the common types of transcription errors made by ASR engines in EMS settings? Answering these research questions can inform the further development and improvement of AI-powered ASR tools to support real-time clinical documentation in dynamic and noisy medical settings.

2 Methods

2.1 Automatic Speech Recognition Engines Evaluated

This study focuses on evaluating the potential of leveraging ASR engines to facilitate clinical documentation in emergency care settings, such as EMS, which are often noisy, dynamic, and involve multiple care providers in the care process, as opposed to quiet care settings such as medical exam rooms or doctor's offices. Based on our research and literature review at the time of this study (April 2023), we identified Google Cloud Speech-to-Text [22] with the medical conversations model, OpenAI automatic speech recognition model Whisper [23], Amazon transcribe medical with conversation model [19], and Microsoft Azure Speech-to-Text [24] as among the most advanced commercial ASR engines publicly available for evaluation. This selection was based on several key factors, including general performance metrics such as accuracy, robustness in handling complex medical vocabulary, support for conversational speech in clinical contexts, and availability of specialized models for medical transcription [25–28]. We refer to these four ASR engines in the following content as "Google ASR," "Amazon ASR," "OpenAI ASR," and "Azure ASR," respectively.

2.2 Evaluation Dataset

The assessment data were generated from 40 audio recordings of high-fidelity EMS simulations that were conducted in a mobile simulation laboratory that was outfitted to resemble the back of an ambulance [29]. The simulation scenarios varied by clinical conditions and required treatments, including a 15-month-old with seizures, a 1-month-old with hypoglycemia, and a 4-year-old with clonidine ingestion. The simulations used high-fidelity patient mannequins, which could be programmed to follow a sequence and even respond to interventions performed by the team of participants. Four to six participants were involved in each simulation. The participants were active, full-time EMS clinicians working for a large fire-based EMS service.

These simulations replicated the dynamic EMS environment, including background noise (i.e., from a distraught parent), overlapping speech from multiple speakers, an unpredictable need for clinical care, and the varied composition of the team working together. The simulation laboratory was outfitted with three cameras (one above the team, one at the foot of the patient, and one at the head of the patient) to capture different angles of the simulation and with a single microphone to capture audio. A facilitator at an audio-visual control station ensured continuous audio and video feeds and monitored participants' audibility.

The simulation recordings were transcribed by professional transcriptionists and verified by medical experts as part of our previous project examining EMS workflow [30, 31]. These transcripts serve as the ground truth for evaluating the ASR engines. The length of the video recordings of simulations varies, ranging between roughly 9 and 14 min (with an average length of 658.925 s, or close to 11 min). The average word count in the transcribed ground truth is about 1307. Although there were only three different clinical scenarios, participants had to manage multiple patient conditions within each scenario, such as cough, congestion, fever, and seizure. Table 1 in the supplemental materials provides detailed information about each recording. The Institutional Review Board of Pace University thoroughly reviewed the study protocol and determined it to fall under the category of nonhuman subjects' research.

2.3 Annotation Process for Classifying the Transcribed Ground Truth into EMS EHR Fields

The main objective of this research is to assess and compare the ability of different AI-powered ASR engines towards accurately transcribing content associated with a set of structured fields within the general EMS EHR. Examples of such structured fields include "age," "gender," "past medical history," and "allergies."

We used a multi-step process to annotate the ground truth transcripts. First, we identified and created a list of typical structured fields in EMS EHR systems and categorized them into five high-level categories (Table 1)—"ePatient," "eHistory," "eVital," "eSituation," and "eMedication"—based on the examination of two EMS EHR systems' structures (ESO and Epic) as well as the National EMS Information System (NEMSIS) data structure [32]. The NEMSIS data structure is a standardized framework for collecting, storing, and sharing EMS data across various agencies and healthcare facilities in the USA, allowing EMS clinicians to document patient care and outcomes in a structured and uniform manner.

After this step, two annotators (ZZ and LZ) independently annotated six ground truth transcripts from three different scenarios (two transcripts for each scenario) by marking content that belongs to these structured fields. In this step, we calculated the inter-rater agreement using Cohen's kappa coefficient; two annotators achieved an "almost perfect agreement" on the mapping between content and EHR fields (kappa value is 0.80). All disagreements between the annotators were resolved in meetings that involved a third annotator (XL).

Afterwards, a data dictionary, as shown in Table 1, was created, which includes five high-level categories along with the corresponding EMS EHR fields under each

Table 1 Categories and fields of data annotation with exemplars

Category	EMS EHR fields	Example
ePatient	Gender	Male
	Age	15 month old
	Family information	Has 3 year old brother
eSituation	Complaint/symptoms	Vomiting, diarrhea
	Lung sounds	Clear
	Mental state	Fussy
	Trauma	No signs of trauma
	Airway	Clear
eHistory	Allergies	Never had any allergies
	Past medical history	ADHD
eVital	AVPU (alert, verbal, pain, unresponsive)	Only responds to painful stimulus
	BGL (blood glucose level)	100
	BP (blood pressure)	58 over 36
	Capillary refill time	6 s
	ECG (heart rate)	190
	Electrolytes	Potassium
	Pupils	A little dilated and slow to respond
	Pulse	190
	RESP (respiration rate)	22 to 26
	SPO2 (oxygen saturation)	100%
	Temperature	98.4
eMedication	Medication	D10 20 ml
	Treatment	Put on mask to help breathing

category, as well as annotated examples. Finally, one annotator (LZ) annotated the rest of the transcripts based on the data dictionary, while another annotator (ZZ) reviewed all the annotations to ensure correctness and consistency of annotation across all 40 transcripts. It is worth noting that the annotation process was iterative and collaborative, involving back-and-forth correction of previous annotations.

2.4 Automated Content Extraction from ASR Transcribed Text for Performance Evaluation

The primary goal of this research is to evaluate whether the ASR system accurately transcribes information for clinical documentation. Medical conditions can be described differently in various contexts; for instance, “high blood pressure” and “hypertension” may appear in separate scenarios. As long as the ASR systems transcribe these terms correctly, they can be documented appropriately. This study does not involve any additional steps for medical term normalization. To determine the ASR engine’s ability in transcribing clinically relevant information

associated with structured fields in the EMS EHR, an automated content extraction framework is designed, shown as Fig. 1. This extraction framework aims to identify transcribed text that aligns with the annotated ground truth transcription. First, the timestamp of the ground truth transcription that contains the annotated content belongs to the EHR fields and is used to locate the searching span of the text analysis. For example, if the ground truth transcription contains annotated age information, based on the timestamp of the transcription, we located the initial timestamp of the ASR transcribed data. From the initial timestamp, we included five ASR transcribed chunks or sentences before and after that timestamp for the text analysis. Then, a sliding window of length n that matches the number of words in the annotated content is used to generate candidate content for extraction from the ASR transcribed sentences for similarity measurements. The sliding window does not cross punctuation tokens. For example, to match annotated age "one month old" in the ground truth transcription, a sliding window of length 3 can be applied to text "All right, so this is Susie, one month old." to generate the following candidates "so this is," "this is Susie," and "one month old."

The string similarity metric Levenshtein distance was employed in this study to identify the most suitable matches in the ASR transcribed data compared to the annotated ground truth transcriptions. The Levenshtein distance serves as a metric for quantifying the dissimilarity between two sequences of words. This metric calculates the minimum number of edits required to transform one word sequence into another. The permissible edits include inserting, deleting, or substituting different letters or words. The FuzzyWuzzy ratio function [33] was used to estimate the Levenshtein distance similarity ratio between a candidate and annotated ground truth. The candidate demonstrating the highest Levenshtein distance similarity ratio was chosen as the best match and utilized for quantitative evaluation. Human evaluation of selected samples indicates that the Levenshtein

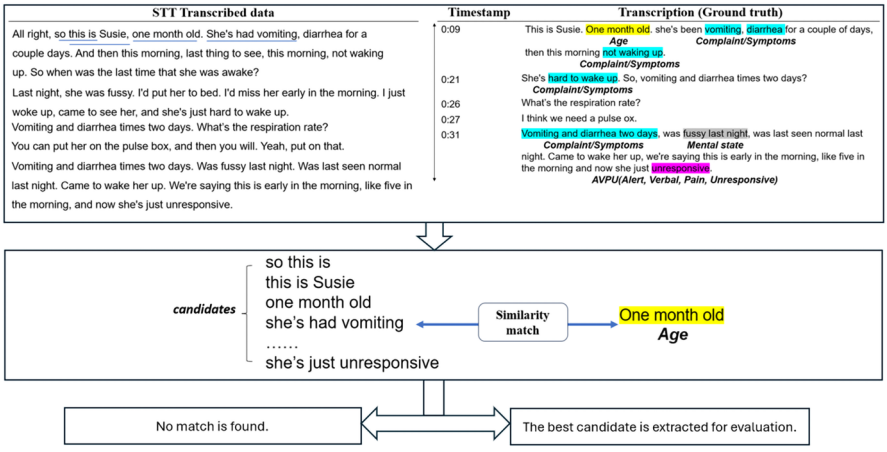

Fig. 1 A text analysis framework to extract the matching content

distance similarity ratio effectively identifies the most similar text strings when they contain the same number of words without considering semantic similarity.

2.5 ASR Performance Evaluation Methods

Two performance evaluation methods were employed to evaluate the extracted content against the annotated ground truth.

Named Entity Recognition-Based Evaluation Metrics After the matching content is extracted, the traditional evaluation metrics that are applied for NER tasks were used for evaluation [34–36]. All extracted content is converted into words. Then, the word-based precision (P), recall (R), and F1 are calculated for each category of the structured fields.

Semantic Similarity-Based Accuracy Measurement For the semantic evaluation, cosine similarity between the embeddings of the annotated ground truth and the best matching candidate is calculated. In this research, the sentence-BERT [23] was used to generate the embeddings. If the cosine similarity is above the threshold 0.8, it is a correct match. The calculated accuracy for each category of the structured fields is based on the frequency of the correct match.

3 Results

3.1 Descriptive Statistics of the Evaluation Data

The ground truth transcription of the 40 audio recordings contained between 744 and 1957 words (mean 1340.5, Std 297.57). The number of speakers in each recording ranged from 4 to 8 (mean 6, Std 1.02). The duration of the reenacted audio recordings was between 528 and 928 s (mean 652.5, SD 86.58).

The detailed statistics of the annotated data can be found in Table 2. One interesting observation is that the total number of occurrences in each EHR field varies. The most frequent fields are “complain/symptoms,” “medication,” and “treatment,” whereas the less frequent fields are “mental state” and “electrolytes.” Some categories occur in more than 87.5% of the recordings, such as “age,” “complain/symptoms,” “past medical history,” “BP,” “ECG,” “RESP,” “medication,” and “medication equipment.” Another interesting observation is that the average length of the annotated content in most fields is no more than three words. The category “trauma” has the longest average description, with 3.89 words, possibly because it often has detailed descriptions about the patient’s situation after the examination.

3.2 ASR Efficiency Comparison

The average length of the audio files is 660 s. We assessed the efficiency of different ASR engines in transcribing these audio files. The average processing times for

Table 2 Statistics of the annotated transcription data

Category	EMS EHR fields	Total # occurrences	% of recordings	Max # occurrences in each recording	Average length by words
ePatient	Gender	26	62.5%	2	1.00
	Age	92	100%	5	2.47
	Family information	22	30%	4	3.59
eSituation	Complaint/symptoms	379	100%	22	2.88
	Lung sounds	37	47.5%	5	2.35
	Mental state	1	2.5%	1	1.00
	Trauma	27	30%	5	3.89
	Airway	14	25%	2	3.07
eHistory	Allergies	21	42.5%	2	2.38
	Past medical history	190	92.5%	15	2.77
eVital	AVPU	105	85%	7	2.87
	BGL	68	80%	4	1.38
	BP	158	92.5%	12	2.61
	Capillary refill time	15	27.5%	4	2.13
	ECG	108	87.5%	11	1.73
	Electrolytes	3	2.5%	3	1.00
	Pupils	11	15%	4	2.54
	Pulse	69	65%	7	1.84
	RESP	119	87.5%	11	2.81
	SPO2	69	72.5%	8	1.55
	Temperature	37	50%	4	1.86
eMedication	Medication	457	100%	26	2.64
	Treatment	417	97.5%	24	2.42

individual audio files using Google ASR, OpenAI ASR, Amazon ASR, and Azure ASR were 211.39, 265.05, 32.38, and 331.21 s, respectively. Notably, Amazon ASR demonstrated significantly faster performance compared to the other three services, potentially due to its more efficient batch processing capabilities.

3.3 ASR Performance Evaluation and Comparison

To evaluate the overall performance of the ASR engines in transcribing the EMS simulation recordings, we first applied the word error rate (WER) [37] metric to each recording. The results showed WER of 4.54, 2.95, 2.65, and 2.44 for Google ASR, OpenAI ASR, Amazon ASR, and Azure ASR, respectively. A higher WER indicates lower performance. Upon further analysis, we discovered that Google ASR transcribed more non-lexical conversational sounds than other ASR engines, such as "Uh-uh" within the content, which contributed to its higher WER. Since WER is a generic evaluation metric without focusing on clinical information, further

comparisons of these ASR engines' capabilities in correctly transcribing clinical terminologies for clinical documentation are needed.

Table 3 presents the performance of the four ASR engines in transcribing content within EMS EHR categories, evaluated using NER-based metrics. The results indicate that Google ASR outperformed the other three ASR engines in most categories and overall. However, compared to Google ASR, the other three ASR engines performed slightly better in transcribing information for certain categories, such as "gender," "age," and "airway." Amazon ASR performed significantly better than the others in transcribing "gender" information. Upon investigation, we found that "male" was often transcribed as "mail" by Google ASR and OpenAI ASR. All ASR engines performed well in transcribing information about "electrolytes." This may be due to fewer instances in the evaluation data, and the relatively short description of "electrolytes." OpenAI ASR demonstrated higher performance in the "airway" category, which typically has an average length of more than three words.

Table 4 shows the performance comparison based on the semantic similarity evaluation. Google ASR still achieved the highest average performance. The performance of Google ASR in categories such as "past medical history," "ECG," "medication," and "treatment" showed greater improvement compared to the other categories when semantic similarity evaluation was used. Through our investigation, we discovered that semantic measurement could identify cases where transcribed words had similar meanings to the ground truth, even if the words were not an exact match. For example, if the ground truth was "diarrhea for the past two days," and the transcribed word was "diarrhea for the past few days," the semantic similarity between these two phrases was more than 0.9, and it was counted as a correct case. On the other hand, the performance in the "pupils" category remained low even though semantic measurements were used.

In addition to the performance on EMS EHR categories, we also assessed the ASR engines' performance in transcribing content within these categories for each individual recording. The number of entities in each EMS EHR category varied between recordings. We calculated precision, recall, and F1 scores using NER-based metrics for all entities in each recording, then averaged these scores across entities to determine the performance for that recording. Finally, we averaged the performance across all 40 recordings to compare the ASR engines. The results are shown in Table 5. Similar to the performance on EMS EHR categories, Google ASR performed better than the others.

3.4 Types of Errors in Transcribing

Given that the ASR did not perform well for some EHR categories, we analyzed the errors and grouped them into three types: substitution (the transcribed word is different from the ground truth), approximation (the transcribed word differs by a few characters from the ground truth, but has a similar pronunciation), and truncation (the transcribed word omits certain characters from the ground truth). These error types were used in the literature [11] to summarize and categorize speech recognition errors.

Table 3 Performance comparison of the ASR engines using NER-based metrics

EMS EHR fields	Google ASR			OpenAI ASR			Amazon ASR			Azure ASR		
	F1	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**
Gender	0.654	0.654	0.654	0.423	0.423	0.423	**0.962**	**0.962**	**0.962**	0.692	0.692	0.692
Age	0.667	0.666	0.671	0.398	0.398	0.398	**0.673**	**0.673**	**0.673**	0.649	0.649	0.649
Family information	**0.708**	**0.704**	**0.714**	0.564	0.564	0.564	0.491	0.491	0.491	0.593	0.593	0.593
Complaint/symptoms	**0.663**	**0.663**	**0.665**	0.573	0.573	0.573	0.500	0.499	0.502	0.544	0.545	0.543
Lung sounds	**0.673**	**0.673**	**0.673**	0.378	0.378	0.378	0.461	0.461	0.461	0.495	0.495	0.495
Mental state	**1.000**	**1.000**	**1.000**	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
Trauma	**0.763**	**0.754**	**0.775**	0.711	0.713	0.710	0.612	0.607	0.620	0.534	0.534	0.534
Airway	0.480	0.479	0.483	**0.524**	**0.524**	0.524	0.345	0.345	0.345	0.399	0.399	0.399
Allergies	**0.912**	**0.912**	**0.912**	0.702	0.702	0.702	0.671	0.671	0.671	0.730	0.730	0.730
Past medical history	**0.789**	**0.789**	**0.789**	0.514	0.514	0.515	0.436	0.435	0.437	0.569	0.570	0.568
AVPU	**0.642**	**0.641**	**0.643**	0.453	0.452	0.454	0.510	0.510	0.511	0.514	0.514	0.513
BGL	**0.806**	**0.805**	**0.807**	0.662	0.663	0.662	0.763	0.763	0.762	0.655	0.655	0.654
BP	**0.597**	**0.596**	**0.597**	0.544	0.544	0.544	0.526	0.526	0.526	0.292	0.292	0.292
Capillary refill time	**0.633**	**0.633**	**0.633**	0.600	0.600	0.600	0.517	0.517	0.517	0.417	0.417	0.417
ECG	**0.777**	**0.777**	**0.778**	0.588	0.588	0.588	0.417	0.416	0.418	0.608	0.608	0.609
Electrolytes	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Pupils	**0.542**	**0.541**	**0.543**	0.416	0.416	0.416	0.422	0.422	0.422	0.543	0.543	0.543
Pulse	**0.766**	**0.766**	**0.766**	0.451	0.451	0.451	0.548	0.548	0.548	0.577	0.576	0.579
RESP	**0.651**	**0.649**	**0.654**	0.452	0.452	0.453	0.560	0.558	0.562	0.607	0.606	0.608
SPO2	**0.641**	**0.641**	**0.641**	0.402	0.402	0.402	0.545	0.545	0.545	0.505	0.505	0.505
Temperature	**0.709**	**0.708**	**0.709**	0.573	0.573	0.573	0.511	0.511	0.511	0.617	0.617	0.617
Medication	**0.576**	**0.577**	**0.576**	0.461	0.462	0.460	0.411	0.411	0.412	0.425	0.426	0.424
Treatment	**0.565**	**0.565**	**0.566**	0.471	0.471	0.471	0.423	0.422	0.424	0.435	0.435	0.435

Table 3 (continued)

EMS EHR fields	Google ASR			OpenAI ASR			Amazon ASR			Azure ASR		
	F1	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**
Average	**0.705**	**0.704**	**0.707**	0.545	0.545	0.545	0.564	0.564	0.565	0.568	0.568	0.568

The numbers in bold highlight the best performance of each category

Table 4 Performance comparison of the ASR engines using semantic similarity-based accuracy (similarity ≥ 0.8)

EMS EHR fields	Google ASR	OpenAI ASR	Amazon ASR	Azure ASR
Gender	0.771	0.778	**0.998**	0.954
Age	0.800	0.639	0.819	**0.838**
Family information	**0.814**	0.801	0.739	0.804
Complaint/symptoms	0.731	**0.759**	0.660	0.749
Lung sounds	**0.741**	0.639	0.639	0.700
Mental state	**1.000**	0.984	0.965	0.984
Trauma	0.753	**0.806**	0.724	0.739
Airway	0.648	**0.733**	0.485	0.586
Allergies	**0.903**	0.834	0.787	0.838
Past medical history	**0.834**	0.695	0.597	0.756
AVPU	**0.736**	0.678	0.686	0.720
BGL	**0.896**	0.800	0.816	0.815
BP	**0.785**	0.657	0.742	0.730
Capillary refill time	0.762	**0.913**	0.779	0.744
ECG	**0.855**	0.736	0.702	0.793
Electrolytes	**1.000**	1.000	1.000	1.000
Pupils	0.594	0.595	0.658	**0.675**
Pulse	**0.807**	0.656	0.668	0.739
RESP	**0.792**	0.622	0.737	0.736
SPO2	**0.728**	0.635	0.726	0.717
Temperature	0.790	0.759	0.724	**0.824**
Medication	**0.711**	0.649	0.616	0.639
Treatment	**0.683**	0.636	0.611	0.636
Average	**0.789**	0.739	0.734	0.770

The numbers in bold highlight the best performance of each category

Table 5 Comparison of average ASR engine performance across all recordings using NER-based metrics

ASR engine	F1	R	P
Google ASR	**0.643**	**0.645**	**0.643**
OpenAI ASR	0.502	0.502	0.502
Amazon ASR	0.490	0.490	0.489
Azure ASR	0.495	0.495	0.496

The numbers in bold highlight the best performance of each metric

Table 6 provides representative examples of these error types. The first five cases show substitution errors, where the transcribed word(s) differ from the ground truth conversation. In other words, most of these errors alter the meaning of the original words, which could result in wrong documentation in the EHR. The next four cases are examples of approximation errors, indicating that the ASR transcribed

Table 6 Examples of the transcribing errors in each type

Error type	Category	Ground truth conversation	Transcribed text
Substitution	AVPU	Responsive	Responding
	AVPU	Crying	Trying
	Complaint/symptoms	Cyanosis	Diagnosis
	Medication	Clonodine	Klonopin
	Treatments	Glucometer	Glucose
Approximation	SPO2	Capnography	Catnography
	Medication	Catapres	Cataphras
	Complaint/symptoms	Hypertension Syncope, bradycardia	Hypotension Syncope, bradycardia
	Complaint/symptoms	Hypoglycemia	Hyperglycemia
Truncation	Complaint/symptoms	Asystolic period	Systolic period
	Treatments	Cardiovert	Cardio
	Treatments	OPA	OP

words into ones that sound similar but have totally different meanings. In the medical domain, transcribing "hypertension" as "hypotension" or "hypoglycemia" as "hyperglycemia" would raise critical concerns. In the truncation error type, as demonstrated in the last three cases, instances of non-exact matches—even when only one character is omitted in the transcription—could potentially lead to a completely different representation of the patient's situation. For instance, the acronym "OPA" stands for oropharyngeal airway, but the transcribed phrase "OP" omits the character "A," leading to an inaccurate representation of the patient's status.

4 Discussion

With the growing interest in leveraging ASR engines to facilitate and automate clinical documentation, a few studies have examined and evaluated their performance in recognizing medical information and automating (part of) clinical documentation during patient-provider encounters. For example, Tran et al. [20] assessed the performance of two ASR engines (Google Speech-to-Text Clinical Conversation and Amazon Transcribe Medical) in recognizing non-lexical conversational sounds (e.g., "Mm-hm," "Uh-uh") during primary care encounters. In a similar vein, a few other studies [11] have used and evaluated ASR engines for transcribing clinical conversations in domains such as primary care [18], orthopedic encounters [38], home hemodialysis [39], and telemedicine [40]. However, these prior studies were primarily conducted in quieter medical settings, such as patient exam rooms, and focused on transcribing and summarizing one-on-one patient-clinician interactions. The performance of ASR engines in more dynamic, noisy, and fast-paced environments involving conversations among multiple individuals (including care providers, patients, and patients' caregivers) has been understudied. To the best of our knowledge, this

study is the first to examine commercially available AI-powered ASR engines for transcribing clinical conversations in dynamic emergency medicine settings.

Our evaluation results show that the overall quality of the transcripts generated by current state-of-the-art ASR engines falls short in accurately transcribing and recognizing medical information in the busy and noisy EMS environment. All of the ASR engines encountered difficulties recognizing information in key EMS EHR fields, such as "airway" and "pupils." Upon investigation, we found that certain words in these categories were not accurately transcribed. For instance, "good chest rise" in the "airway" category was transcribed as "good chest wise," and "nonreactive when I looked" in the "pupils" category became "nonreactive where there was." After investigation of two human annotators, these transcription errors could be attributed to the noisy EMS environment and the speakers' volume levels. Since all recordings involved four or more EMS clinicians and had relatively equivalent complexity in medical scenarios, we did not find a direct relationship between the ASR transcription performance and the number of involved EMS clinicians or the complexity of the medical scenarios.

For all four ASR engines, a number of pieces of content containing clinically relevant information were truncated, with many being replaced by irrelevant words, particularly within categories "complaint/symptoms," "treatment," and "medication." This finding aligns with the existing literature [41–43], which suggests that without adding clinical context to refine the models, ASR engines are likely to have difficulties in recognizing "complaint/symptoms," "treatment," and "medication," even when some of it appears frequently in the recordings. Our examination of the 40 EMS audio recordings indicates that over 87.5% of communications during EMS scenes contain clinical information within these categories ("complaint/symptoms," "treatment," and "medication"). Accurate documentation in these fields is therefore essential, especially for downstream tasks such as generating real-time decision support to enhance EMS care. Failure to capture such information correctly could lead to inaccuracies in clinical documentation and potentially adverse patient safety incidents.

The results of this study indicate substantial potential for improving state-of-the-art ASR engines in accurately recognizing clinical information in EMS communication, especially with the advance of AI and advanced large language models (LLMs), such as GPT-4 [44], Me-Llama [45], and others. More specifically, LLMs can be finetuned to correct transcription errors by leveraging medical knowledge and context to accurately extract the clinical information [46, 47]. For example, LLMs could be fine-tuned to modify transcribed text from "heart rate 90 to 105. Fifty-five to 70. Blood pressure is systolic. Hypertensive, tachycardia, just dehydrated" to "heart rate, 90 to 105. Fifty-five to 70. Blood pressure is systolic. Hypotensive, tachycardia, just dehydrated," based on range of blood pressure according to the health guidelines. Additionally, LLMs can be finetuned to identify the correct numeric value (e.g., "fifty-three over twenty-eight") when provider spoke a wrong value followed by a correct one (e.g., "Fifty-three over 25 … Twenty-eight"). Our future strategy involves creating a training dataset that encompasses the common errors made by ASR engines in the EMS context. This will enable us to fine-tune LLMs to identify errors based on context, ensuring not only accurate transcription

but also capturing the intended words and conveyed meaning effectively. Additionally, our future work also includes field testing to assess the performance of fine-tuned LLMs in transcribing real-time clinical conversations in dynamic and noisy environments such as EMS.

However, despite the great potential of ASR engines and LLMs in transcribing clinical conversations and automating EHR documentation, there are several considerations for fully implementing such technology in real clinical settings. For example, the accuracy of current ASR engines in transcribing spontaneous speech and extracting relevant clinical information from fragmented conversations has been a concern for many healthcare clinicians [48, 49]. Therefore, human validation is still required in the process of automated EMS clinical documentation to prevent any errors or biases that may arise from the system or LLMs [50–52]. Future work can look into whether the human validation task increased the documentation time, as prior work found that editing ASR-generated reports took longer than traditional dictation and transcription [53]. Another consideration is related to user acceptance and usability of such tools, that is, whether these tools are perceived to be easy to use and useful, and how adopting them can impact clinicians' workflow. Limited work has engaged end-users in evaluating ASR engines, with only two exceptions where a qualitative evaluation with end-users was conducted to elicit clinicians' experience with using such tools [39, 40]. Future work should take a human-centered approach to systematically evaluate the tools' usability, clinical validity, and impact on clinical workflow. Finally, as some medical settings (e.g., EMS) do not have a dedicated role for documentation and require clinicians to be hands-on almost all the time [7], it is critical to consider how to facilitate the use of ASR tools in such dynamic and hands-busy environment. Recent work has proposed using wearable devices (e.g., smart glasses [54]) to enable hands-free clinical documentation. Future work can explore the potential of combining ASR engines with hands-free, wearable devices integrated with EHR systems to better support real-time transcription and documentation in dynamic and hands-busy medical settings.

There are several limitations in this study. First, only three clinical scenarios were used in the simulations, and only 40 audio recordings from EMS training simulations were utilized to assess the ASR engines. The relatively limited sample size might affect the generalizability of our findings. However, the use of three distinct clinical scenarios and having multiple groups of EMS clinicians perform each scenario may better isolate ASR performance issues arising from the technology's intrinsic limitations rather than scenario-specific factors. Second, our results might not accurately reflect the actual performance of ASR systems when applied to real EMS environments, which are subject to various complicated factors, including possible more interruptions, variations in communications between patients and clinicians, and the potential for increased outdoor noise levels, and operating in a moving vehicle. These conditions are expected to present greater challenges and may lead to lower ASR performance. Third, our study only evaluated four ASR engines. We acknowledge the existence of other initiatives led by academic institutions and companies (such as tecdoc.ai) aimed at developing next-generation ASR engines tailored for clinical documentation. However, some ASR engines did not provide an open API for evaluation. Finally, our assessment focused on evaluating the transcription

of conversations specifically in the pediatric EMS context. Further research is needed in other noisy and dynamic settings, such as emergency departments and trauma resuscitations, where more complex teams are present. Additionally, evaluation of generalizability to the adult EMS context is necessary, given that clinical terms in different medical contexts may have varying meanings.

5 Conclusion

We assessed the performance of four contemporary ASR engines in analyzing conversations within EMS care to determine the potential of using such technology to facilitate and even automate real-time clinical documentation. Our results indicate that the Google ASR engine outperformed the other three ASR engines in recognizing clinical information across most categories in EMS EHR. However, the overall performance of all engines remains suboptimal, with significant clinical information often omitted or replaced with irrelevant words. Such errors can result in missing or incorrectly interpreted patient information and treatment during EMS care. Therefore, there is a need to improve the performance of ASR engines to ensure accurate recognition of all critical clinical information, enabling effective automated documentation in the EMS setting. Future work should focus on enhancing ASR accuracy to minimize recognition errors and reduce patient safety risks associated with EMS clinical documentation technology.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41666-025-00193-w.

Acknowledgements We want to thank Colorado EMS for Children which supported the original data collection during simulations.

Author Contribution XL, KA, and ZZ conceptualized and designed the study. KA and ZZ provided data resources. XL and ZL conducted data analysis. XL, ZL, and ZZ wrote the main manuscript text. KA reviewed and revised the manuscript.

Funding This work was supported in part by funding from the National Science Foundation (Award# 2237097) and National Institute of Health (Award# 1R15LM014556-01).

Data Availability No datasets were generated or analysed during the current study.

Declarations

Ethical Approval Not applicable.

Competing Interests Dr. Zhan Zhang is an Associate Editor of the Journal of Health Informatics Research. Other authors declare that they have no conflicts of interest.

References

1. Arndt BG et al (2017) Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 15(5):419–426

2. Sinsky C et al (2016) Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 165(11):753–760
3. Pelland KD, Baier RR, Gardner RL (2017) 'It is like texting at the dinner table': a qualitative analysis of the impact of electronic health records on patient–physician interaction in hospitals. BMJ Health Care Inf 24(2).
4. Sarcevic A, Ferraro N (2017) On the use of electronic documentation systems in fast-paced, time-critical medical settings. Interact Comput 29(2):203–219
5. Park SY, Chen Y (2012) Adaptation as design: learning from an EMR deployment study. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems
6. Park SY, Lee SY, Chen Y (2012) The effects of EMR deployment on doctors' work practices: a qualitative study in the emergency department of a teaching hospital. Int J Med Inf 81(3):204–217
7. Zhang Z et al (2022) Characteristics and challenges of clinical documentation in self-organized fast-paced medical work. Proc ACM Hum Comput Interact 6(CSCW2):1–21
8. Pilerot O, Maurin Söderholm H (2019) A conceptual framework for investigating documentary practices in prehospital emergency care. In Proceedings of the Tenth International Conference on Conceptions of Library and Information Science, Ljubljana, Slovenia. Information Research, 24(4), paper colis1931. Retrieved from http://InformationR.net/ir/24-4/colis/colis1931.html
9. Laudermilch DJ et al (2010) Lack of emergency medical services documentation is associated with poor patient outcomes: a validation of audit filters for prehospital trauma care. J Am Coll Surg 210(2):220–227
10. Holzman TG (1999) Computer-human interface solutions for emergency medical care. Interactions **6**(3):13–24
11. van Buchem MM et al (2021) The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med 4(1):57
12. Blackley SV et al (2019) Speech recognition for clinical documentation from 1990 to 2018: a systematic review. J Am Med Inform Assoc 26(4):324–338
13. Vorbeck F et al (2000) Report generation using digital speech recognition in radiology. Eur Radiol 10:1976–1982
14. Alapetite A (2008) Speech recognition for the anaesthesia record during crisis scenarios. Int J Med Inf 77(7):448–460
15. Prevedello LM et al (2014) Implementation of speech recognition in a community-based radiology practice: effect on report turnaround times. J Am Coll Radiol 11(4):402–406
16. Bhan SN, Coblentz CL, Ali SH (2008) Effect of voice recognition on radiologist reporting time. Can Assoc Radiol J 59(4):203
17. Issenman RM, Jaffer IH (2004) Use of voice recognition software in an outpatient pediatric specialty practice. Pediatrics 114(3):e290–e293
18. Shafran I, et al (2020) The medical scribe: corpus development and model performance analyses. In: Proceedings of the Twelfth Language Resources and Evaluation Conference
19. Amazon Comprehend Medical. [cited 2024 06/21]; Available from: https://aws.amazon.com/comprehend/medical/
20. Tran BD et al (2023) "Mm-hm","Uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? J Am Med Inform Assoc 30(4):703–711
21. Biswas A, Talukdar W (2024) Intelligent clinical documentation: harnessing generative AI for patient-centric clinical note generation. arXiv preprint arXiv:2405.18346
22. Google speech to text. [cited 2024 06/21]; Available from: https://cloud.google.com/speech-to-text?hl=en.
23. Introducing Whisper - OpenAI. [cited 2024 06/21]; Available from: https://openai.com/index/whisper.
24. Azure speech to text. [cited 2024 06/21]; Available from: https://azure.microsoft.com/en-us/products/ai-services/speech-to-text.
25. Mani A, Palaskar S, Konam S (2020). Towards understanding ASR error correction for medical conversations. In: Proceedings of the first workshop on natural language processing for medical conversations
26. Zielonka M, et al (2023) A survey of automatic speech recognition deep models performance for Polish medical terms. In: 2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA). IEEE
27. Tran BD, et al (2022) Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician

conversations. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association
28. Schultz BG et al (2021) Automatic speech recognition in neurodegenerative disease. Int J Speech Technol 24(3):771–779
29. Kothari K et al (2021) Effect of repetitive simulation training on emergency medical services team performance in simulated pediatric medical emergencies. AEM Educ Train 5(3):e10537
30. Zhang Z et al (2021) Data work and decision making in emergency medical services: a distributed cognition perspective. Proc ACM Hum Comput Interact 5(CSCW2):1–32
31. Ozkaynak M, et al (2020) Simulating teamwork for better decision making in pediatric emergency medical services. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association
32. NEMSIS Data Dictionary EMS Data Standard. [cited 2024 06/21]; Available from: https://nemsis.org/media/nemsis_v3/release-3.5.0/DataDictionary/PDFHTML/EMSDEMSTATE/index.html.
33. Bosker HR (2021) Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies. Behav Res Methods 53(5):1945–1953
34. Wu Y, et al (2017) Clinical named entity recognition using deep learning models. In: AMIA annual symposium proceedings. American Medical Informatics Association
35. Luo X, et al (2021) A deep language model for symptom extraction from clinical text and its application to extract COVID-19 symptoms from social media. IEEE J Biomed Health Inf
36. Navarro DF, et al (2023) Clinical named entity recognition and relation extraction using natural language processing of medical free text: a systematic review. Int J Med Inf: 105122
37. Ali A, Renals S (2018) Word error rate estimation for speech recognition: e-WER. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).
38. Enarvi S, et al (2020) 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
39. Lacson RC, Barzilay R, Long WJ (2006) Automatic analysis of medical dialogue in the home hemodialysis domain: structure induction and summarization. J Biomed Inform 39(5):541–555
40. Joshi A, et al (2020) Dr. summarize: global summarization of medical dialogue by exploiting local structures. In: Findings of the association for computational linguistics: EMNLP 2020.
41. Miner AS et al (2020) Assessing the accuracy of automatic speech recognition for psychotherapy. NPJ Digit Med 3(1):82
42. Du N, et al (2019) Extracting symptoms and their status from clinical conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
43. Kodish-Wachs J, et al (2018) A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association
44. Achiam J, et al (2023) Gpt-4 technical report. arXiv preprint arXiv:2303.08774
45. Xie Q, et al (2024) Me LLaMA: foundation large language models for medical applications. arXiv preprint arXiv:2402.12749
46. Gundabathula SK, Kolar SR (2024) PromptMind team at MEDIQA-CORR 2024: improving clinical text correction with error categorization and LLM ensembles. arXiv preprint arXiv:2405.08373
47. Abacha AB, et al (2024) Overview of the mediqa-corr 2024 shared task on medical error detection and correction. In: Proceedings of the 6th Clinical Natural Language Processing Workshop
48. Quiroz JC et al (2019) Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ Digit Med 2(1):114
49. Lin SY, Shanafelt TD, Asch SM (2018) Reimagining clinical documentation with artificial intelligence. In: Mayo Clinic Proceedings. Elsevier
50. Yu P, et al (2023) Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. In: Healthcare. MDPI
51. Ong JCL et al (2024) Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 6(6):e428–e432
52. Gerke S, Minssen T, Cohen G (2020) Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial intelligence in healthcare. Elsevier, pp 295–336
53. Mohr DN et al (2003) Speech recognition as a transcription aid: a randomized comparison with standard transcription. J Am Med Inform Assoc 10(1):85–93

54. Zhang Z, et al (2022) Hands-free electronic documentation in emergency care work through smart glasses. In: International Conference on Information. Springer