Evaluation of Commercially Available Machine Interpretation Applications for Simple Clinical Communication

Won Lee, MD, ScM[1], Elaine C. Khoong, MD, MS[1,2], Billy Zeng, MD[1], Francine Rios-Fetchko, BA[1], YingYing Ma, BS[1], Kirsten Liu, MSW[1,3], and Alicia Fernandez, MD[1,2]

[1]University of California San Francisco, 513 Parnassus Ave, Room S-436, San Francisco, CA 94143, USA; [2]Zuckerberg San Francisco General Hospital, San Francisco, CA, USA; [3]University of California Berkley, Berkely, CA, USA

BACKGROUND: Accessing professional medical interpreters for brief, low risk exchanges can be challenging. Machine translation (MT) for verbal communication has the potential to be a useful clinical tool, but few evaluations exist.

OBJECTIVE: We evaluated the quality of three MT applications for English–Spanish and English-Mandarin two-way interpretation of low complexity brief clinical communication compared with human interpretation.

DESIGN: Audio-taped phrases were interpreted via human and 3 MT applications. Bilingual assessors evaluated the quality of MT interpretation on four assessment categories (accuracy, fluency, meaning, and clinical risk) using 5-point Likert scales. We used a non-inferiority design with 15% inferiority margin to evaluate the quality of three MT applications with professional medical interpreters serving as gold standards.

MAIN MEASURES: Proportion of interpretation exchanges deemed acceptable, defined as a composite score of 16 or greater out of 20 based on the four assessment categories.

KEY RESULTS: For English to Spanish, the proportion of MT-interpreted phrases scored as acceptable ranged from 0.68 to 0.84, while for English to Mandarin, the range was from 0.62 to 0.76. Both Spanish/Mandarin to English MT interpretation had low acceptable scores (range 0.36 to 0.41). No MT interpretation met the non-inferiority threshold.

CONCLUSION: While MT interpretation was better for English to Spanish or Mandarin than the reverse, the overall quality of MT interpretation was poor for two-way clinical communication. Clinicians should advocate for easier access to professional interpretation in all clinical spaces and defer use of MT until these applications improve.

Key Words: Machine translation; Machine interpretation; Two-way interpretation; Non-inferiority study

J Gen Intern Med 38(10):2333–9
DOI: 10.1007/s11606-023-08079-6

Prior Presentations *This study was presented at the 2022 International Anesthesia Research Society (IARS) annual meeting (held virtually) on March 19th, 2022.*

Received October 15, 2022
Accepted January 30, 2023
Published online February 13, 2023

INTRODUCTION

For 25 million individuals with limited English proficiency (LEP) in the USA, language barriers limit equitable access to healthcare, which results in worse clinical outcomes and decreased therapeutic engagement.[1–5] Clinical communication extends beyond the transference of information or instruction; it helps to build rapport and interpersonal relationships between patients and clinicians. While certified medical interpretation remains an indispensable tool for communicating with language-discordant patients, these resources are often impractical or unfeasible in certain clinical settings and are therefore underutilized.[6] In the perioperative setting, the busy workflow, the sterile environment, varying levels of patient consciousness, and sporadic and brief conversation exchanges make it challenging to utilize formal medical interpretation services. Consequently, clinicians may forgo using medical interpreters and instead rely on nonverbal communication, which poses a significant challenge to safe and high-quality care.[7]

Machine translation (MT) has the potential to fill the gaps of communicating in language-discordant clinical situations. MT refers to automated software with the capacity for two-way translation (text) and interpretation (speech) between languages. MT products are widely available on mobile devices with small infrastructural cost, making them a tempting pragmatic resource for clinicians. However, the evaluation of MT for healthcare remains limited, and MT use in clinical settings has raised safety concerns.[8,9] MT has been evaluated for translating patient portal messages, discharge instructions, and public health information with mixed results depending on the language translated,[10–12] but only a few have evaluated the use of MT for interpretation.[13] Previous studies have shown that MT interpretation is accurate in limited settings.[9,14]

Machine interpretation is necessarily more complex than machine translation. Proper speech recognition, transcription (speech into written form), and language synthesis (speech generation) are necessary for MT to function as a two-way interpreter. To determine whether MT interpretation is useful for brief and low-stakes two-way communication encounters, we designed a non-inferiority study to compare the accuracy and safety of three commercially available MT applications against professional interpreters between English and

Spanish, as well as between English and Mandarin Chinese, the two most common non-English languages in the United States.[15]

METHODS

Study Design

We designed a non-inferiority study to evaluate the quality of MT interpretation for two-way communication between patients and clinicians. Professional medical interpreter services served as a gold standard. Three MT applications, Google Translate (GT), Apple iTranslate (AT), and Microsoft Translator (MS), were selected based on their availability without cost to users across multiple devices and operating systems. All three applications utilize machine learning algorithm based on artificial neural networks that can improve with aggregation of more data.[16–18]

Recognizing that the perioperative setting is one where professional interpretation is often not used, we formulated study phrases that simulate conversation between English-speaking clinicians and patients with LEP using input from anesthesiologists and perioperative nurses. Each study phrase consisted of one to three sentences in a standard language, devoid of slang or excessive colloquialism, such as "Can you please point to where it hurts the most?" Additional examples of the study phrases are available in Appendix A. Using the conventional, predetermined 15% non-inferiority margin, we developed 105 provider-to-patient and 105 patient-to-provider phrases.

To assess MT interpretation (speech to speech), study phrases were first audio recorded; provider-to-patient phrases were recorded in English, and patient-to-provider phrases were recorded in Spanish and Mandarin by native bilingual speakers. These recordings were played into each MT application, and the resulting interpretations were captured as audio files. Professional medical interpreters were provided with the same audio recording of the study phrases, and their interpretations were also captured as audio files. Transcriptions of study phrases were not provided to simulate live two-way interpretation (Fig. 1).

Each audio recording was reviewed for sound clarity, and volume was adjusted to comparable decibel levels using WavePad Audio Editor (Version 11.33, Canberra, Australia). We downloaded the MT applications from Apple AppStore onto an iPhone running iOS 14.3 for consistency across the device hardware and software versions (GT: 6.16.x, AT: 14.1.x, MS: 4.049x). All machine interpretations and audio records occurred between February 5th and 7th of 2021. Data were collected in a quiet room. A desktop computer with dedicated speakers and a high-fidelity microphone was used to record and capture MT interpretations.

Evaluation Metrics and Outcome Measures

For each language, two bilingual assessors evaluated the quality of MT interpretation, with a third bilingual assessor adjudicating the difference in scores if necessary. The six assessors (3 per language) were a mix of clinician (4) and non-clinical (2) volunteers. Assessors were instructed to listen to interpretation audio files and score one interpretation at a time. The order in which the four interpretations (human, GT, AT, and MS) were presented was randomized for each phrase to mitigate habituation bias. Assessors were instructed to take frequent breaks to minimize fatigue bias. Assessors were also instructed to describe the types of errors encountered in their evaluation process. Errors were classified as omission, abbreviation (inability to accurate identify abbreviation), syntactic (word order and/or sentence structure), lexical (related to vocabulary), nonsense interpretation, and phonemic (distinguishing one word from another, such as pad, pat, bad, and bat).

Due to a lack of consensus on evaluation metrics for MT interpretation, we modified and adapted four assessment categories commonly used for evaluating MT translation.[19,20] "Accuracy" evaluated for a loss of information (omission), "Fluency" assessed grammar, "Meaning" assessed unnecessary additions or changes that impacted meaning, and "Clinical Risk" assessed whether that

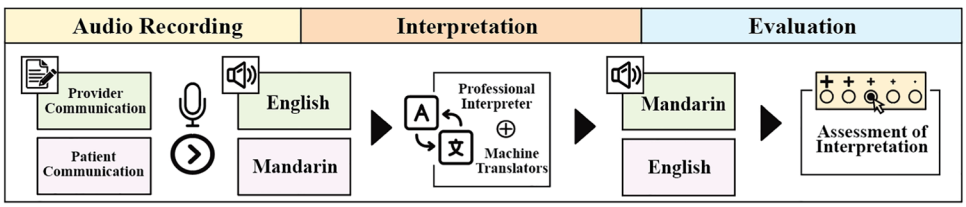

Figure 1 Diagram of study workflow. Study phrases simulating two-way communication between English-speaking providers (English) and patients with limited English proficiency. Provider communications (in green) were recorded in English and patient communications (in pink) were recorded in either Mandarin or Spanish. These recordings were then played onto three MT applications and resulting interpretations were captured as audio files. A professional medical interpreter also provided interpretations, serving as a gold standard. The interpretations were then evaluated by bilingual assessors based on four categories (Fluency, Accuracy, Meaning, and Clinical Risk) using 5-point Likert scale. In this figure, English–Mandarin interpretation workflow is shown. Same steps were taken for evaluating English–Spanish interpretations

Table 1 Composite scores and proportions of acceptable interpretations. Median and interquartile range (IQR) of composite scores and the proportion of interpretations that have met the acceptability criteria (composite score of 16 or higher) is presented with its 95% confidence interval (CI)

Interpreter by language	Composite score Median [IQR]	Acceptable interpretation Rate [95% CI]
English to Spanish		
Apple iTranslate	18.0 [15.0–19.0]	0.68 [0.59–0.77]
Google Translate	19.0 [17.0–19.0]	0.84 [0.77–0.91]
Microsoft Translator	19.0 [16.0–19.0]	0.74 [0.66–0.83]
Human Interpreter	20.0 [20.0–20.0]	1.0
Spanish to English		
Apple iTranslate	14.0 [12.0–19.0]	0.38 [0.29–0.48]
Google Translate	14.0 [11.3–18.3]	0.41 [0.31–0.51]
Microsoft Translator	14.0 [11.0–18.0]	0.37 [0.28–0.47]
Human Interpreter	20.0 [20.0–20.0]	1.0
English to Mandarin		
Apple iTranslate	17.7 [14.7–19.0]	0.62 [0.53–0.71]
Google Translate	19.0 [15.0–20.0]	0.74 [0.66–0.83]
Microsoft Translator	19.0 [16.0–20.0]	0.76 [0.68–0.85]
Human Interpreter	20.0 [20.0–20.0]	1.0
Mandarin to English		
Apple iTranslate	13.0 [9.0–19.0]	0.39 [0.30–0.49]
Google Translate	13.0 [9.0–20.0]	0.36 [0.27–0.46]
Microsoft Translator	14.0 [9.0–19.0]	0.39 [0.30–0.49]
Human Interpreter	20.0 [20.0–20.0]	1.0

change in meaning could lead to a poor patient outcome.[21] Each category was scored on a 5-point Likert scale; Clinical Risk was inversely coded such that a high number indicated less (no) risk. Only the clinicians scored the Clinical Risk category.

The outcome was the acceptability of MT interpretation based on a composite score of the 4 assessment categories. We defined an interpretation as acceptable if it scored 16 or higher out of 20 possible points (four 5-point Likert categories). We also examined each category separately, defining acceptability as a score of 4 or greater on the 5-point Likert scale.

Statistical Analysis

Descriptive statistics (proportions with 95% confidence interval [CI]) were used to characterize the proportion of phrases with acceptable interpretations. Paired *t*-tests were used to compare each MT application to the human interpreter. MT applications were not compared with each other. A *p*-value of less than 0.05 was considered statistically significant for all analyses. Cronbach's alpha was used to measure inter-assessor agreement.

RESULTS

Six assessors evaluated 105 phrases from English to Spanish/Mandarin and 105 phrases from Spanish/Mandarin to English. The inter-assessor reliability was high for both Spanish (alpha: 0.80) and Mandarin (alpha: 0.86). Figure 2 presents the proportion of interpretations that met the acceptability criteria by language and direction of interpretations. For English to Spanish, the proportion of MT-interpreted phrases scored as acceptable ranged from 0.68 to 0.84. Only the GT algorithm came close to the non-inferiority criteria (0.84, 95% CI: 0.77–0.91). For English-to-Mandarin interpretation, the proportion of MT-interpreted phrases scored as acceptable ranged from 0.62 to 0.76; no MT interpretation met the non-inferiority threshold (Table 1). Both Spanish-to-English and Mandarin-to-English interpretations had a lower composite score (median range 13.0 to 14.0 out of 20), and a low proportion of MT-interpreted phrases scored as acceptable (range 0.36–0.41). Every interpretation by professional medical interpreters, both to and from English, was rated highly and scored as acceptable.

Figure 3 shows the proportions of interpreted phrases scored as being acceptable by individual assessment categories. For English to Spanish, scores of the accuracy (range 0.83 to 0.96) and clinical risk (0.82 to 0.90) categories were

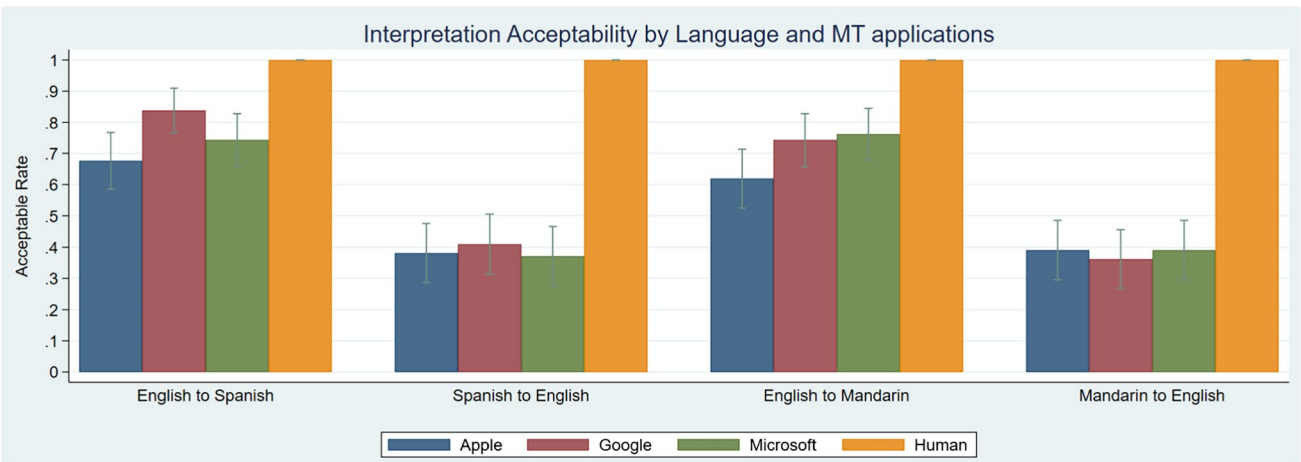

Figure 2 Proportion of interpreted phrases deemed acceptable based on the composite scores of 4 assessment categories

higher than fluency (0.60 to 0.81) and meaning (0.75 to 0.85) for MT applications. For Spanish-to-English interpretations, accuracy scored 0.70 to 0.76, but the other three categories scored lower (0.40 to 0.51). For English to Mandarin, MT applications scored better in the accuracy category (0.88 to 0.91) than the other three categories (0.68 to 0.86). For Mandarin to English, all four categories scored low (0.36 to 0.59).

Assessors described the types of errors they encountered during their evaluation of MT interpretations. Table 2

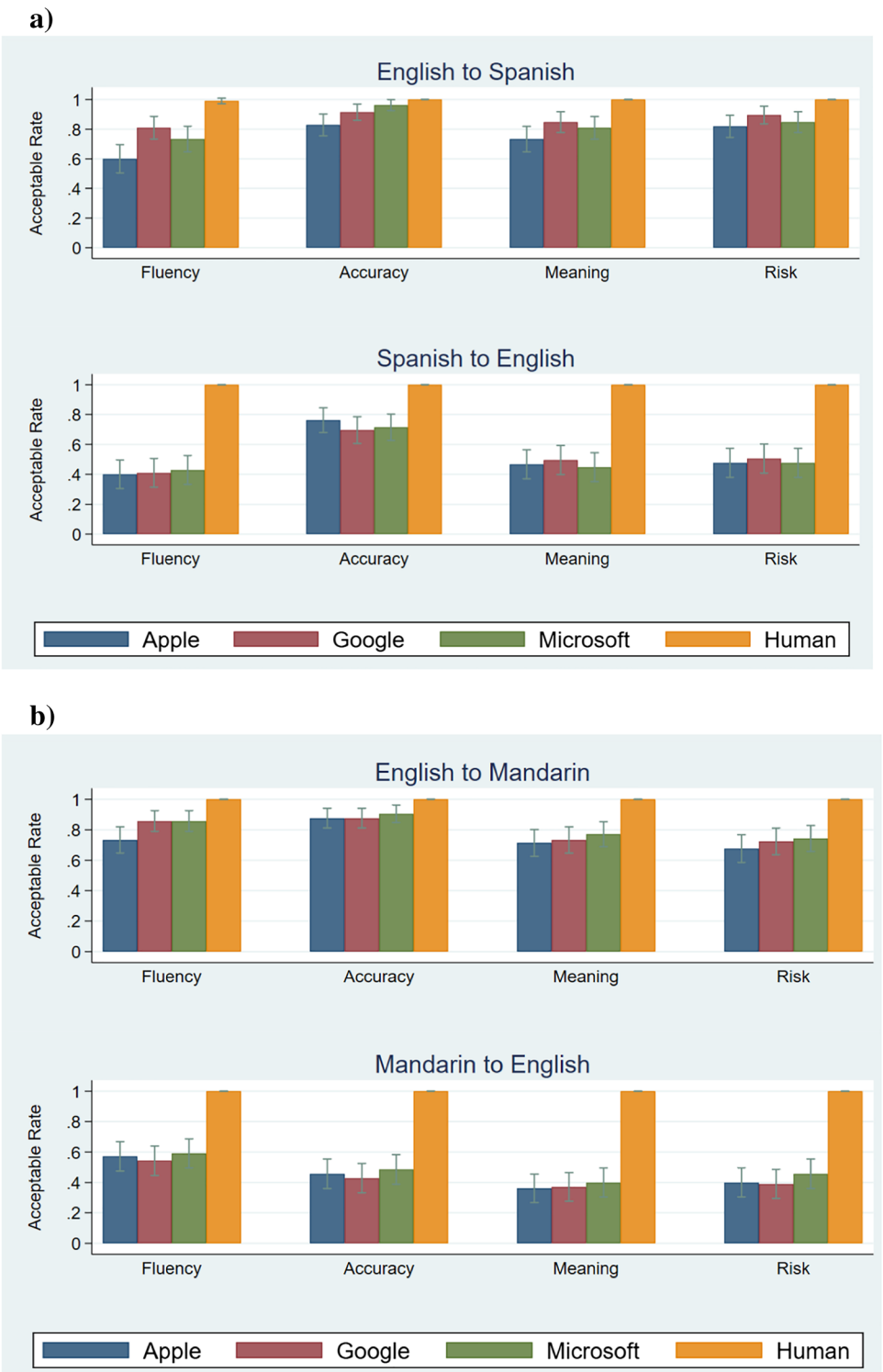

Figure 3 Proportions of interpreted phrases deemed acceptable (defined as score of 4 or greater on 5-point Likert scale) by individual assessment categories. a English–Spanish interpretation, b English–Mandarin interpretation

presents examples of the errors. Errors of syntactic parsing (i.e., word order and/or sentence structure issues) and differentiating statements from questions were common. Commonly used abbreviations sometimes posed challenges; while two MT applications correctly recognized "I.V." as "intravenous," one MT application understood it as "ivy," resulting in a significant error in the interpretation of the overall phrase.

DISCUSSION

In this study of three widely available MT applications, we found the overall quality of MT interpretation to be poor for two-way clinical communication use for conversations, even in low-stakes settings. In general, MT applications performed significantly better at interpreting English to Mandarin/Spanish than vice versa. All MT applications were inferior to professional human interpretation, and only English-to-Spanish interpretation using GT came close to meeting the non-inferiority threshold.

Previous studies have reported fewer Spanish MT inaccuracies compared to those of Chinese translations.[10,12] However, this study found similar quality for Spanish and Mandarin interpretations. As machine interpretation requires appropriate transcription and speech synthesis in addition to translation, challenges in either domain may have impacted the accuracy and quality of Spanish interpretation seen in this study. This may also explain the lower quality of MT interpretation from either Spanish/Mandarin to English than from English to other languages, as the current machine algorithm may be better adapted to handle English transcription than other languages with distinct inherent challenges in each language, such as tonation for Mandarin.[22]

All three MT applications performed poorly when interpreting phrases containing medical abbreviations, regardless of the direction of interpretation. This may be due to language ambiguity when using abbreviations, medical jargons, or uncommon phrases. Language ambiguity can influence pronunciations and connotations, thereby increasing the risk of improper interpretation.[23] In this study, MT had difficulty differentiating between "por que? (why)" and "porque (because)." Intonation and context would allow the human interpreter to distinguish between the two but may pose challenges for machines.

Disfluency (such as fillers, stutters, or pauses) may also impact MT interpretation. Examples of these fillers include "um," "well," and "you know," which professional interpreters would ignore, but MT applications may either incorporate them into their interpretation or stop the interpretation even before the statement was completed.[14] Anxiety is common among hospitalized patients, and communicative anxiety may generate a higher prevalence of language disfluencies.[24]

The results of this study should be interpreted in the context of its limitations. Although the order in which the human and the three MT interpretations were presented was randomized, the human voice clearly differs from MT audio outputs. The absence of established criteria to evaluate MT interpretation led us to adapt metrics created by the Advanced Research Project Agency (ARPA) for evaluating MT translation.[19] However, we did not test whether MT interpretations were comprehensible to patients. Comprehensibility, defined as the extent to which an interpretation is understandable, takes into consideration the fact that recipients may be able to infer the original content even if interpretation is deficient in lexical, grammatical, stylistic accuracy, or fluency. Performing a specific action following the interpretation of an instruction could serve as a reasonable test of MT comprehension.[25] Finally, in the real world, a person using MT applications would notice issues with MT interpretation (i.e., if the application stopped transcribing mid-sentence) and would repeat the statement using the visual cues provided by the applications.

The critical role of professional interpretation in healthcare is well documented. Executive Order 13166 mandates that federally funded healthcare institutions provide access to professional medical interpretation for patients with limited English proficiency.[1] Professional interpreters (compared to no interpretation) improve patient satisfaction, quality of care, many outcomes, and patient safety.[2] Hospital systems, several of which have undergone litigation related to patient

Table 2 Examples and types of interpretation errors

Types of error	Original phrase	Interpreted version
Omission	I need to pee right now. Where is the bathroom?	Where is the bathroom?
Abbreviation	We are going to connect you to some IV fluids to keep you hydrated	We are going to connect you to some ivy fluids *(líquidos de hiedra)* to keep you hydrated
Syntactic	You are just waking up. Relax and breathe	You are waking up, relaxing and breathing
Lexical	Why *(por qué)* does that machine keep beeping?	Because *(porque)* that machine keeps making noise
Nonsense Interpretation	Did I need any blood transfusions during surgery?	During the operation, I was not mathematically
Phonemic	Do you have any medicine in your bag?	Do you have any medicine in your back?

Errors are classified as omission, abbreviation (inability to accurate identify abbreviation), syntactic (word order and/or sentence structure), lexical (related to vocabulary), nonsense interpretation and phonemic (distinguishing one word from another, such as pad, pat, bad, and bat)

safety or quality of care events, also promote the use of professional interpretation.[3–6] Although this study compared MT interpretation to that of professional medical interpretation, we are aware that the most common alternative in low-stakes communication is, unfortunately, no interpretation at all. Nevertheless, our findings do not currently support a recommendation for use of MT interpretations in clinical settings. Instead, we encourage clinicians to use professional interpretation and advocate for hardware (speaker phones and video interpretation) in all settings, at least until MT improves significantly for two-way communication.[26]

In conclusion, three common MT programs demonstrated inferior quality in interpreting two-way verbal communication between English–Spanish and English–Mandarin, even in simple, brief encounters when compared to a professional medical interpreter. Until the quality of MT interpretation significantly improves, clinicians must ensure safe, effective, and equitable care by working with professional medical interpreters whenever possible.

Supplementary Information *The online version contains supplementary material available at https://doi.org/10.1007/s11606-023-08079-6.*

Acknowledgements *We want to thank Drs. Alex Chu and Elodia Caballero, Joshua Cristantiello, and Olga Maria Londoño for their contributions to the project.*

Corresponding Author: *Won Lee, MD, ScM; , University of California San Francisco, 513 Parnassus Ave, Room S-436, San Francisco, CA, 94143, USA (e-mail: Won.Lee@ucsf.edu).*

Funding *Research reported in this publication was supported by the National Research Service Award Institutional Research Training Grant (T32) of the NIH under Award Number T32GM008440 for Dr. Won Lee and by NHLBI K23HL157750 for Dr. Elaine Khoong. Dr. Fernandez and Ms. Rios-Fetchko efforts were partly supported by HRSA 5D34HP31878.*

Data Availability *Data available upon request.*

Declarations

Conflict of Interest *The authors declare no competing interests.*

Disclaimer *The content is solely the responsibility of the authors and do not necessarily represent the official views of the NIH or HRSA.*

REFERENCES

1. **Ngo-Metzger Q, Sorkin DH, Phillips RS, et al.** Providing high-quality care for limited english proficient patients: The importance of language concordance and interpreter use. J Gen Intern Med. 2007;22(SUPPL. 2):324-330. https://doi.org/10.1007/s11606-007-0340-z
2. **Manson A.** Language concordance as a determinant of patient compliance and emergency room use in patients with asthma. Med Care. 1988;26(12):1119-1128. https://doi.org/10.1097/00005650-198812000-00003
3. **Fernandez A, Schillinger D, Warton EM, et al.** Language barriers, physician-patient language concordance, and glycemic control among insured latinos with diabetes: The diabetes study of Northern California (DISTANCE). J Gen Intern Med. 2011;26(2):170-176. https://doi.org/10.1007/s11606-010-1507-6
4. **Diamond L, Izquierdo K, Canfield D, Matsoukas K, Gany F.** A Systematic review of the impact of patient–physician non-english language concordance on quality of care and outcomes. J Gen Intern Med. 2019;34(8):1591-1606. https://doi.org/10.1007/s11606-019-04847-5
5. **Chandrashekar P, Zhang R, Leung M, Jain SH.** Impact of patient-physician language concordance on healthcare utilization. J Gen Intern Med. 2022;37(8):2120-2122. https://doi.org/10.1007/S11606-021-06998-W/TABLES/2
6. **Schulson LB, Anderson TS.** National estimates of professional interpreter use in the ambulatory setting. J Gen Intern Med. 2022;37(2):472-474. https://doi.org/10.1007/s11606-020-06336-6
7. **Patel DN, Wakeam E, Genoff M, Mujawar I, Ashley SW, Diamond LC.** Preoperative consent for patients with limited English proficiency. J Surg Res. 2016;200(2):514-522. https://doi.org/10.1016/J.JSS.2015.09.033
8. Commonwealth of massachusetts board of registration in medicine quality and patient safety division. 2016:1–7. https://www.mass.gov/doc/july-2016-clinical-translation-advisory/download. Accessed April 27, 2022.
9. **Dew KN, Turner AM, Choi YK, Bosold A, Kirchhoff K.** Development of machine translation technology for assisting health communication: A systematic review. J Biomed Inform. 2018;85:56-67. https://doi.org/10.1016/J.JBI.2018.07.018
10. **Taira BR, Kreger V, Orue A, Diamond LC.** A Pragmatic assessment of google translate for emergency department instructions. J Gen Intern Med. 2021;36(11):3361-3365. https://doi.org/10.1007/S11606-021-06666-Z
11. **Rodriguez JA, Fossa A, Mishuris R, Herrick B.** Bridging the language gap in patient portals: an evaluation of google translate. J Gen Intern Med. 2021;36(2):567-569.https://doi.org/10.1007/s11606-020-05719-z
12. **Khoong EC, 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(4):580-582. https://doi.org/10.1001/jamainternmed.2018.7653
13. **Khoong EC, Rodriguez JA.** A Research Agenda for using machine translation in clinical medicine. J Gen Intern Med. 2022;37(5):1275-1277. https://doi.org/10.1007/S11606-021-07164-Y
14. **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(2):422-429. https://doi.org/10.1136/BMJINNOV-2019-000347
15. U.S. Census Bureau. detailed household language by household limited english speaking status (B16002); from American Community Survey - 2019: 5 -year estimates . https://data.census.gov/cedsci/table?q=Language Spoken at Home&tid=ACSDT1Y2019.B16002&hidePreview=false. Accessed January 18, 2021.
16. Microsoft Translator launching neural network based translations for all its speech languages . Microsoft Translator Blog. https://www.microsoft.com/en-us/translator/blog/2016/11/15/microsoft-translator-launching-neural-network-based-translations-for-all-its-speech-languages/. Published November 15, 2016. Accessed January 10, 2023.
17. iTranslate uses neural networks for translations . iTranslate Blog. https://blog.itranslate.com/machine-learning/itranslate-uses-neural-networks/. Accessed January 10, 2023.
18. **Caswell I, Liang B.** Recent advances in google translate. Google Research. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html. Published June 8, 2020. Accessed January 10, 2023.
19. **White JS, O'Connell T, O'Mara F.** The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proc 1994 Conf Assoc Mach Transl Am. 1994:193–205. http://mt-archive.info/AMTA-1994-White.pdf.
20. **Moorkens J, Castilho S, Gaspari F, Doherty S.** Translation quality assessment from principles to practice. 2018.

21. **Napoles AM, Santoyo-Olsson J, Karliner LS, Gregorich SE, Perez-Stable EJ.** Inaccurate language interpretation and its clinical significance in the medical encounters of spanish-speaking latinos. Med Care. 2015;53(11):940-947. https://doi.org/10.1097/MLR.0000000000000422
22. **Yuan J.** Perception of intonation in Mandarin Chinese. J Acoust Soc Am. 2011;130(6):4063. https://doi.org/10.1121/1.3651818
23. **Sproat R, Black AW, Chen S, Kumar S, Ostendorf M, Richards C.** Normalization of non-standard words. Comput Speech Lang. 2001;15:287-333. https://doi.org/10.1006/csla.2001.0169
24. **Bergmann G, Forgas JP.** Situational variation in speech dysfluencies in interpersonal communication. Lang Soc Situations. 1985:229–252. https://doi.org/10.1007/978-1-4612-5074-6_13
25. **Kapoor R, Truong AT, Vu CN, Truong D-T.** Successful verbal communication using google translate to facilitate awake intubation of a patient with a language barrier. A A Pract. 2020;14(4):106-108. https://doi.org/10.1213/xaa.0000000000001158
26. **Khoong EC, Fernandez A.** Addressing gaps in interpreter use: time for implementation science informed multi-level interventions. J Gen Intern Med. 2021;36(11):3532-3536. https://doi.org/10.1007/S11606-021-06823-4