VIEWPOINT

A Research Agenda for Using Machine Translation in Clinical Medicine

Elaine C. Khoong, MD MS[1,2] *and Jorge A. Rodriguez, MD*[3]

[1]Division of General Internal Medicine at Zuckerberg San Francisco General Hospital, Department of Medicine, UCSF, San Francisco, CA, USA; [2]UCSF Center for Vulnerable Populations at Zuckerberg San Francisco General Hospital, San Francisco, CA, USA; [3]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

J Gen Intern Med 37(5):1275–7
DOI: 10.1007/s11606-021-07164-y

Provision of linguistically appropriate care remains a challenge to achieving health equity. Language barriers impact 25.6 million limited English proficient (LEP) individuals in the USA. LEP patients experience worse healthcare access, quality, and outcomes, partly because systems frequently fail to engage patients in their preferred language.[1] Though interpreters are essential for language-discordant communication, they are underused[2] and may be unavailable in under-resourced settings or during times of increased demand, such as during a pandemic. Consequently, clinicians may rely on ad hoc interpreters or defer interpretation. Barriers to interpreter use are compounded by limited language diversity among clinicians. Ultimately, inadequate language access has led to inequities that require urgent attention.

Machine translation (MT) presents a tempting solution. MT refers to software that translates (text) or interprets (speech) from one language to another. These tools are available at low cost through websites and mobile apps, making them pragmatic resources for clinicians, particularly when certified (or even ad hoc) translators or interpreters are not available. Though not formally quantified, the use of MT tools in clinical care is likely frequent, reflecting this convenience, and has prompted safety concerns.[3] While the accessibility of MT tools suggests it may solve the challenge of inadequate language resources, the limited real-world healthcare evaluation of MT has prevented broader uptake. While there is preliminary evidence supporting MT accuracy for translating public health information, discharge instructions, and patient portal messages,[4–6] there is scant evidence on MT use for interpretation.[7]

In this paper, we propose an agenda to harness the potential of MT to improve clinical care by expanding research along four domains: communication scenarios, populations, machine translation algorithms, and outcomes (Fig. 1). Our proposed agenda advocates for research that explores the risk of MT use in various clinical scenarios, increases the diversity in training and evaluation of algorithms, compares performance variation among different MT algorithms, and expands the outcomes used in MT evaluations.

Prior Presentations: *none*
Received March 19, 2021
Accepted September 24, 2021
Published online February 7, 2022

COMMUNICATION SCENARIOS: EVALUATE VARIOUS INTERACTIONS

Healthcare contains a wide range of communication complexity. Consider the verbal and written communication that occurs for an office visit—from scheduling and confirming the visit to the check-in process to the patient-clinician interaction to the post-visit follow-up tasks. Extant MT research has only focused on a narrow scope of healthcare-related communication: written communication (e.g., patient education, discharge instructions, and portal messages). Future research should include assessment of MT for interpretation during real-time, synchronous communication. Although written communication improves patient understanding, language access for verbal communication is necessary. Several companies have advertised their software's ability to interpret live interactions, but the accuracy of interpretation for health-related interactions is unknown.

Moreover, within clinical medicine, there is wide variability in the risks of miscommunication. If miscommunication occurs during acquisition of consent for a procedure, there is potential for severe harm. Few clinicians would use machine translation tools to acquire consent. However, healthcare has many lower-risk interactions, such as when inpatients inquire about basic needs (e.g., water, toileting) or outpatients schedule appointments. Patients with language barriers experience friction at these encounters. MT could supplement the interaction in these situations, particularly since patients with language barriers are less likely to interact with the healthcare team.

These lower-risk interactions often involve everyday conversations that are likely well represented in training data for commercial MT tools. We hypothesize these tools are likely to accurately interpret/translate these encounters in comparison

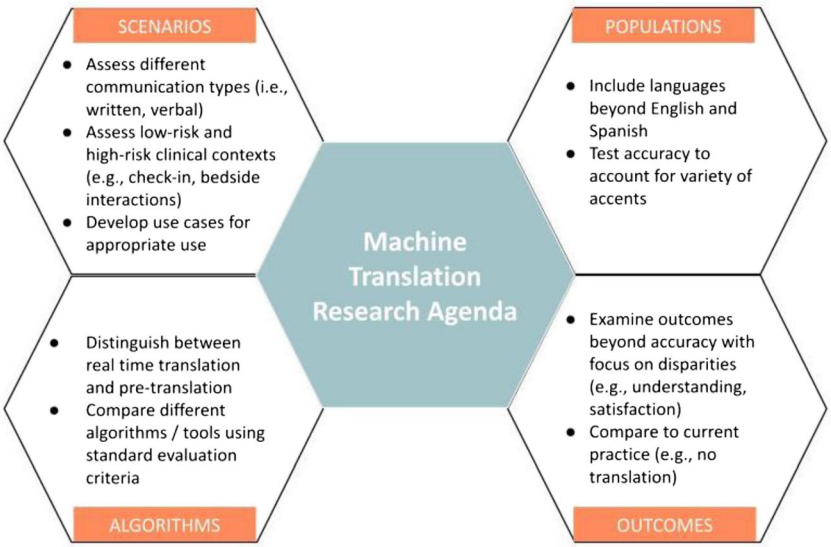

Figure 1 A research agenda for machine translation in healthcare.

to interactions replete with medical terminology. Research may show that MT is inadequate when communicating medically focused content, but for non-healthcare-focused communication, MT provides value.

POPULATIONS: FOCUS ON EQUITY FOR LINGUISTICALLY DIVERSE POPULATIONS

Prior research has also focused on translation between English and Spanish.[4] Although Spanish is the most common non-English language in the USA, other languages, including Vietnamese and Korean, are underrepresented in the healthcare work force.[8] Studies must evaluate other languages, as accuracy is often worse.[5,6] Globally, many migrants emigrate to locations where English is not the dominant language; therefore, researchers should also evaluate MT use between non-English languages.

Beyond language diversity, we need to evaluate the impact of accents. English may be spoken with an accent among native- and foreign-born English speakers. The variable experience of using digital assistants among fluent English speakers demonstrates this bias in technologies.[9]

ALGORITHMS: ASSESS MULTIPLE MACHINE TRANSLATION ALGORITHMS

We suspect there is variation in the quality of MT tools. Most studies have evaluated Google Translate, but other tools exist. Future evaluations should include multiple algorithms to clarify if any perform better. Establishing standard evaluation criteria will facilitate the comparison of tools.

It is equally important to distinguish MT tools used for real-time translation through machine learning from tools that present pre-translated phrases. These latter products include mobile applications (e.g., Canopy Speak) that contain phrase libraries with healthcare-related phrases. One could select the phrase "I am a doctor"; then, the app would provide the written translation and read the phrase in the desired language. Pre-translated phrase libraries may have less flexibility but can guarantee accuracy. Much like we compare different chronic disease medications, we should explicitly compare different tools.

OUTCOMES: EXPAND OUTCOME EVALUATION

Most research has evaluated only accuracy. While this outcome is important, the goal of harnessing MT for healthcare is to improve the quality of care for patients with language barriers. With this mindset, evaluations need to include outcomes with known importance, such as clinical, patient-reported, and utilization outcomes. By focusing on known disparities experienced by LEP patients, we can evaluate if MT will reduce inequities.

A limitation of prior studies is the evaluation of accuracy against a certified medical interpreter. While there is value in using this comparison, in real-life clinical practice, certified medical interpreters are underutilized,[2] and patients frequently receive no written communication in their preferred language. We believe researchers must evaluate MT against current practice, recognizing that practice patterns vary. For example, rather than evaluating the impact of machine-translated discharge instructions against certified translated discharge instructions, studies should evaluate machine-translated instructions against usual practice, which frequently is English-only instructions. Similarly, the use of MT to help with low-risk bedside interactions should be compared against usual care, which may include non-verbal communication, ad hoc interpreters, and/or not asking patients about their concerns.

A PATH TO IMPLEMENTATION

Advancing our proposed agenda requires cross-disciplinary collaboration, clarification on legal considerations, adequate funding, and research standards. First, MT research must include patients, researchers, industry leaders, and MT experts to identify the combination of clinical scenarios, patient populations, algorithms, and/or outcomes that hold the most promise to improve care. This multidisciplinary group will also facilitate inclusion of diverse perspectives on determining research questions that maximize benefit and limit safety concerns. Second, research must develop in conjunction with a nuanced discussion about the legal issues surrounding use of MT and how adverse events should be addressed; these conversations should be integrated into the broader discussion of the liability of artificial intelligence in healthcare. Third, private and public funders should establish funding mechanisms that incentivize rigorous evaluation of all the potential uses of MT in healthcare. Aligned incentives will allow academic-industry partnerships that foster innovation and focus development of these digital tools for underserved populations.[10] Finally, to ensure research endeavors build on each other, multidisciplinary stakeholders must develop evaluation standards and performance benchmarks to facilitate comparisons of findings from multiple studies.

CONCLUSION

Prior healthcare-focused machine translation research has evaluated a smattering of areas without a cohesive vision to advance the field. As practicing clinicians, researchers, informaticists, and advocates for language access, we believe that although there is excitement for MT tools to reduce inequities for patients with language barriers, we need more definitive evidence on its benefits or harms. We believe our proposed research agenda focused on appropriate scenarios, diverse populations, multiple algorithms, and expanded outcomes will ensure the progress of this potentially valuable field. Without advancements in each of these domains, we will be unable to bring the promise of MT into reality.

Contributors: *None*

Corresponding Author: *Elaine C. Khoong, MD MS; Division of General Internal Medicine at Zuckerberg San Francisco General Hospital, Department of Medicine, UCSF, San Francisco, CA, USA (e-mail: elaine.khoong@ucsf.edu).*

Funding *Dr Khoong is supported by the National Heart Lung and Blood Institute of the NIH under Award Number K12HL138046 and K23HL157750. Dr. Rodriguez is supported by the National Institute on Minority Health and Health Disparities of the NIH under Award Number K23MD016439.*

Declarations:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. **Diamond LC, Jacobs EA, Karliner L.** Providing equitable care to patients with limited dominant language proficiency amid the COVID-19 pandemic. *Patient Education and Counseling*. 2020;103(8):1451-1452. https://doi.org/10.1016/j.pec.2020.05.028
2. **Schulson LB, Anderson TS.** National Estimates of Professional Interpreter Use in the Ambulatory Setting. *Journal of General Internal Medicine*. Published online November 2, 2020:1-3. https://doi.org/10.1007/s11606-020-06336-6
3. Commonwealth of Massachusetts Board of Registration in Medicine Quality and Patient Safety Division. *Clinical Translation Advisory.*; 2016:1-7. https://www.mass.gov/doc/july-2016-clinical-translation-advisory/download
4. **Dew KN, Turner AM, Choi YK, Bosold A, Kirchhoff K.** Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics*. 2018;85:56-67. https://doi.org/10.1016/j.jbi.2018.07.018
5. 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 Internal Medicine*. 2019;179(4):580. https://doi.org/10.1001/jamainternmed.2018.7653
6. **Taira BR, Kreger V, Orue A, Diamond LC.** A Pragmatic Assessment of Google Translate for Emergency Department Instructions. *Journal of General Internal Medicine*. Published online March 5, 2021. https://doi.org/10.1007/s11606-021-06666-z
7. **Panayiotou A, Gardner A, Williams S, et al.** Language Translation Apps in Health Care Settings: Expert Opinion. *JMIR Mhealth Uhealth*. 2019;7(4):e11316. https://doi.org/10.2196/11316
8. **Diamond L, Grbic D, Genoff M, et al.** Non–English-Language Proficiency of Applicants to US Residency Programs. *JAMA*. 2014;312(22):2405. https://doi.org/10.1001/jama.2014.15444
9. **Rangarajan S.** Hey Siri—Why Don't You Understand More People Like Me? Mother Jones. Published 2021. https://www.motherjones.com/media/2021/02/digital-assistants-accents-english-race-google-siri-alexa/
10. **Lyles, Courtney, Horn, Ivor, Sarkar, Urmimala**. In Digital Health, Partnerships Between Business And Academia Are Needed To Advance Health Equity. Health Affairs Blog. Published April 16, 2021. https://www.healthaffairs.org/do/10.1377/hblog20210413.13025/full/