There's not an APP for That: Comparing Interpretation Through an AI Voice App and Qualified Medical Interpreters in Real World Clinical Settings

Iris Feinberg *, Heewon Lee-Laminack, Elizabeth L. Tighe, Ifedola Owoeye

Posted Date: 25 March 2026

doi: 10.20944/preprints202603.2013.v1

Keywords: qualified medical interpreter; AI-voice assisted app; interpretation; limited English proficiency; health literacy; language access; language barriers

 doi:10.20944/preprints202603.2013.v1

Article

There's not an APP for That: Comparing Interpretation Through an AI Voice App and Qualified Medical Interpreters in Real World Clinical Settings

Iris Feinberg [1,*], Heewon Lee-Laminack [1], Elizabeth L. Tighe [2] and Ifedola Owoeye [1]

[1] Adult Literacy Research Center, Department of Learning Sciences, Georgia State University

[2] Department of Psychology, Georgia State University

* Correspondence: ifeinberg2@gsu.edu

Highlights

What are the main findings?

- AI voice apps had significantly higher linguistic and clinical error rates than qualified medical interpreters (33.3% vs. 4.8%).
- Error rates were worse for less commonly spoken languages, amplifying equity concerns.

What are the implications of the main findings?

- Qualified in-person interpreters remain essential for safe, accurate clinical communication with linguistically minority patients.
- Hybrid models blending professional interpretation with AI technology may balance access and safety but need organizational oversight.

Abstract

Background/Objectives: AI voice interpretation applications are increasingly used in clinical settings to address language access challenges, yet evidence comparing their performance to qualified in-person medical interpreters in authentic clinical encounters remains limited. This study compares the linguistic and clinical accuracy of AI-based voice interpretation and certified in-person medical interpretation using recorded real-world clinical encounters. **Methods:** Outpatient clinical encounters involving patients with limited English proficiency were audio-recorded. Fourteen physician speech segments (mean length 78.5 words) representing common diagnostic, treatment, and counseling content were extracted and translated into seven languages using both an AI voice interpretation application and certified in-person medical interpreters. Bilingual reviewers and a clinician evaluated translations for accuracy, completeness, and clinical fidelity. Qualitative analyses examined error patterns and contextual loss; quantitative comparisons assessed error rate differences across languages and interpretation conditions. **Results:** AI voice app translations exhibited significantly higher linguistic errors ($\chi 2[1] = 19.78$, $p < .001$) and clinical accuracy errors ($\chi 2[1] = 45.07$, $p < .001$) than qualified medical interpreter translations. Clinical error rates were 33.3% for AI-generated versus 4.8% for interpreter-generated translations. Error rates were also higher for less commonly spoken languages compared to commonly spoken languages when using the AI voice app (42.9% vs. 14.3%). **Conclusions:** Qualified in-person interpreters remain essential for safe, accurate clinical communication. Hybrid models integrating professional interpretation with appropriately deployed AI technology may offer a balanced approach to expanding language access while maintaining communication safety and equity.

 doi:10.20944/preprints202603.2013.v1

Keywords: qualified medical interpreter; AI-voice assisted app; interpretation; limited English proficiency; health literacy; language access; language barriers

1. Introduction

Language barriers remain a persistent and well-documented challenge in healthcare systems worldwide for patients with limited English proficiency (LEP). Inadequate communication between clinicians and patients with LEP is consistently associated with lower comprehension of diagnoses and treatment plans, reduced adherence to medical recommendations, increased risk of medical error, and poorer overall health outcomes [1–3]. Communication difficulties can compromise shared decision-making, limit patients' ability to express symptoms accurately, and undermine trust in healthcare providers. As healthcare delivery becomes increasingly complex and technologically mediated, the consequences of ineffective communication for LEP populations have become more pronounced, reinforcing language access as a core component of healthcare quality and equity. Recent studies further indicate that language discordance contributes to disparities in hospitalization rates, emergency department utilization, and patient-reported experiences of care [4–7]. These patterns highlight the central role of language access in advancing health outcomes and health equity.

Professional medical interpreters have been shown to substantially improve communication quality, patient safety, and patient satisfaction in clinical encounters involving patients with LEP [4]. Interpreters facilitate accurate information exchange, promote patient engagement, and reduce misunderstandings that can lead to confusion about treatment or delayed care. Empirical studies indicate that interpreter-mediated visits are associated with higher rates of preventive service use, better chronic disease management, and improved adherence to prescribed therapies [1]. Despite this evidence, access to qualified interpreters remains uneven across healthcare settings. Rural hospitals, community clinics, and safety-net systems frequently report limited interpreter availability, particularly for languages spoken by smaller immigrant or refugee populations. Structural constraints—including financial limitations, workforce shortages, scheduling challenges, and inadequate reimbursement mechanisms—continue to hinder consistent provision of interpreter services [1]. Workforce pipeline challenges and limited certification infrastructure further constrain interpreter availability in underserved regions [8,9]. These systemic barriers create persistent gaps that perpetuate reliance on ad hoc or informal interpretation practices that compromise quality and safety.

In response to these persistent gaps in language access, digital and artificial intelligence (AI)–based language technologies have increasingly been introduced into clinical environments. AI voice interpretation systems, mobile translation applications, and machine translation platforms offer rapid, on-demand access to multilingual communication support and are often promoted as scalable and cost-efficient alternatives to traditional interpreter services [10]. These technologies have been adopted in outpatient clinics, emergency departments, inpatient units, and telehealth settings, particularly in resource-constrained environments. Technology-mediated communication is considered a potential integral component of contemporary language access strategies. Healthcare systems increasingly incorporate these tools into electronic health record platforms and patient portals to support routine interactions [11]. However, institutional adoption often occurs without comprehensive evaluation of translation and clinical accuracy.

While AI voice-assisted technologies hold promise for expanding access to language services, their widespread implementation has outpaced the development of rigorous evidence regarding their safety, accuracy, and impact on communication quality [12]. Many healthcare organizations have adopted commercial translation platforms without systematic evaluation of their clinical appropriateness or limitations [13–15]. Vendors frequently emphasize speed, convenience, and cost savings, while less attention is given to error rates, contextual understanding, and performance variability across clinical scenarios. This imbalance raises concerns about the extent to which technological solutions are being substituted for, rather than complementing, professionally trained

 doi:10.20944/preprints202603.2013.v1

interpreters. Health technology assessments have noted that few translation systems undergo independent validation prior to clinical deployment [9]. Consequently, decision-making is often driven by operational pressures rather than evidence-based quality benchmarks.

Existing research on technology-assisted interpretation has produced mixed findings. Studies evaluating machine translation and mobile translation applications suggest that these tools can achieve acceptable performance for simple or standardized content in highly used languages, such as Spanish and Vietnamese [12–16]. For routine instructions, appointment scheduling, or basic symptom descriptions, automated systems may provide reasonably accurate translations. However, accuracy declines substantially for complex medical terminology, nuanced counseling, emotionally sensitive discussions, and culturally embedded communication [10]. Subtle shifts in tone, modality, or contextual meaning may be lost, potentially altering clinical intent. These limitations restrict the appropriateness of automated tools in high-stakes or ethically sensitive encounters.

Performance disparities are particularly pronounced for less commonly spoken languages, reflecting limited training data and algorithmic bias within AI systems [17]. Languages with fewer digital resources such as Amharic, Arabic, Burmese, and Swahili are often underrepresented in training datasets, leading to higher error rates and inconsistent output quality. These limitations raise serious concerns about the potential for clinically meaningful misinterpretations, omissions, and distortions when AI tools are used without professional oversight [12]. Even minor translation errors can have significant consequences in contexts involving medication dosing, informed consent, or discharge planning. Studies have shown that error rates increase substantially for indigenous, refugee, and regional languages [18]. Such disparities risk reinforcing existing structural inequities in healthcare delivery [19,20].

Liability for communication errors resulting from automated translation remains poorly defined. Clinicians and healthcare organizations face uncertainty regarding responsibility when inaccurate translations contribute to adverse outcomes [21]. Regulatory frameworks have not kept pace with technological developments, leaving gaps in oversight and enforcement [21–24]. Patients with LEP are disproportionately affected by these risks, as they have limited capacity to independently verify or contest erroneous translations. Without clear governance structures, transparency standards, and accountability mechanisms, the integration of AI interpretation tools may inadvertently exacerbate existing inequities in healthcare access and quality [22]. Legal scholars have emphasized the need for clearer malpractice standards in technology-mediated communication [11].

In contrast, qualified medical interpreters are trained to convey not only linguistic meaning but also contextual, cultural, and relational aspects of communication. Qualified interpreters receive instruction in medical terminology, ethical principles, and standardized interpretation techniques designed to preserve clinical intent and minimize distortion [25,26]. They are skilled in managing ambiguity, clarifying misunderstandings, and facilitating bidirectional dialogue between patients and providers. These competencies enable interpreters to adapt communication strategies to individual patient needs and preferences [9]. Such relational skills remain difficult to replicate through automated systems. Prior research demonstrates that interpreter-mediated encounters are associated with improved patient comprehension, reduced disparities in care, and lower rates of adverse events [1]. Nevertheless, interpreter services are not always available at the point of care, particularly for rare languages, after-hours visits, or emergency situations. Studies indicate that delays in interpreter access are associated with longer visit times and deferred clinical decision-making [7].

Emerging implementation research suggests that the effectiveness of language technologies depends heavily on organizational context, provider training, and integration with existing clinical workflows [11,13,27]. Studies indicate that clinicians often lack formal guidance on when and how to appropriately use automated translation tools, resulting in inconsistent practices and potential overreliance in high-risk encounters [11,13]. Time pressures, staffing constraints, and limited awareness of best practices may encourage providers to use digital tools as default solutions.

 doi:10.20944/preprints202603.2013.v1

Variability in institutional policies further contributes to fragmented implementation patterns [9]. Without standardized protocols, quality assurance and clinical accuracy remains difficult to achieve. Models that combine technological support with structured training, interpreter collaboration, and institutional oversight demonstrate greater potential to enhance communication while preserving safety and trust [11]. Organizational leadership, reimbursement policies, and information technology infrastructure also influence successful implementation. Interdisciplinary governance committees have been shown to improve accountability and alignment across departments [13]. These models promote responsible innovation while safeguarding clinical outcomes.

Translation errors in clinical encounters involving patients with limited English proficiency (LEP) are not merely linguistic inaccuracies; they carry substantial clinical significance with measurable implications for patient safety, quality of care, and health outcomes. Empirical studies have demonstrated that interpretation errors—particularly omissions, additions, substitutions, and distortions—can alter diagnostic information, medication instructions, informed consent discussions, and treatment planning in ways that directly affect clinical decision-making [2,4]. Errors of clinical consequence have been shown to occur more frequently when ad hoc interpreters, including untrained staff or family members, are used compared with professional medical interpreters [2]. Such errors have been associated with increased risk of adverse events, misunderstanding of discharge instructions, reduced adherence to treatment, and decreased patient satisfaction [1]. Moreover, even seemingly minor inaccuracies—such as incorrect medication dosing, failure to convey symptom duration, or omission of qualifiers (e.g., "intermittent" vs. "constant")—can lead to misdiagnosis, inappropriate testing, or unsafe prescribing. Systematic reviews indicate that professional interpreter services are associated with improved comprehension, safer care processes, and, in some cases, reduced health disparities among LEP populations [1,4]. As healthcare systems increasingly integrate digital and AI-based translation tools, concerns remain regarding variable accuracy for complex medical terminology and culturally nuanced communication, underscoring the continued importance of clinically trained human interpreters in high-stakes interactions. Collectively, the evidence suggests that translation accuracy is not a peripheral quality issue but a core determinant of clinical safety and equity for linguistically diverse patients.

Despite growing reliance on AI-based interpretation, direct comparisons between these tools and qualified in-person interpreters in real-world clinical encounters remain limited [10,28]. Much of the existing literature relies on simulated interactions, written translation tasks, or narrowly defined clinical scenarios. Such designs may not adequately capture the complexity of authentic patient–provider communication. Contextual factors such as time pressure, emotional distress, and competing clinical priorities are rarely incorporated into experimental designs [29]. As a result, current evidence may overestimate real-world performance. This study explores the translation accuracy of AI-based voice interpretation and qualified in-person medical interpretation using 14 recorded real-world clinical encounters in six languages (Amharic, Arabic, Burmese, Spanish, Swahili, Vietnamese) which we grouped as common (Spanish and Vietnamese) and less common (Amharic, Arabic, Burmese, Swahili). Specifically, we sought to describe not only translation accuracy between qualified translators and AI-assisted voice technologies, but if the errors displayed clinical significance that could impact patient care and health outcomes.

To address these aims, this study examined translation accuracy across both modality and language grouping. First, it compared APP and QMI translations in terms of overall error rates. Second, it examined whether error patterns differ between common and less common languages within each translation condition. Finally, it assessed the likelihood of clinically accurate translations across language groups for both APP and QMI conditions. Here were the research questions for this study:

1. Is there a significant difference in linguistic or clinical error rates between APP and QMI translations?
2. Within each condition, is there a significant difference in linguistic or clinical error rates between common and less common languages?

 doi:10.20944/preprints202603.2013.v1

3. What are the odds of APP translations being clinically accurate in common versus less common languages?
4. What are the odds of QMI translations being clinically accurate in common versus less common languages?

2. Materials and Methods

2.1. Sample

Fourteen English-speaking physician speech segments were sampled from an existing corpus of audio-recorded outpatient clinical encounters [30]. Selected segments represented common diagnostic, treatment, and counseling content and contained medically relevant terminology. Sentences were chosen to reflect realistic patient-facing medical comments and instructions in outpatient settings. The mean length was 78.5 words per segment (range: 33–140 words), representing a range of common clinical communication contexts including diagnostic explanation, treatment planning, and patient counseling. See Table 1 for a sample of selected speech segments. All source audio recordings were de-identified prior to sentence extraction. No patient-identifying information was included in translated materials. The study protocol adhered to ethical standards for secondary data analysis of clinical communication.

Table 1. Sample of Physician Speech Segments.

Physician Speech Segments
Okay so what I'm going to do now is I'm going to listen to your heart and lungs okay because you're a new patient and then I really just want to do an exam on your ah muscles okay. The musculoskeletal exam and I want to look at your left leg a little bit closely okay.
it could be ah a cyst which I'll see if you have any cysts sometimes it's called a baker's cyst in the back of your knee which could cause pain ah like I said you got that study, the DVT so that's good you don't have any clots and then ah it could be other things like ah of course we want to rule out any signs of infection which I don't think it is since it's been happening for 6 months and then just other signs of Edema cause this burning pain usually if you have a burning pain it can be due to swelling and you're saying it feels swollen so that could explain that and then we could talk about if we need to do any blood work or anything like that and possible treatments as well.
But yeah if your doing fine at home then probably we don't need to adjust your medications too much. But, uh, again, if at home your blood pressure is running like 160, 170s, we do need to increase the dose of the lisinopril. As long as we can control your blood pressure, and can control your (.) and control your lipid, that can help to prevent other the more severe =events down the road.
So you know how to recognize the signs of hypoglycemia. Like low blood sugars, say you are sweating a lot, you are having a period of dizziness , those are the signs of low blood sugar and you may need to take a look at your meter and see what is the blood sugar level. If it's lower than 80, you may want to get something to combat your blood sugar a lit bit. Otherwise, it's also dangerous. Just like DKA is dangerous, and low blood sugar is also dangerous. Does that make sense?
Okay. Yeah, according to the guideline, because you have type 1 diabetes for more than 5 years is recommended to uh check the microalbumin to see hows your kidney function annually. If its, say if it's damaged we can add medication to you to help you to prevent further damage of your kidneys because we know that high blood pressure and diabetes those are two most common causes of the kidney damages in the United States. Have you ever heard about that?

 doi:10.20944/preprints202603.2013.v1

2.2. Study Design

This study employed a comparative accuracy design to evaluate translation accuracy and clinical accuracy of speech segments translated by AI-assisted voice translation applications and by qualified medical interpreters (QMIs). A within-item paired design was used so that each English sentence served as its own control across translation conditions. Accuracy outcomes were analyzed using paired statistical tests appropriate for matched binary data. Each sentence was translated into six target languages: Amharic, Arabic, Burmese, Spanish, Swahili, and Vietnamese. These languages were selected based on prevalence within local clinical populations and availability of local QMIs.

Each of the 14 sentences was translated under two primary conditions:

- Condition 1: AI-assisted Voice Translation Applications (App Condition)

Translations were produced using the AI-based voice translation application Google Translate. Sentences were spoken into Google Translate by one of the study investigators. The translated output in the target language was recorded verbatim.

- Condition 2: Qualified Medical Interpreters (QMI Condition)

Qualified in-person medical interpreters fluent in the respective target language translated the same 14 sentences. Interpreters were credentialed and experienced in clinical medical interpretation.

To assess translation accuracy while minimizing rater bias from direct source comparison, each translated sentence (App and QMI conditions) underwent blind back-translation into English by professional translators who had not participated in the original translation. Back-translators were blinded to the original English sentence.

2.3. Measures

Back-translated English versions were systematically compared with the original English source sentences for linguistic accuracy. Two bilingual researchers independently reviewed each back-translated segment alongside its original English counterpart. Using a predefined coding guide with operational definitions and examples, each segment was coded as:

1. Correct
2. Partially wrong (minor inaccuracies, omissions, or distortions that did not fundamentally alter clinical meaning)
3. Totally wrong (substantive inaccuracies, omissions, or distortions)

The bilingual reviewers assessed translations for their fidelity to original medical meaning, completeness of content, and linguistic clarity. The study PI reviewed the segments to ascertain fidelity to the coding guide; researcher analysis was 100% accurate.

In the second stage, a clinician with 25 years of clinical practice experience reviewed the partially or totally wrong sentence segments to assess potential impact on clinical interpretation, patient understanding, or patient safety, comparing original English to back-translated English for both APP and QMI conditions. The clinician was blinded to condition. The clinician evaluated whether identified discrepancies met the threshold for clinical significance. Specifically, segments were coded as:

1 = clinically wrong (presence of medically meaningful inaccuracy, distortion, omission, or unsafe instruction)

0 = not clinically wrong

3. Results

All tabulated error rates are presented in Table 2 and separated by translation condition (APP and QMI), language (individual and by common and less common groups), and linguistic and clinical error rates. Below we describe the analyses that could be computed based on the error rates, please note that low error rates, particularly in the clinically wrong for the QMI conditions precluded us running some of the analyses or may have lead to more unstable estimates even with corrections for

 doi:10.20944/preprints202603.2013.v1

low counts. These descriptives in Table 2 are crucial for illustrating differences in error rates between the two translation conditions.

Table 2. Linguistic and Clinical accuracy of APP and QMI Translations by language (n=14 per language per condition.

LANGUAGE	APP				QMI			
	Linguistically Wrong		Clinically Wrong		Linguistically Wrong		Clinically Wrong	
	n	%	n	%	n	%	n	%
Commonly Spoken								
Spanish	3	21.4%	2	14.3%	0	0.0%	0	0.0%
Vietnamese	3	21.4%	2	14.3%	2	14.3%	1	7.1%
Sub-Total	6	21.4%	4	14.3%	2	7.1%	1	3.6%
Less Commonly Spoken								
Amharic	6	42.9%	4	28.6%	9	64.3%	3	21.4%
Arabic	7	50.0%	5	35.7%	0	0.0%	0	0.0%
Burmese	4	28.6%	8	57.1%	3	21.4%	0	0.0%
Swahili	5	35.7%	7	50.0%	3	21.4%	0	0.0%
Sub-Total	22	39.3%	24	42.9%	15	26.8%	3	5.4%
TOTAL	28	33.3%	28	33.3%	17	20.2%	4	4.8%

To assess linguistic error rates, we used a McNemar's chi-square test to compare error rates on the same 14 sentences between the APP and QMI conditions. The analysis revealed a significant difference ($\chi^2[1] = 19.78$, $p < .001$, $OR = 3.29$, 95% CIs = 1.91-5.67), suggesting an overall higher linguistic error rate, increased odds of 3.29 for APP translations (see Table 2 for error rates by translation condition).

We used a chi-square test to compare proportions of linguistic errors among independent language groups (commonly spoken [Spanish, Vietnamese] vs. less commonly spoken [Amharic, Arabic, Burmese, Swahili]) within the APP translation. This analysis revealed a non-significant difference in proportions of errors among language groups using APP ($\chi^2[1] = 2.68$, $p = .101$, $OR = 2.37$, 95% CIs = 0.83-6.78; see Table 2 for all error rates by language group). Linguistic error rates between common and less commonly spoken languages were similar when using the APP. We did not run a statistical significance test comparing language groups within the QMI condition because the error rates were very low for the commonly spoken languages (7.3%).

We used a McNemar's chi-square test to compare linguistic error rates on the same 14 sentences between APP and human interpreters by language group. For less common languages, the analysis revealed a significant difference ($\chi^2[1] = 7.37$, $p = .009$, $OR = 2.27$, 95% CIs = 1.23-4.16, suggesting an overall higher linguistic error rate, increased odds of 2.27 for APP translations (see Table 2 for all error rates). The odds of APP generating linguistic error rates are 2.27 times more than a QMI for less commonly spoken languages. For common languages, the analysis revealed a significant difference ($\chi^2[1] = 16.67$, $p < .001$, $OR = 11.00$, 95% CIs = 2.59-46.78), suggesting an overall higher linguistic error rate, increased odds of 11.00 for APP translations; however caution is warranted given the low error rate (7.3%) for commonly spoken languages in the QMI condition.

To assess clinical error rates, we used a McNemar's chi-square test to compare error rates on the same 14 sentences between the APP and QMI conditions for clinical correctness. The analysis revealed a significant difference ($\chi^2[1] = 45.07$, $p < .001$, $OR = 14.00$, 95% CIs = 5.08-38.61), suggesting an overall higher clinical error rate, increased odds of 14.00 for APP translations. Although significant, caution is warranted given the wide CI, which can suggest uncertainty in the estimated effect and is likely due to the lower number of error rates in the QMI condition. Specifically, human

 doi:10.20944/preprints202603.2013.v1

interpreter generated 4.8% clinically wrong phrases; the APP generated 33.3% clinically wrong phrases over the same sentence segment sample (Table 2).

We used a Fisher's Exact Test to compare proportions of clinical errors among independent language groups (commonly spoken [Spanish, Vietnamese] vs. less commonly spoken [Amharic, Arabic, Burmese, Swahili]) because of the small (<5) error rates in the some of the groups. For APP, there was a significant difference in error rates (p = .013, OR = 4.10, 95% CIs = 1.32-12.75), suggesting a 4.10 increased odds of errors for the less common language group using APP translation. Although significant, caution is warranted given the wide CI, which can suggest uncertainty in the estimated effect and may also be due to the sample size imbalance among language groups (commonly spoken n = 28 sentences; less commonly spoken n = 56 sentences; see Table 2 for all error rates by language group).

We did not run statistical analysis tests within QMI condition or comparing language groups between translation conditions because of the very small error rates in the QMI condition (less common = 4.8%; more common = 3.6%).

4. Discussion

Language access remains a foundational determinant of healthcare quality, safety, and equity for patients with limited English proficiency. The findings of this study reinforce longstanding evidence that communication barriers contribute to diagnostic errors, compromised treatment adherence, and diminished patient engagement [1,6]. Although healthcare systems have increasingly turned to technological solutions to address interpreter shortages, this study demonstrates that such approaches must be carefully evaluated within real-world clinical contexts. The high prevalence of linguistically incorrect and clinically significant errors observed in AI-assisted translations underscores the persistent risks associated with substituting automated tools for human expertise. These findings align with prior research cautioning against uncritical adoption of digital language technologies in high-stakes medical environments [9,11].

The comparative results of this study highlight meaningful differences in accuracy between qualified in-person interpreters and AI-assisted voice applications. While interpreter-mediated encounters were not error-free, they demonstrated substantially lower rates of linguistically and clinically significant errors than automated systems. This pattern is consistent with existing literature documenting the capacity of trained interpreters to preserve clinical intent, manage ambiguity, and facilitate culturally responsive communication [1]. In contrast, the disproportionately high error rates associated with AI-based tools—particularly across less commonly spoken languages—reflect ongoing limitations in algorithmic training, contextual understanding, and linguistic representation [28]. Consistent with these limitations, the present study found that AI-based interpretations frequently involved omissions, mistranslations of medical terminology and medication names, altered sentence meanings, and, at times, severe distortions that conveyed unintended or opposite meanings. These disparities raise concerns about the potential for automated systems to exacerbate existing inequities in healthcare delivery.

The clinical significance of translation errors identified in this study further emphasizes the ethical and safety implications of technology-mediated interpretation. Nearly half of all incorrect translations were judged to have meaningful implications for patient care, with the majority originating from AI-assisted systems. Errors affecting medication instructions, symptom descriptions, or treatment planning may result in adverse events, delayed care, or compromised informed consent [21]. Patients with LEP are uniquely vulnerable to these risks, as they often lack alternative mechanisms to verify the accuracy of clinical information. Without appropriate safeguards, reliance on automated interpretation may unintentionally undermine patient autonomy and trust.

These findings also underscore the importance of organizational context and governance in shaping the effectiveness of language access strategies. Prior research demonstrates that the impact of digital tools depends heavily on institutional policies, provider training, and integration with

 doi:10.20944/preprints202603.2013.v1

professional interpreter services [11,31]. In the absence of standardized protocols, clinicians may default to automated solutions even in complex or sensitive encounters, increasing the likelihood of clinically significant errors. Health systems must therefore prioritize evidence-based guidelines that delineate appropriate use cases for technology, reinforce interpreter engagement, and establish mechanisms for quality assurance and continuous monitoring [9].

From a policy and regulatory perspective, this study highlights the need for clearer standards governing the deployment of AI-based translation tools in healthcare. Current regulatory frameworks offer limited guidance regarding validation, liability, and accountability for technology-mediated communication [21,22]. Establishing minimum performance benchmarks, independent evaluation requirements, and transparent reporting mechanisms is essential to ensure patient safety. Reimbursement structures and workforce development initiatives must also be strengthened to expand access to qualified medical interpreters, particularly for underserved and linguistically diverse populations [9]. Such reforms are necessary to prevent cost-saving imperatives from superseding clinical quality.

Future research should build on the findings of this study by conducting larger, multi-site investigations that incorporate diverse clinical settings, languages, and patient populations. Longitudinal designs and mixed-method approaches may provide deeper insight into how communication errors affect health outcomes, patient experiences, and healthcare utilization over time [10,28]. Comparative effectiveness studies examining hybrid models that combine AI tools with professional interpreter oversight are also warranted. Ultimately, while technological innovations may enhance access in limited circumstances, they cannot replace the relational, ethical, and contextual competencies of trained interpreters. A balanced, evidence-based approach that integrates technology as a supportive—not substitutive—resource offers the most promising path toward equitable and safe communication for patients with limited English proficiency.

Limitations

This study has several limitations that should be considered when interpreting its findings. First, the sample size was relatively small, consisting of 14 recorded clinical encounters across six languages. While the use of authentic, real-world interactions strengthens ecological validity, the limited number of cases restricts the generalizability of the results. The findings may not fully represent the wide range of clinical contexts, provider communication styles, patient characteristics, and linguistic variations present in broader healthcare settings. Larger, multi-site studies are needed to confirm the observed patterns across diverse populations and institutional environments.

Second, the study focused on a selected set of languages, including Amharic, Arabic, Burmese, Spanish, Swahili, and Vietnamese. Although these languages reflect important immigrant and refugee populations in our catchment area, they do not capture the full spectrum of linguistic diversity encountered in healthcare systems. Performance may differ for other commonly spoken or regionally specific languages, dialects, and lesser spoken languages. Additionally, variations in accent, speech rate, and colloquial expression were not systematically analyzed, which may have influenced translation accuracy for both human and automated interpreters.

Third, translation accuracy and clinical significance were assessed through expert review, which necessarily involves a degree of bias. Although standardized criteria were used to evaluate partial and clinically significant errors, interpretations of clinical impact may vary among reviewers. Differences in professional background and clinical specialty could influence assessments. The absence of formal interrater reliability testing further limits the ability to quantify agreement among evaluators and may introduce measurement bias.

Fourth, this study evaluated a limited number of AI-assisted voice translation applications and did not examine variability across different platforms, software versions, or system updates. Performance may differ substantially among commercial products and may change over time as algorithms are refined. Consequently, the results should not be interpreted as representative of all AI-based interpretation technologies. Similarly, interpreter performance may vary based on

 doi:10.20944/preprints202603.2013.v1

certification level, experience, fatigue, and familiarity with specific clinical domains, factors that were not systematically controlled in this analysis.

Finally, contextual and environmental factors that may influence communication quality were not systematically measured. Variables such as time pressure, background noise, clinician workload, emotional distress, and technological infrastructure may affect both interpreter performance and AI accuracy. In addition, the study design did not evaluate hybrid models combining AI tools with professional oversight, which may represent an important emerging approach to language access. Future research should incorporate more comprehensive contextual measures and experimental designs to better understand how organizational and situational factors shape the effectiveness and safety of language interpretation strategies.

Author Contributions: Conceptualization, Iris Feinberg and Heewon Lee-Laminack; Methodology, Iris Feinberg and Heewon Lee-Laminack; Validation, Elizabeth L. Tighe; Formal analysis, Elizabeth L. Tighe; Investigation, Iris Feinberg and Heewon Lee-Laminack; Data curation, Heewon Lee-Laminack and Ifedola Owoeye; Writing – original draft, Iris Feinberg; Writing – review & editing, Heewon Lee-Laminack, Elizabeth L. Tighe and Ifedola Owoeye; Supervision, Iris Feinberg; Project administration, Iris Feinberg. All authors have read and agreed to the published version of the manuscript.

Funding Information: This research received no external funding.

Institutional Review Board Statement: The ethical approval is not required for this study as it utilized data derived from a previously conducted study [30]. No direct contact with participants was involved in the present research.

Informed Consent Statement: Informed consent was not required for this study due to its retrospective design and the exclusive use of data obtained from a previously conducted study [30], with no direct participant involvement.

Data Availability Statement:.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Al Shamsi, H.; Almutairi, A.G.; Al Mashrafi, S.; Al Kalbani, T. Implications of language barriers for healthcare: A systematic review. *Oman Med. J.* **2020**, *35*, e122. https://doi.org/10.5001/omj.2020.40
2. Flores, G.; Laws, M.B.; Mayo, S.J.; Zuckerman, B.; Abreu, M.; Medina, L.; Hardt, E.J. Errors in medical interpretation and their clinical consequences. *Pediatrics* **2003**, *111*(1), 6–14. https://doi.org/10.1542/peds.111.1.6
3. Lindholm, M.; Hargraves, J.L.; Ferguson, W.J.; Reed, G. Professional language interpretation and inpatient length of stay and readmission rates. *J. Gen. Intern. Med.* **2012**, *27*, 1294–1299. https://doi.org/10.1007/s11606-012-2041-5
4. Karliner, L.S.; Jacobs, E.A.; Chen, A.H.; Mutha, S. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. *Health Serv. Res.* **2007**, *42*, 727–754. https://doi.org/10.1111/j.1475-6773.2006.00629.x
5. Escobedo, L.E.; Cervantes, L.; Havranek, E. Barriers in healthcare for Latinx patients with limited English proficiency: A narrative review. *J. Gen. Intern. Med.* **2023**, *38*(5), 1264–1271. https://doi.org/10.1007/s11606-022-07995-3
6. Schwei, R.J.; Hoang, L.; Wilson, P.; Greene, M.Z.; Lor, M.; Shah, M.N.; Pulia, M.S. Patient-centered care outcomes for patients in the emergency department with a non-English language preference: A scoping review. *Patient Educ. Couns.* **2023**, *114*, 107875. https://doi.org/10.1016/j.pec.2023.107875
7. Sudore, R.L.; Schillinger, D.; Katen, M.T.; Shi, Y.; Boscardin, W.J.; Osua, S.; Barnes, D.E. Engaging diverse English- and Spanish-speaking older adults in advance care planning: The PREPARE randomized clinical trial. *JAMA Intern. Med.* **2018**, *178*, 1616–1625.

 doi:10.20944/preprints202603.2013.v1

8. Feinberg, I.; Ogrodnick, M.; Zeidan, A. Building a qualified medical interpreter workforce program for lesser-spoken languages. *J. Health Care Poor Underserved* **2025**, *36* (3 Suppl.). https://doi.org/10.1353/hpu.2025.a967358
9. National Academies of Sciences, Engineering, and Medicine. *Integrating Health Literacy, Cultural Competence, and Language Access Services: Workshop Summary*. National Academies Press: Washington, DC, USA, 2016.
10. Panayiotou, A.; Hwang, K.; Williams, S.; Chong, T.W.H.; LoGiudice, D.; Haralambous, B.; Lin, X.; Zucchi, E.; Mascitti-Meuter, M.; Goh, A.M.Y.; You, E.; Batchelor, F. The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. *J. Clin. Nurs.* **2020**, *29(17-18), 3516–3526.* https://doi.org/10.1111/jocn.15390
11. Blease, C.; Kaptchuk, T.J.; Bernstein, M.H.; Mandl, K.D.; Halamka, J.D.; DesRoches, C.M. Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. *J. Med. Internet Res.* **2019**, *21*, e12802. https://doi.org/10.2196/12802
12. Khoong, E.C.; Steinbrook, E.; Brown, C.; Fernandez, A. Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA Intern. Med.* **2019**, *179*, 580–582. https://doi.org/10.1001/jamainternmed.2018.7653
13. Kreienbrinck, A.; Hanft-Robert, S; Mösko, M. Usability of technological tools to overcome language barriers in healthcare: A scoping review. *BMJ Open.* **2024**, *14*(3):e079814. https://doi.org/10.1136/bmjopen-2023-079814
14. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. https://doi.org/10.1056/NEJMra1814259
15. Wang, Y.; Patel, M.; Chen, L.; Roberts, K. Evaluating commercial medical translation platforms. *J. Am. Med. Inform. Assoc.* 2024, *31*, 389–398.
16. Brewster, R.C.; Tse, G.; Fan, A.L.; Elborki, M.; Newell, M.; Gonzalez, P.; Khan, A. Evaluating human-in-the-loop strategies for artificial intelligence-enabled translation of patient discharge instructions: A multidisciplinary analysis. *NPJ Digit Med* **2025**, *8*, 629. https://doi.org/10.1038/s41746-025-02055-6
17. Ta, R.; Turner Lee, N. How language gaps constrain generative AI development. Brookings Institution. Available online: https://www.brookings.edu/articles/how-language-gaps-constrain-generative-ai-development/ (accessed on 17 March 2026).
18. Mehandru, N.; Robertson, S.; Salehi, N. Reliable and safe use of machine translation in medical settings. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Baltimore, MD, USA, 21–24 June 2022; ACM: New York, NY, USA, 2022; pp. 1–10. https://doi.org/10.1145/3531146.353324
19. Birhane, A.; Prabhu, V.U.; Kahembwe, E. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, Virtual, **202**1. https://doi.org/10.48550/arXiv.2110.01963
20. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, Virtual, 2021; pp. 610–623. https://doi.org/10.1145/3442188.344592
21. Gerke, S.; Minssen, T.; Cohen, G. Ethical and legal challenges of AI-driven healthcare. *Camb. Q. Healthc. Ethics* 2020, *29*, 1–12. https://doi.org/10.1016/B978-0-12-818438-7.00012-5
22. Price, W.N.; Cohen, I.G. Privacy in the age of medical big data. *Nat. Med.* 2019, *25*, 37–43. https://doi.org/10.1038/s41591-018-0272-7
23. He, J.; Baxter, S.; Xu, J.; Xu, J.; Zhou, X.; Zhang, K. The practical implementation of AI technologies in medicine. *Nat. Med.* 2019, *25*, 30–36. https://doi.org/10.1038/s41591-018-0307-0
24. FDA. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan. U.S. Food and Drug Administration: Silver Spring, MD, USA, 2021. Available online: https://www.fda.gov/media/145022/download (accessed on 24 March 2026).
25. Hsieh, E. Bilingual Health Communication: Working with Interpreters. Routledge: New York, NY, USA, 2018.
26. National Council on Interpreting in Health Care (NCIHC). *National Standards of Practice for Interpreters in Health Care*. NCIHC: Washington, DC, USA, 2005. Available online:

 doi:10.20944/preprints202603.2013.v1

https://www.ncihc.org/assets/z2021Images/NCIHC%20National%20Standards%20of%20Practice.pdf (accessed on 24 March 2026).

27. Damschroder, L.J.; Reardon, C.M.; Widerquist, M.A.; Lowery, J. The updated CFIR framework. *Implement. Sci.* 2022, *17*, 75.
28. Wekenborg, M.K.; Gilbert, S.; Kather, J.N. Examining human-AI interaction in real-world healthcare beyond the laboratory. *npj Digit. Med.* 2025, *8*, 169. https://doi.org/10.1038/s41746-025-01559-5
29. Flores, G. The impact of medical interpreter services on the quality of health care: A systematic review. *Med. Care* 2005, *43*, 753–760. https://doi.org/10.1177/1077558705275416
30. Feinberg, I.; Ogrodnick, M.; Hendrick, R.; Bates, K.; Johnson, K.; Wang, B. Perception vs. reality in use of teach-back by medical residents. *Health Lit. Res. Pract.* 2019, *3*, e117–e126. https://doi.org/10.3928/24748307-20190501-01
31. Nápoles, A.M.; Santoyo-Olsson, J.; Karliner, L.S.; Gregorich, S.E.; Pérez-Stable, E.J. Inaccurate language interpretation and its clinical significance in the medical encounters of Spanish-speaking Latinos. *Med. Care* 2015, *53*, 940–947. https://doi.org/10.1097/MLR.0000000000000422