Review Article | Data Sciences

Artificial intelligence in clinical settings: a systematic review of its role in language translation and interpretation

Ariana Genovese[1]^, Sahar Borna[1]^, Cesar A. Gomez-Cabello[1]^, Syed Ali Haider[1]^, Srinivasagam Prabha[1]^, Antonio J. Forte[1,2]^, Benjamin R. Veenstra[3]

[1]Division of Plastic Surgery, Mayo Clinic, Jacksonville, FL, USA; [2]Center for Digital Health, Mayo Clinic, Rochester, MN, USA; [3]Division of Advanced Gastrointestinal and Bariatric Surgery, Mayo Clinic, Jacksonville, FL, USA

Contributions: (I) Conception and design: A Genovese, AJ Forte, BR Veenstra; (II) Administrative support: S Borna, CA Gomez-Cabello, SA Haider; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: A Genovese, S Borna, CA Gomez-Cabello, SA Haider, S Prabha; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Antonio J. Forte, MD, PhD. Division of Plastic Surgery, Mayo Clinic, 4500 San Pablo Rd, Jacksonville, FL 32224, USA; Center for Digital Health, Mayo Clinic, Rochester, MN, USA. Email: ajvforte@yahoo.com.br.

Background: Addressing language barriers through accurate interpretation is crucial for providing quality care and establishing trust. While the ability of artificial intelligence (AI) to translate medical documentation has been studied, its role for patient-provider communication is less explored. This review evaluates AI's effectiveness in clinical translation by assessing accuracy, usability, satisfaction, and feedback on its use.

Methods: A systematic search was conducted on July 11, 2024, across Cumulated Index in Nursing and Allied Health Literature (CINAHL), Institute of Electrical and Electronics Engineers (IEEE) Xplore, PubMed, Scopus, Web of Science, and Google Scholar. Inclusion criteria required AI to translate clinical information for a real or theoretical consultation. Exclusion criteria included reviews, correspondence, educational materials, non-peer-reviewed or retracted reports, non-English translations, pre-2016 publications, and reports on sign language or patient education. Search strings representing AI, language interpretation, and healthcare were used. Two investigators independently conducted the screening, extraction, synthesis of results, and bias assessments using Risk Of Bias In Non-randomized Studies - of Interventions (ROBINS-I), Mixed Methods Appraisal Tool (MMAT), and the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Qualitative Research. A third investigator resolved conflicts.

Results: Of 1,095 reports, 9 studies were analyzed, evaluating AI translation platforms Google Translate, Microsoft Translator, Apple iTranslate, AwezaMed, Pocketalk W, and the Asynchronous Telepsychiatry (ATP) App. Investigations occurred in the US, France, Switzerland, and South Africa, with publications from 2019–2024. AI medical translation shows promise, typically providing accurate translations for brief communications in limited languages, though human translation is often necessary. Accuracy scores ranged from 83–97.8% when translating from English, and 36–76% when translating to English. Usability scores were 76.7–96.7%. Patients were more satisfied than clinicians, with 84–96.6% and 53.8–86.7% satisfied, respectively. Clinicians were hesitant to use AI due to questions of respect, quality, reliability, and misunderstanding. AI is being used as a last-resort option, to assist fluent, non-certified providers and lay interpreters, and for brief communications.

Conclusions: Limitations include few languages tested, unidirectional translation, simulation, and evolving translation tools. AI shows promise in clinical translation, but the complexity of medical consultations requires a balanced approach combining AI and human translation services for quality care.

Keywords: Artificial intelligence (AI); machine intelligence; communication barrier; limited English proficiency

^ ORCID: Ariana Genovese, 0009-0000-9678-2163; Sahar Borna, 0000-0002-7845-7356; Cesar A. Gomez-Cabello, 0009-0008-0603-3192; Syed Ali Haider, 0009-0007-5621-2861; Srinivasagam Prabha, 0000-0003-2573-8493; Antonio J. Forte, 0000-0003-2004-7538.

Submitted Sep 12, 2024. Accepted for publication Dec 02, 2024. Published online Dec 17, 2024.
doi: 10.21037/atm-24-162
View this article at: https://dx.doi.org/10.21037/atm-24-162

Introduction

Background

As the world's "melting pot", a significant percentage of United States residents don't speak English as their primary language, evidenced by the 21.7% of individuals older than age 5 years that speak a language other than English at home (1). Furthermore, research shows that language barriers can significantly and negatively affect health and healthcare (2). Addressing the need for high-quality healthcare access in this population is challenged by potential language barriers, making accurate and efficient translation services key to providing equitable, patient-centered healthcare services to persons of all backgrounds (3). This obstruction to care expands beyond the United States, particularly in linguistically diverse regions of the world.

Precise language interpretation in the clinical domain is fundamental in preventing misunderstandings, preventable errors, adverse events, and harm to the patient (4). For this reason, many institutions utilize the services of those professionally trained in providing medical language interpretation (5). While these services can be offered in-person or remotely (6), translators are not always readily available (7), which poses a significant issue in the setting of a medical emergency where timely communication is vital (8). Additionally, it has been shown that the use of a language interpreter lengthens the interview time (9), but providers are typically given the standard amount of time for the consult (10). This has led to providers avoiding calling translation services, gauging the medical urgency of a situation versus the time it takes to call an interpreter, and at times, avoid using translation services (10). It was found that providers in multiple clinical settings used professional interpreters for less than 20% of patients who had limited English proficiency (11).

Professionally trained medical interpreters are not without error, as a study evaluating medical language interpretation recorded an average of 27 errors per encounter, 7.1% of which were either moderately or highly clinically significant (12). Additionally, human translators are costly for institutions (13) and remote services rely on adequate internet access for both parties across multiple locations, posing another barrier to high quality, efficient care.

A sector of computer science, artificial intelligence (AI) is the making of machines designed to perform activities that usually need human intelligence (14). Addressing the deficit in human availability and providing a more cost-effective solution that can yield rapid results, AI's role in medical language interpretation is currently being explored. Neural machine translation (NMT), a type of machine learning (ML), is a form of AI holding promise in this respect (15). Google Translate, an NMT platform, began utilizing AI in November of 2016 (16), and has gained regard in household and nonclinical settings. More recently, Google Translate's ability to provide translation services in the clinical field is

Highlight box

Key findings

- Artificial intelligence (AI) often provides sufficiently accurate translations for brief, simple communications when translating from English. When translating to English, non-European languages, and for complex discussions, inaccuracies become more common, necessitating human intervention. While AI was generally viewed as user-friendly, clinicians were less satisfied than patients. Despite these concerns, AI is currently being used in clinical settings when other translation options are unavailable, for short interactions, and to assist lay interpreters.

What is known and what is new?

- AI has been studied for language translation across various fields and has gained attention in healthcare for its potential. Research shows that AI can often provide acceptable translations for clinical documents such as discharge instructions and electronic health record data, though inaccuracies remain common.
- This manuscript contributes to the existing body of knowledge by focusing on AI's role in patient-provider communication, rather than documentation. It evaluates popular machine translation (MT) applications, such as Google Translate and Microsoft Translator, which have potential to improve the efficiency of healthcare visits and reduce costs for institutions.

What is the implication, and what should change now?

- Our research highlights that while AI can assist with language translation, it cannot replace human interpretation in its current state. However, as an additional tool, AI has the potential to enhance healthcare access and quality of care for limited English proficiency individuals. Clinicians should continue utilize the services of trained human interpreters, and if AI is utilized, a system must be used to assess accuracy and gauge understanding.

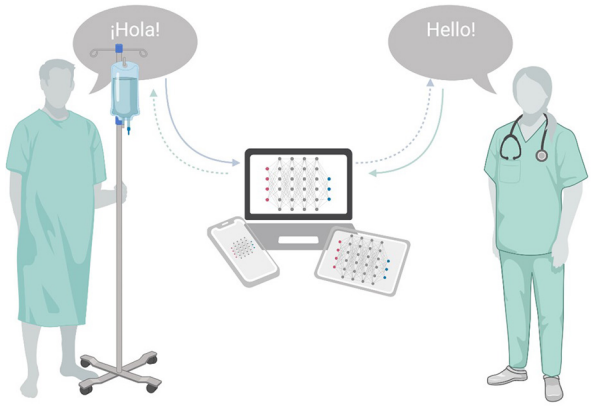

Figure 1 Application of machine learning for clinical language translation: a visual overview. The figure illustrates how artificial intelligence can facilitate real-time language translation during clinician-patient interactions. Machine learning algorithms in readily available devices, such as computers, tablets, and smartphones, can allow both parties to communicate despite language barriers. This technology has the potential to eliminate the need for human interpreters, optimizing visit efficiency and costs to healthcare institutions. This figure was generated using BioRender.

being studied (17).

Effective medical translation plays a key role in improving health outcomes and enhancing patient satisfaction (5). Clear communication between healthcare providers and patients is essential for the proper understanding of medical instructions, adherence to treatment regimens, and overall management of health conditions (18). Furthermore, it helps build trust between patients and providers (19), which in turn boosts patient compliance and satisfaction (20). Regarding health equity, ensuring that translation services are accessible helps guarantee that patients who do not speak English receive the same quality of care as those who speak English (21), thereby addressing disparities in healthcare access and outcomes. Therefore, the exploration of advanced translation technologies, such as AI, is promising not only for increasing efficiency, but also for advancing health equity and delivering patient-centered care in diverse populations.

Rationale and knowledge gap

With the need for prompt, affordable medical language interpretation services and the vast potential of AI in this realm, obtaining an understanding of its current benefits and limitations will allow us to meet this need with greater efficiency. While research has been conducted to test Google Translate, as well as other forms of AI, in their ability to translate medical information in the form of documents (17) and electronic health record information (22), AI's potential to serve as a medical translator for a patient and medical provider who speak different languages is a more novel idea in the literature. *Figure 1* provides a visual representation of how ML can be applied in clinical settings for language interpretation.

Objective

A thorough systematic review to analyze AI models in this respect will address this gap by aiming to answer the following questions:

(I) How accurate are the medical translations rendered by AI, and to what extent are they deemed acceptable in clinical settings?
(II) How user-friendly is AI for language translation?
(III) Among users of AI for medical language translation, how satisfied are they with its performance, and are they inclined to use it again?
(IV) What feedback have users provided after utilizing AI for medical language translation?
(V) How is AI currently being utilized in the medical field as a language interpreter?

To answer these questions, we will identify and collect relevant reports in the literature on the use of AI as a medical translator to bridge the communication gap between a patient and provider in the clinical setting. First, we will evaluate the accuracy and acceptability of AI-driven language translation in clinical settings. Following this, we will examine the usability, user satisfaction, perspectives of both patients and providers regarding the use of AI for medical translation, and how it is currently being used in this context. Finally, we will explore next steps in research to address the limitations of past studies. We present this article in accordance with the PRISMA reporting checklist (available at https://atm.amegroups.com/article/view/10.21037/atm-24-162/rc) (23).

Methods

Search strategy

Our search was conducted on July 11th, 2024, and a

systematic search strategy was employed to optimize the accuracy of the results yielded. The research units included AI models, language interpretation, and the healthcare industry, and these units were combined using the Boolean operators "AND", "OR". Two final search strings were composed to execute the literature search:

("Artificial Intelligence" OR "Machine Translation" OR "Natural Language Processing" OR "Deep Learning" OR "ChatGPT" OR "Large Language Models" OR "Neural Machine Translation") AND ("Language Translat*" OR "Language Interpret*" OR "Medical Translator" OR "Language Barrier" OR "Limited English Proficiency" OR "Google Translate" OR "Clinical Communication") AND ("Health care" OR "Clinical" OR "Medical" OR "Hospital" OR "Surgery" OR "Patient").

("Artificial Intelligence" OR "Machine Translation" OR "Natural Language Processing" OR "Deep Learning" OR "ChatGPT" OR "Large Language Models" OR "Neural Machine Translation") AND ("Language Translat*" OR "Language Interpret*" OR "Medical Translator" OR "Google Translate") AND ("Health care" OR "Clinical" OR "Medical" OR "Hospital").

The former was utilized for every database except for the IEEE Xplore database, which used the latter to narrow the results to relevant publications.

Databases searched

Two investigators conducted the literature search across five, digital bibliographic databases and one search engine to facilitate a more comprehensive review and avoid omitting relevant reports. Searches were conducted in the databases Cumulated Index in Nursing and Allied Health Literature (CINAHL), Institute of Electrical and Electronics Engineers (IEEE) Xplore, PubMed, Scopus, and Web of Science, and the search engine, Google Scholar.

All searches were conducted using the basic search bar. The filter "2016–2024" was applied after the search for CINAHL, Google Scholar, IEEE Xplore, PubMed, and Scopus, and the filter "2019 or 2020 or 2021 or 2022 or 2023 or 2024" was applied after the search for Web of Science, which did not yield results for years 2015, 2016, 2017, or 2018. In Scopus, the search was done within "Article title, Abstract, Keywords".

Due to the extensive number of results from Google Scholar [15,800], only the initial 100 papers from the search results were reviewed. This selection was based on Google Scholar's default relevance-based sorting algorithm, which prioritizes the most pertinent studies, as relevance decreased significantly beyond the first 100 results. This method provided a balance between managing the vast number of results and maintaining a transparent and reproducible selection process.

Selection of reports and study eligibility

Inclusion criteria included the use of AI to serve as a language interpreter, the translation was used for communication between a real or theoretical healthcare worker and patient, and clinical information must be translated by AI.

Exclusion criteria included review articles, letters to the editor, commentaries, book chapters, lectures, retracted articles, tutorials, reports that were not peer-reviewed, reports that did not have an English translation available, reports published prior to 2016, reports that focused on sign language translation, and reports that focused on translating patient instructions or education material.

In this review, we applied stringent inclusion criteria to ensure that only high-quality reports that were relevant to patient-provider interactions were included. This process, emphasizing quality over quantity, focused on diverse methodologies across multiple settings and medical specialties to increase applicability.

Articles were imported into the Endnote software (version 21.3) where they were labeled and organized by database. After the removal of duplicate reports, title and abstract screening was thoroughly conducted chronologically by two independent reviewers, whereafter full text retrieval and further reports were removed based on the inclusion and exclusion criteria. *Figure 2* demonstrates this process in accordance with the PRISMA (23) flow diagram.

Data quality

The team employed a three-step process to evaluate the quality of the selected reports. A title and abstract screening were conducted to determine its relevance to our research questions. Next, we retrieved the full texts of reports that had passed the title and abstract screening and thoroughly reviewed them. Lastly, we conducted a comprehensive review of the selected papers in their entirety to ensure their applicability and ability to address our research questions.

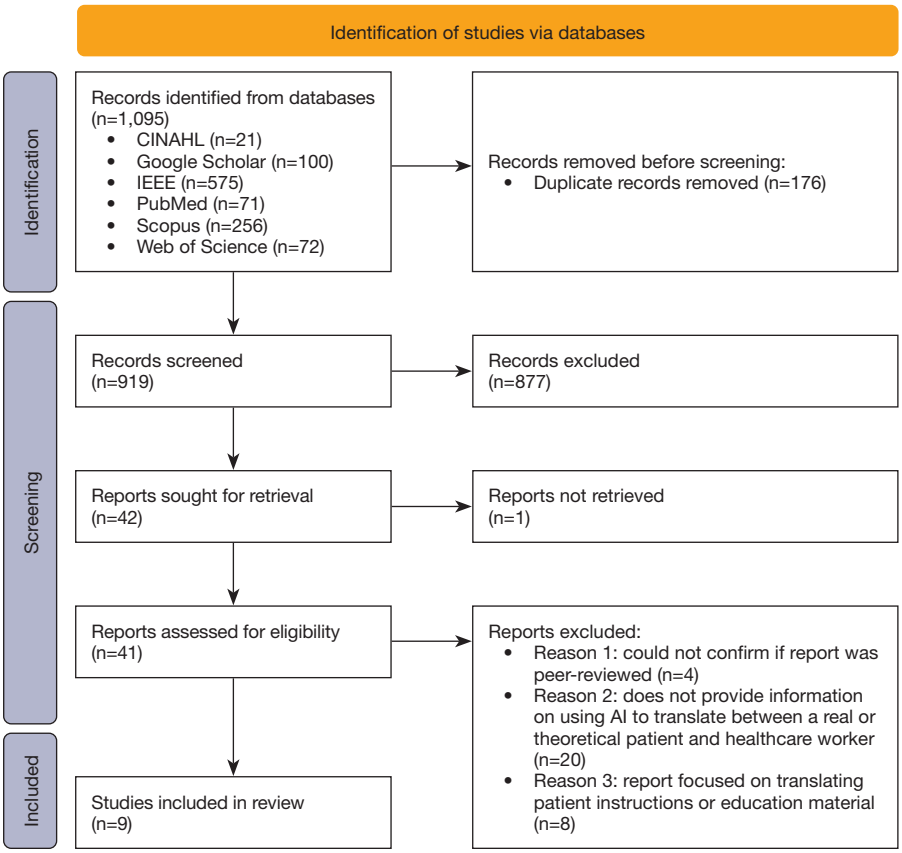

Figure 2 PRISMA flow diagram (23)—study selection process. Our methods for selecting reports for the systematic review were based on the PRISMA 2020 statement. CINAHL, Cumulated Index in Nursing and Allied Health Literature; IEEE, Institute of Electrical and Electronics Engineers; AI, artificial intelligence.

Risk of bias assessment

To assess bias in the selected papers, two researchers conducted an independent assessment using the Risk Of Bias In Non-randomized Studies - of Interventions (ROBINS-I) tool (24) for each study, except for those by Mehandru *et al.* (10) and Tougas *et al.* (9), which were evaluated with Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Qualitative Research (25) and the Mixed Methods Appraisal Tool (MMAT) version 2018 (26), respectively. The JBI Critical Appraisal Checklist for Qualitative Research was better suited for Mehandru *et al.* (10) due to its focus on evaluating the rigor and credibility of qualitative research, while the MMAT was employed to accommodate both quantitative and qualitative components of Tougas *et al.* (9), which are beyond the scope of ROBINS-I. Discrepancies in assessments between the first two authors were resolved by the independent decision of the third author. Upon completion of the bias evaluation, a chart and summary of the ROBINS-I (24) assessments were created using Microsoft Excel and Microsoft Word.

Data synthesis and analysis

The data collected from each article included the authors, year of publication, study location, type of AI used, translation software employed, and information relevant to translation accuracy, usability of AI tools, user satisfaction, and feedback on AI for medical language translation. Accuracy scores were reported as the percentage of translation deemed as accurate by the authors, while usability scores were reported as a percentage, encompassing visits meeting consultation goals, visits with successful use of AI, and visits able to use AI to answer questions without a human interpreter present. Reports that had information such as accuracy and usability percentages,

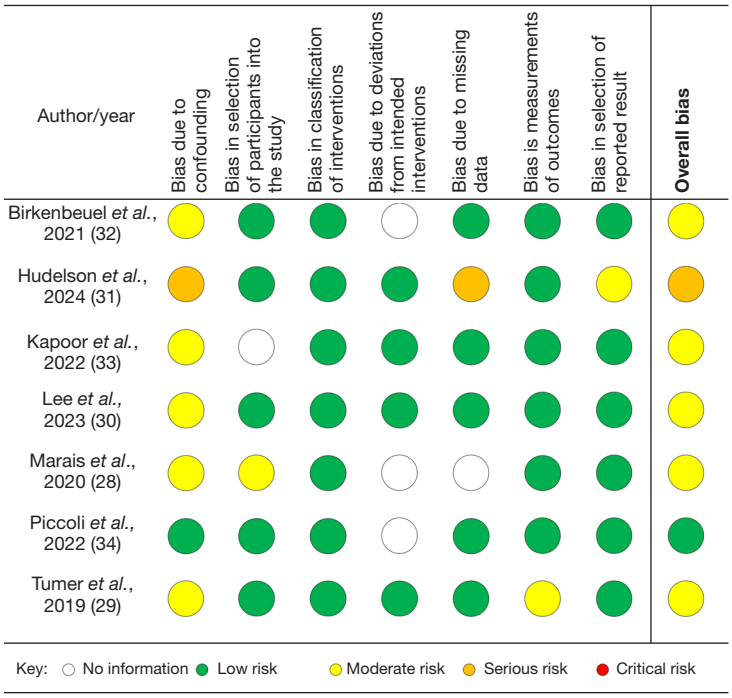

Figure 3 Risk of bias of included studies using ROBINS-I (24). ROBINS-I was utilized for the bias assessment of seven studies (28-34) in this review. The result of these studies, evaluating seven key factors with potential to contribute to bias, are demonstrated. One study was found to have an overall low risk of bias, five with moderate risk, one with serious risk, and none with critical risk. This figure was generated using Microsoft Word. ROBINS-I, Risk Of Bias In Non-randomized Studies - of Interventions.

as well as descriptive data pertaining to our outcomes, were eligible for synthesis. Two reviewers conducted the data extraction. If a report answered any of our research questions, data were collected. If a report partially answered a research question, the data presented was obtained and no assumptions were made over the data omitted. A visual summary of the results was prepared using Microsoft Word.

A quantitative meta-analysis was not performed due to variations in study designs and outcome measures, and different metrics for assessing accuracy, usability, and satisfaction. This heterogeneity would prevent a strong statistical analysis. Therefore, a meaningful narrative synthesis of the results of a systematic search was chosen to provide context-specific interpretation of the findings.

Results

Number of reports in review

The search conducted yielded a total of 1,095 reports across five databases and one search engine. Upon applying the inclusion and exclusion criteria, nine studies focusing on the use of AI in medical language interpretation were selected for the systematic review. Of note, one study (27) was excluded despite meeting many of the criteria because both reviewers agreed that it lacked a substantial clinical focus, concentrating instead on other community settings.

Risk of bias

Figure 3 shows a chart of included studies evaluated using ROBINS-I (24) and a summary of the results can be found in *Figure 4*. As for the two studies not assessed using ROBINS-I (24), for Mehandru *et al.* (10), the outcome using the JBI Critical Appraisal Checklist for Qualitative Research (25) was an overall appraisal of "Include", with every response being "Yes", except for items 6 and 7, which asked if there was a statement locating the researcher culturally or theoretically and if the influence of the researcher were addressed, and the responses were "No". The outcome of the MMAT (26), evaluating Tougas *et al.* (9), was 100% "Yes" for screening questions in the qualitative, quantitative non-randomized, and mixed methods sections.

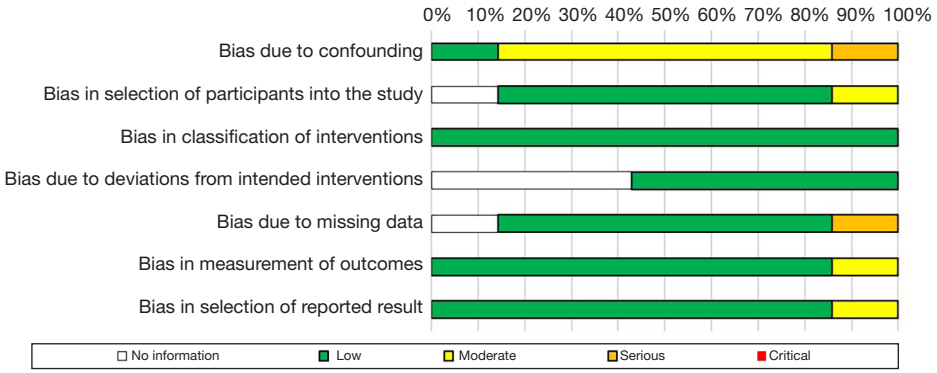

Figure 4 ROBINS-I (24) risk of bias summary. This figure illustrates that the primary source of bias across the seven studies evaluated using the ROBINS-I tool stemmed from confounding, followed by missing data, participant selection, outcome measurement, and selection of reported results. In contrast, there was a low risk of bias associated with the classification of interventions. This figure was generated using Microsoft Excel. ROBINS-I, Risk Of Bias In Non-randomized Studies - of Interventions.

These findings allowed us to conclude that neither study presented a significant risk of bias.

Data obtained

The nine studies encompassed a variety of methodological approaches, including a non-controlled pilot implementation study (28), a scenario-based simulation study (29), a comparative evaluation, a non-inferiority study (30), three non-controlled, observational studies (31-33), one of which was a cohort study (33), a qualitative interview study (10), a case study (34), and a cross-sectional study nested within a randomized controlled trial (9).

The included studies were conducted in various regions around the world. Six took place in the United States (9,10,29,30,32,33), one in South Africa (28), one in Switzerland (31), and one in France (34).

While our search parameters allowed for multiple types of AI in the setting of language translation, only machine translation (MT) studies were yielded that met our criteria for the systematic review. While all nine studies discuss MT, the translation program utilized in the studies varied from Google Translate (10,29,30,32-34), Microsoft Translator (30,31), Apple iTranslate (30), Pocketalk W (31), AwezaMed (28), and the Asynchronous Telepsychiatry (ATP) App (9).

The data is presented in *Table 1*.

Accuracy and acceptability

One of the primary concerns regarding the use of AI as a medical language translator is the accuracy of translations, and acceptability, defined by a score of four or higher on the five-point Likert scale, and these parameters were assessed by three of the nine studies in this review. While Google Translate showed a high overall accuracy at 92.2% (32), human medical interpretation remains superior in acceptability of translation, outperforming Google Translate, Apple iTranslate, and Microsoft Translator (30). However, human translation did not demonstrate superiority when translating figurative language devices (9). The results reveal that AI may perform better when translating one sentence at a time (32), translating from English (30), and translating European languages (31). While medical terminology did not seem to affect translation accuracy, it did impact speech synthesis accuracy (32), which is crucial when using a text-to-speech program. There was no significant difference for MT over zoom compared to in-person translation (9).

Usability

The usability of MT in medical settings has shown promising results, with 82.7% of consultations meeting their goals when using MT (31). Further speaking to AI's serviceability, a study reported that 87.8% of patients found that communicating through MT was easy (31), reflecting the tool's ability to transition into medical settings. Regarding symptom communication, 76.7% of Spanish-speaking patients successfully used MT to articulate their postoperative pain and nausea (33). In the same study, 83.3% of patients were able to use MT successfully on their first attempt, and 96.7% managed to answer all questions at least once without needing a human interpreter (33). In

Table 1 Characteristics and results of included studies

Author and year	Study type	Clinical setting	AI model and program evaluated	Languages translated	Results
Birkenbeuel *et al.*, 2021, (32)	Non-controlled, observational study	Theoretical interaction between a patient and clinician	Machine Translation, Google Translate	English to Spanish	Accuracy and acceptability 92.2% overall translation accuracy Higher translation accuracy for 1 sentence compared to 2 or 3 (P<0.001) 93.9% translation accuracy for 1 sentence less than 8 words 97.8% translation accuracy for 1 sentence greater than 8 words 81.0% translation accuracy for 2 sentences 83.9% translation accuracy for 3 sentences The presence of medication terms Did not significantly affect translation accuracy (P=0.29) Did significantly impact speech synthesis accuracy (P<0.001)
Hudelson *et al.*, 2024, (31)	Non-controlled, observational study	Primary care	Machine Translation, Microsoft Translator, Pocketalk W	Arabic, English, Portuguese, Romanian, Spanish, Turkish, Ukrainian, Russian	Usability 82.7% met consultation goals 87.8% of patients found MT communication to be easy Satisfaction and willingness to use 53.8% of communications were found satisfactory by HCPs 73% of HCPs would be willing to use the application again 84% of patients would be willing to use the application again Challenges with speech recognition and translation quality were noted in non-European languages, making it less favorable to use User feedback 77% of patients preferred the app or found it as effective as a human interpreter for sensitive topics

Table 1 *(continued)*

Table 1 *(continued)*

Author and year	Study type	Clinical setting	AI model and program evaluated	Languages translated	Results
Kapoor *et al.*, 2022, (33)	Non-controlled, observational cohort study	Surgery	Machine Translation, Google Translate	English, Spanish	Usability 76.7% of Spanish-speaking patients successfully used MT to communicate their postoperative pain and nausea 83.3% of patients successfully used MT on their first attempt 96.7% of patients were able to answer all questions at least once without a human interpreter Satisfaction and willingness to use 96.6% of patients were satisfied/highly satisfied with MT's symptom evaluation 83.3–86.7% of nurses were satisfied or very satisfied with MT's speed and effectiveness
Lee *et al.*, 2023, (30)	Comparative evaluation, non-inferiority study	Theoretical interaction between a patient and clinician	Machine Translation, Google Translate, Apple iTranslate, and Microsoft Translator	English, Spanish, Mandarin	Accuracy and acceptability Accuracy scores were higher when translating from English than to English English to Spanish: 0.83 to 0.96 English to Mandarin: 0.88 to 0.91 Spanish to English: 0.70 to 0.76 Mandarin to English: 0.36 to 0.59 Human interpreters surpassed all evaluated machine translation platforms in acceptability
Marais *et al.*, 2020, (28)	Non-controlled, pilot implementation study	Maternal health	Machine Translation, AwezaMed	English, Afrikaans, IsiXhosa, isiZulu	Usability Patients typically reported that the synthetic voice used was clear User feedback Some providers found using MT devices to be less respectful to the patient compared to using human interpreters Some patients felt MT translation into their native language provided comfort and respect

Table 1 *(continued)*

Table 1 *(continued)*

Author and year	Study type	Clinical setting	AI model and program evaluated	Languages translated	Results
Mehandru *et al.*, 2022, (10)	Qualitative interview study	Interactions of patients with physicians, surgeons, nurses, and midwives	Machine Translation, Google Translate	English, Marshallese, Spanish, Russian, sign language, and others	Satisfaction and willingness to use HCPs were less willing to use MT when obtaining patient consent User feedback HCPs had concerns about MT's reliability and quality of translations There is no protective accountability with MT Current use by providers MT was used a last-resort option or when human translators weren't available HCPs used Google Translate when they knew a language but did not have a translation certification HCPs used Google Translate for brief conversations (i.e., rounding)
Piccoli *et al.*, 2022, (34)	Case study	General practice	Machine Translation, Google Translate	Albanian, English, French	User feedback Using Google Translate may increase cognitive demands on lay interpreters and marginalize the patient The absence of context in Google Translate may contribute to misunderstanding Current use by providers To translate with lay interpreter
Tougas *et al.*, 2022, (9)	Cross-sectional study	Psychiatry	Machine Translation, ATP App	English, Spanish	Accuracy and acceptability There was no significant difference between the translation accuracy of AI or human interpretation of figurative language devices (P=0.17) Zoom compared to in-person translations (P=0.39) MT accuracy was nearly significantly better over Zoom than in-person (P=0.06)
Turner *et al.*, 2019, (29)	Scenario-based simulation study	Theoretical emergency medical services	Machine Translation, Google Translate	English, Spanish, Mandarin	Usability Google Translate had a significantly lower usability score compared to QuickSpeak (P=0.04) Satisfaction and willingness to use 92% of Chinese and 86% of Spanish-speaking users preferred QuickSpeak over Google Translate

MT, machine translation; HCP, healthcare professional; ATP, Asynchronous Telepsychiatry; AI, artificial intelligence; QuickSpeak, fixed-response language translation emergency response system.

a pilot study using an AI-produced voice to communicate, patients generally found the synthetic voice clear, which contributed to the overall positive experience and user-friendliness (28). However, there are variations in usability among different tools, and MT, while exhibiting potential, may not be the best option for all clinical situations in its current state. For instance, Google Translate had a significantly lower usability score compared to QuickSpeak (P=0.04), a fixed-response translation system used for language translation in emergency settings, indicating that the choice of application can impact user experience and effectiveness (29).

Satisfaction and willingness to use

Healthcare providers' satisfaction with MT is unclear, as a study evaluating Microsoft Translator and Pocketalk W found that only 53.8% of communications were deemed adequate by health professionals, but 73% were willing to use the application again (31). In contrast, 83.3–86.7% of nurses using Google Translate to evaluate postoperative pain and nausea were satisfied with the effectiveness and speed of the application in assessing patient symptoms (33). A barrier to healthcare provider satisfaction to consider is MT's inability to provide legal support in high-stakes interactions, such as obtaining patient consent for a procedure (10). Conversely, patients appear to be more satisfied with MT for medical communication, with 84% willing to use it again (31) and 96.6% expressing satisfaction with MT's symptom evaluation (33). Additionally, there is evidence that other language translation software outside of MT might be preferred in certain situations, as 92% and 86% of Chinese and Spanish speakers, respectively, had a preference for another software over Google Translate for emergency communication (29).

User feedback

User feedback on MT in medical settings is mixed. Notably, 77% of patients preferred the app or found it as effective as a human interpreter for discussing sensitive topics (31), citing comfort and respect when their native language was used (28). However, some healthcare providers felt that using MT devices was less respectful to patients compared to human interpreters (28), expressing concerns about the reliability and quality of translations (10). Moreover, it was noted that using Google Translate to assist a lay interpreter may increase cognitive demands on the lay interpreters and further marginalize the patient (34).

Current use of AI as a medical language interpreter

AI-based language translation is currently utilized in medical settings primarily as a supplementary tool when human translators are not available. Healthcare providers often turn to MT as a last-resort option, particularly in situations where certified human translators are unavailable (10). Google Translate is frequently employed by healthcare professionals (HCPs) who have some knowledge of a language but lack official translation certification, aiding in the translation of medical terms (10). Additionally, providers reported that they use programs such as Google Translate to facilitate brief conversations during activities, such as patient rounding (10), where the risk of misinterpretation may be lower. Google Translate has also been used to assist lay interpreters in translating medical information, ensuring that communication is maintained even in the absence of professional interpreters (34).

Discussion

Key findings

This systematic review found that using AI models for medical language translation has future promise. First, it may provide accurate enough translations to meet consultation goals if used for short (32), simple, clinical communication (33), but a professionally trained human medical language interpreter may be preferred for longer (32), more in-depth discussions. Additionally, human translation may be needed to meet acceptability standards when translating to English (30), or when translating non-European languages (31). MT tended to be viewed as user-friendly (28,31,33), and overall, patients were more satisfied with their AI-translated encounters than their clinical counterparts (10,31,33). Next, HCPs reported reservations with AI-translations due to issues with accuracy (10), respect to the patient (28), and lack of liability protection (10). Despite these concerns, MT services have already been implemented in the clinical setting to bridge communication gaps for brief conversations or as a last-resort option when human interpreters are not accessible (10,34). *Figure 5* demonstrates the key components of AI-driven medical language interpretation as evaluated in this study.

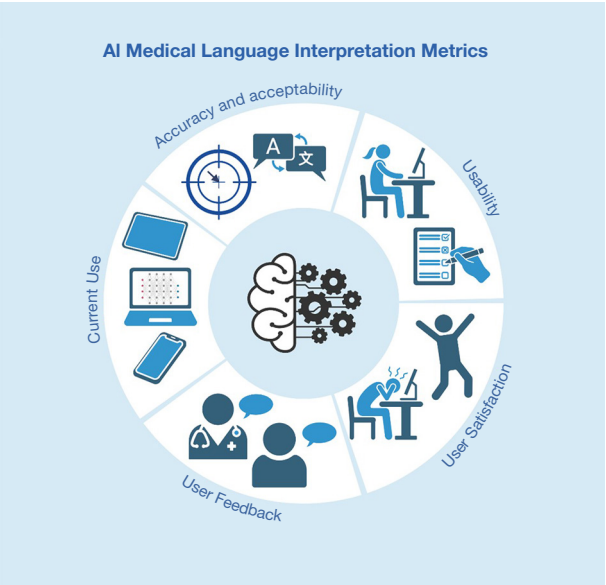

Figure 5 Medical language interpretation metrics evaluated. This systematic review focused on the accuracy, acceptability, and usability of language translation performed by artificial intelligence in the clinical setting. Furthermore, it gives information on its current use, how satisfied users are, and their feedback on its application in medical language translation. This figure was generated using BioRender. AI, artificial intelligence.

Translation platforms evaluated

Each AI translation platform in this review offers distinct strengths and limitations. General-use tools like Google Translate and Microsoft Translator each support over 100 languages and use NMT (35,36), often achieving higher accuracy for European languages (31). Apple's iTranslate similarly supports multiple languages, but may be better suited for conversational use, as errors were identified with complex sentences (37). Tools designed for clinical use, like Pocketalk W and AwezaMed, address specific needs. Pocketalk W provides on-demand translation in 84 languages during the visit (31,38), while AwezaMed specializes in utterances tailored to specific medical scenarios for patient-provider communication (28). The ATP App enhances languages accessibility in mental health by supporting remote consultations through language translation (39). These distinctions emphasize that the choice of tool depends on communication needs and language specifics.

Potential of large language models (LLMs) in translation

LLMs, like ChatGPT, are increasingly used in general language translation, demonstrating strengths and limitations for future integration in healthcare. ChatGPT-4 achieved comparable accuracy to human translation at 0.68 when translation Persian to English (40), and ChatGPT's translations were found to be similar or superior to commercial translations for high-resource languages (41). Jiao *et al.* reported similar findings, noting struggles with low-resource languages but comparable performance to Google Translate for European languages (42). These findings highlight the potential of LLMs in healthcare translation and the need for more expansive language coverage across AI models.

Practical applications in clinical settings

Despite its need for improvement, AI translation has already demonstrated utility in clinical settings. During a medical consultation with an Albanian family in France, Google Translate effectively bridged language barriers despite occasional misunderstandings (34). Google Translate was also successfully used for diagnosis and management in a psychiatric visit when a human alternative was not available (43). These examples illustrate that while further development is necessary, AI translation tools already have meaningful applications in clinical practice, supporting communication in critical scenarios.

However, structural limitations in clinical practice impact AI translation effectiveness, as each environment presents unique demands. In intensive care units, where clear and empathetic communication with family members is crucial, AI tools may miss emotional nuances or struggle with complex medical terminology. While progress has been made in recognizing human emotion through speech elements (44), these capabilities require further improvement and integration into translation platforms. Fast-paced environments like emergency rooms demand rapid, context-specific communication, where AI may produce inaccuracies, particularly in medical language. Although medical terms did not significantly affect translation accuracy, their inclusion affected speech synthesis (32). Further examining translation of medical terminology, while "pneumonia" could be accurately translated between English and Arabic, errors like incorrect word order and literal translations were

observed (45). These examples emphasize the need for ongoing development to ensure AI translation tools meet specific communication needs across varied environments.

Ethical and legal implications

The integration of AI translation tools into healthcare raises critical legal and ethical considerations, particularly concerning patient confidentiality, transparency, and data protection. The use of AI for language translation may be underreported due to these implications. While federal mandates require language access for patients with limited English proficiency (46), there is a lack of comprehensive legal regulation regarding AI in clinical settings (47). Emerging frameworks like the White House's AI principles, state-level legislative endorsements, and Food and Drug Administration (FDA) oversight aim to address issues like transparency and public trust in AI (48), but gaps remain in regulations specifically addressing AI-driven applications like medical translation.

Major ethical concerns include the need for informed consent regarding the use of patient data, transparency in how AI systems generate outputs, and accountability for decisions made by models (47). Robust data privacy protections are needed, as beaches to sensitive health information can lead to identify theft, emotional harm, or financial loss (49). Addressing these challenges will require comprehensive regulations to ensure AI translation tools are implemented responsibly, prioritizing patient safety.

Limitations of evidence

In this systematic review, we acknowledge several limitations inherent in the evidence gathered, which may influence the interpretation and applicability of our findings. One such limitation is the modest quantity of languages translated and translations assessed across the studies in this review, reducing the generalizability of our findings. Furthermore, not all studies evaluated bidirectional translation, as Birkenbeuel *et al.* (32) assessed one-way translation from English to Spanish while Tougas *et al.* (9) used AI to translate an interview with a patient conducted in Spanish by a researcher which was later provided to a physician. Another limitation of evidence is that some evaluations were theoretical or simulated scenarios as opposed to real-world encounters between patients and healthcare providers (29,30,32). Additionally, we excluded reports that were not in English, and more evidence may be present in other languages. Other limitations identified in the included reports include using one human interpreter to evaluate the retention of meaning (32), HCPs not obtaining patient feedback after an encounter (31), and researchers defining their own standards for assessing MT interpretation (30). Finally, this review contained results from studies published between 2019–2024, and evolution of translation tools may result in decreased relevance of data extracted.

Strengths and limitations of this review

We discussed the use of AI in the setting of medical language interpretation, and evaluated the literature for evidence of accuracy, usability, user satisfaction, user feedback, and current use. Furthermore, we employed a rigorous and clear methodology to identify, select, and assess pertinent reports to enhance the credibility and reliability of our study. Additionally, the inclusion of multiple AI translation platforms contributes to the strength of this review. However, the limited number of studies on the subject and various methodologies employed across the included studies provided a barrier to assembling a final verdict on AI-based language interpretation. Although, the limited evidence reflects the current emerging nature of AI in clinical translation and the rigorous criteria applied to enhance the reliability of evidence presented. The result of this selective approach emphasizes the need for further research to build upon these foundational findings and explore applications across more clinical environments.

Future directions

The future of AI translation in healthcare lies in its ability to deliver real-time accurate communication across diverse clinical scenarios. Emerging advancements are likely to enhance processing speeds and contextual accuracy, allowing integrating with electronic health records (22) to translate key documents like education material (50,51) and discharge instructions (17,52) as needed. Innovations in speech recognition systems, such as those demonstrated by smart speakers for hands-free communication in clinical settings (53), highlight the potential for integrating similar technology into language translation platforms, enabling real-time, hands-free communication in critical environments like emergency rooms. AI may also facilitate efficient collaboration between multilingual teams, in in resource-limited settings, providing translation services when human interpreters are unavailable. Advancements

will allow for bridging linguistic barriers in real time, ultimately improving patient safety, enhancing provider efficiency, reducing costs for institutions, and expanding access to care across varied healthcare environments.

The findings in this review hold significant implications for HCPs, guiding them in utilizing AI as a supportive tool, rather than a replacement, for language translation, thereby enhancing the quality of care provided across language barriers. Based on these results, future directions and recommendations include:

(I) Clinicians should utilize professionally trained human medical language interpreters until the accuracy of bidirectional translations increase across more, non-European languages.
(II) In resource-limited settings where human interpreters are not accessible, AI may provide sufficiently accurate translations for brief consultations that use short sentences and involve European languages.
(III) Use a standardized metric, such as the Bilingual Evaluation Understudy (BLEU) score, for assessing the accuracy of translations.
(IV) If an AI-based translation is utilized, use the teach-back method, where the patient repeats the information back to the provider in their words, to evaluate their level of understanding.
(V) Future research should be conducted to increase the sample size across multiple European and non-European languages, test MT in real-world settings, and create a system for Health Insurance Portability and Accountability Act (HIPAA)-compliant translations.

Conclusions

Our systematic review of nine reports provided a comprehensive overview of the use of AI to serve as a medical language interpreter and concluded that AI holds considerable potential for future applications in clinical settings. While AI translation can sufficiently meet the needs for simple, straightforward clinical communications, they are currently inadequate for extensive discussions and remains inferior to human interpretation. Despite an array of impressive results in accuracy, usability, and user satisfaction, the lack of consistency among findings suggests a need for more comprehensive research to ease the hesitations among healthcare providers. Overall, while patient satisfaction with AI-translated encounters is generally high, the requirements of medical consultations necessitate a balanced integration of AI and human translation services to ensure high-quality, comprehensive care.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the PRISMA reporting checklist. Available at https://atm.amegroups.com/article/view/10.21037/atm-24-162/rc

Peer Review File: Available at https://atm.amegroups.com/article/view/10.21037/atm-24-162/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-24-162/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. U.S. Census Bureau. QuickFacts: United States 2024. Available online: https://www.census.gov/quickfacts/
2. Kirkman-Liff B, Mondragón D. Language of interview: relevance for research of southwest Hispanics. Am J Public Health 1991;81:1399-404.
3. Bischoff A, Denhaerynck K. What do language barriers cost? An exploratory study among asylum seekers in Switzerland. BMC Health Serv Res 2010;10:248.
4. Divi C, Koss RG, Schmaltz SP, et al. Language proficiency

and adverse events in US hospitals: a pilot study. Int J Qual Health Care 2007;19:60-7.
5. Heath M, Hvass AMF, Wejse CM. Interpreter services and effect on healthcare - a systematic review of the impact of different types of interpreters on patient outcome. J Migr Health 2023;7:100162.
6. Locatis C, Williamson D, Gould-Kabler C, et al. Comparing in-person, video, and telephonic medical interpretation. J Gen Intern Med 2010;25:345-50.
7. Bischoff A, Hudelson P. Access to healthcare interpreter services: where are we and where do we need to go? Int J Environ Res Public Health 2010;7:2838-44.
8. Lundin C, Hadziabdic E, Hjelm K. Language interpretation conditions and boundaries in multilingual and multicultural emergency healthcare. BMC Int Health Hum Rights 2018;18:23.
9. Tougas H, Chan S, Shahrvini T, et al. The Use of Automated Machine Translation to Translate Figurative Language in a Clinical Setting: Analysis of a Convenience Sample of Patients Drawn From a Randomized Controlled Trial. JMIR Ment Health 2022;9:e39556.
10. Mehandru N, Robertson S, Salehi N. Reliable and safe use of machine translation in medical settings. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; 2022:2016-25.
11. Hsieh E. Not just "getting by": factors influencing providers' choice of interpreters. J Gen Intern Med 2015;30:75-82.
12. Nápoles AM, Santoyo-Olsson J, Karliner LS, et al. Inaccurate Language Interpretation and Its Clinical Significance in the Medical Encounters of Spanish-speaking Latinos. Med Care 2015;53:940-7.
13. Jacobs EA, Shepard DS, Suaya JA, et al. Overcoming language barriers in health care: costs and benefits of interpreter services. Am J Public Health 2004;94:866-9.
14. University of Illinois Chicago. What is (AI) Artificial Intelligence? | Online Master of Engineering 2024. Available online: https://meng.uic.edu/news-stories/ai-artificial-intelligence-what-is-the-definition-of-ai-and-how-does-ai-work
15. Han L, Gladkoff S, Erofeev G, et al. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. Front Digit Health 2024;6:1211564.
16. Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv e-prints 2016. arXiv:1609.08144.
17. Khoong EC, Steinbrook E, Brown C, et al. Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. JAMA Intern Med 2019;179:580-2.
18. Zolnierek KB, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care 2009;47:826-34.
19. Fiscella K, Meldrum S, Franks P, et al. Patient trust: is it related to patient-centered behavior of primary care physicians? Med Care 2004;42:1049-55.
20. Martin LR, Williams SL, Haskard KB, et al. The challenge of patient adherence. Ther Clin Risk Manag 2005;1:189-99.
21. Al Shamsi H, Almutairi AG, Al Mashrafi S, et al. Implications of Language Barriers for Healthcare: A Systematic Review. Oman Med J 2020;35:e122.
22. Soto X, Perez-de-Viñaspre O, Labaka G, et al. Neural machine translation of clinical texts between long distance languages. J Am Med Inform Assoc 2019;26:1478-87.
23. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
24. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.
25. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015;13:179-87.
26. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Education for Information 2018;34:285-91.
27. Qian N, Rey C, Catalano A, et al., editors. AI-Assisted Translation and Speech Synthesis for Community Services Supporting Limited English Proficient Individuals. Communications in Computer and Information Science; Springer, Cham; 2024: vol 2120.
28. Marais L, Louw JA, Badenhorst J, et al., editors. AwezaMed: A multilingual, multimodal speech-to-speech translation application for maternal health care. Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020; Institute of Electrical and Electronics Engineers Inc. 2020.
29. Turner AM, Choi YK, Dew K, et al. Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study. JMIR Public Health Surveill 2019;5:e11171.
30. Lee W, Khoong EC, Zeng B, et al. Evaluation of

Commercially Available Machine Interpretation Applications for Simple Clinical Communication. J Gen Intern Med 2023;38:2333-9.
31. Hudelson P, Chappuis F. Using Voice-to-Voice Machine Translation to Overcome Language Barriers in Clinical Communication: An Exploratory Study. J Gen Intern Med 2024;39:1095-102.
32. Birkenbeuel J, Joyce H, Sahyouni R, et al. Google translate in healthcare: preliminary evaluation of transcription, translation and speech synthesis accuracy. BMJ Innov 2021;7:422-9.
33. Kapoor R, Corrales G, Flores MP, et al. Use of Neural Machine Translation Software for Patients With Limited English Proficiency to Assess Postoperative Pain and Nausea. JAMA Netw Open 2022;5:e221485.
34. Piccoli V. Plurilingualism, multimodality and machine translation in medical consultations: A case study. Transl Interpreting Stud 2022;17:42-65.
35. Doss Mohan K, Skotdal J. Microsoft Research [Internet]. Microsoft2021. Available online: https://www.microsoft.com/en-us/research/blog/microsoft-translator-now-translating-100-languages-and-counting
36. Johnson M, Schuster M, Le Q, et al. Google's Multilingual Neural Machine Translation System: Enabling. Transactions of the Association for Computational Linguistics 2017;5:339-51.
37. Chen X, Acosta S, Barry AE. Machine or Human? Evaluating the Quality of a Language Translation Mobile App for Diabetes Education Material. JMIR Diabetes 2017;2:e13.
38. Pocketalk. Healthcare 2024. Available online: https://www.pocketalk.com/healthcare
39. Parish MB, Gonzalez A, Hilty D, et al. Asynchronous Telepsychiatry Interviewer Training Recommendations: A Model for Interdisciplinary, Integrated Behavioral Health Care. Telemed J E Health 2021;27:982-8.
40. Ghassemiazghandi M. An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score. Theory and Practice in Language Studies 2024;14:985-94.
41. Gao Y, Wang R, Hou F. How to Design Translation Prompts for ChatGPT: An Empirical Study. ArXiv e-prints 2023. arXiv:2304.02182.
42. Jiao W, Wang W, Huang JT, et al. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. ArXiv e-prints 2023. arXiv:2301.08745.
43. Leite FO, Cochat C, Salgado H, et al. Using Google Translate^© in the hospital: A case report. Technol Health Care 2016;24:965-8.
44. Kapoor A, Verma V. Emotion AI: understanding emotions through artificial intelligence. International Journal of Engineering Science and Humanities 2024;14:223-32.
45. Al-Jarf R. Translation of Medical Terms by AI: A Comparative Linguistic Study of Microsoft Copilot and Google Translate. 1st International Conference on Artificial Intelligence and its Applications in the Age of Digital Transformation; Nouakchott, Mauritania; 2024.
46. Chen AH, Youdelman MK, Brooks J. The legal framework for language access in healthcare settings: Title VI and beyond. J Gen Intern Med 2007;22 Suppl 2:362-7.
47. Naik N, Hameed BMZ, Shetty DK, et al. Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? Front Surg 2022;9:862322.
48. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial Intelligence in Healthcare 2020:295-336.
49. Pressman SM, Borna S, Gomez-Cabello CA, et al. AI and Ethics: A Systematic Review of the Ethical Considerations of Large Language Model Use in Surgery Research. Healthcare (Basel) 2024;12:825.
50. Chen X, Acosta S, Barry AE. Evaluating the Accuracy of Google Translate for Diabetes Education Material. JMIR Diabetes 2016;1:e3.
51. Muraj Z, Ugas M, Tse K, et al. Evaluating the Feasibility and Utility of Machine Translation for Radiation Therapy Patient Education Materials. Journal of Medical Imaging & Radiation Sciences 2024;55:S9.
52. Brewster RCL, Gonzalez P, Khazanchi R, et al. Performance of ChatGPT and Google Translate for Pediatric Discharge Instruction Translation. Pediatrics 2024;154:e2023065573.
53. Yoo TK, Oh E, Kim HK, et al. Deep learning-based smart speaker to confirm surgical sites for cataract surgeries: A pilot study. PLoS One 2020;15:e0231322.

Cite this article as: Genovese A, Borna S, Gomez-Cabello CA, Haider SA, Prabha S, Forte AJ, Veenstra BR. Artificial intelligence in clinical settings: a systematic review of its role in language translation and interpretation. Ann Transl Med 2024;12(6):117. doi: 10.21037/atm-24-162