Reliable and Safe Use of Machine Translation in Medical Settings

NIKITA MEHANDRU, University of California, Berkeley, USA
SAMANTHA ROBERTSON, University of California, Berkeley, USA
NILOUFAR SALEHI, University of California, Berkeley, USA

Language barriers between patients and clinicians contribute to disparities in quality of care. Machine Translation (MT) tools are widely used in healthcare settings, but even small mistranslations can have life-threatening consequences. We study how MT is currently used in medical settings through a qualitative interview study with 20 clinicians–physicians, surgeons, nurses, and midwives. We find that clinicians face challenges stemming from lack of time and resources, cultural barriers, and medical literacy rates, as well as accountability in cases of miscommunication. Clinicians have devised strategies to aid communication in the face of language barriers including back translation, non-verbal communication, and testing patient understanding. We propose design implications for machine translation systems including combining neural MT with pre-translated medical phrases, integrating translation support with multimodal communication, and providing interactive support for testing mutual understanding.

Additional Key Words and Phrases: Machine Translation in Medicine, Reliability of AI Systems, Clinical Decision-Making

ACM Reference Format:
Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3531146.3533244

1 INTRODUCTION

The risks and harms of machine translation (MT) technologies disproportionately fall on vulnerable people who depend on it for access to healthcare, employment, and social support [34, 50, 55]. In recent years, researchers have studied the dangers of NLP technology, including bias and environmental impact [2, 4, 25, 31, 42, 45, 47, 48]. We build on this work and study how and when MT systems might lead to miscommunication and misinformation, particularly in high-stakes scenarios. People usually use MT because they do not know the source or target language, which makes MT very difficult to evaluate in practice. Mistranslations can cause frustration, conversational breakdowns, and even human rights violations [5, 23, 57, 58].

In this paper, we focus on one high-stakes context where people rely on MT: healthcare. Patient-clinician communication is a crucial aspect of providing healthcare, and can be negatively impacted in the presence of language barriers, which contributes to disparities in quality of care. MT is used by many clinicians in the U.S. as a low-cost and efficient way to communicate with patients, but reliability varies. One study found that common medical discharge information was incorrectly translated by Google Translate 8% of the time for Spanish and 19% for Chinese and that 2% of those translations could cause clinically significant harm for Spanish and 8% for Chinese [27]. Another study found that only 45% of common medical phrases were correctly translated to two African languages [41]. Therefore, the continued use of MT in medical settings poses great risk, particularly to vulnerable people [55]. Learning whether and how MT might

Manuscript submitted to ACM

be used reliably in these settings requires first understanding how clinicians currently navigate language barriers with patients.

We conducted 20 semi-structured interviews with clinicians in the U.S. across 7 specialties who regularly face language barriers with patients. Our goal was to understand how clinicians currently navigate language barriers, including when and why they use MT, if at all; what concerns they have about using such systems; and how they assess the quality of translations, whether by working with a medical interpreter or using an MT tool. We conducted interpretive data analysis and open coding to analyze the data.

We found that in the presence of language barriers, clinicians faced challenges stemming from lack of time and resources, cultural barriers, medical literacy, and accountability in cases of miscommunication. We found that machine translation systems save clinician time and aid medical interpreters in providing translation services, particularly for less widely spoken languages. In such cases, clinicians leveraged MT tools to speak with patients and also provided written discharge instructions in the patient's primary language. Medical interpreters played the role of facilitating cross-lingual communication, but medical literacy rates drove patient understanding and cultural differences impact patient-clinician trust. These three areas can serve to exacerbate disparities for the most vulnerable populations if not accounted for. We found that clinicians have navigated these challenges by employing a variety of tactics including back-translations, non-verbal communication, and testing patient understanding. We propose design implications of machine translation systems that can help reduce the patient-clinician communication gap and discuss the ethical implications of MT used in high-stakes situations.

Based on these findings, we identify opportunities for language technology to more reliably support cross-lingual patient-clinician communication. First, general purpose MT systems fall short of meeting clinician's needs, therefore systems need to be developed that are highly accurate when translating medical language. One avenue can be systems that combine pre-translated dictionaries with neural MT. Second, the development of such systems can benefit from moving beyond one-to-one mapping of phrases, to recognizing the context of and reciprocal nature of communication. Finally, we argue that any tools such as MT that are deployed in medical settings should go through rigorous evaluation (RCTs) and endorsement by a relevant governing body.

2 RELATED WORK

In this work, we expand on three major areas of related work: cross-lingual patient-clinician communication, reliability metrics in AI systems, and machine translation system deployment in medical settings.

2.1 Patient-Clinician Communication

Language barriers can cause miscommunication between patients and clinicians, impeding the quality of care a patient receives. It can also lead to lower satisfaction levels for clinicians and patients alike. Non-English speaking patients can often be less satisfied with the care they receive compared to their English-speaking counterparts, and are typically less inclined to return to a medial setting if a health problem persists [11].

Medical professionals have also expressed that language barriers can be an obstruction to providing quality patient care and serve as a source of workplace stress[6]. The effects tend to be more pronounced in nurses than physicians, with the former reporting higher levels of stress when interacting with patients in the presence of language barriers.

In addition to negatively impacting quality of patient care, language barriers can also increase healthcare costs. Increased allocation of resources and emergency department durations have been associated with non-English-speaking patients compared to English-speaking ones [22].

Existing work on AI in clinical settings has focused primarily on clinical decision-making such as how AI can be leveraged to assist physicians in diagnoses or treatment recommendations [51]. We instead focus on how AI can assist clinicians in communicating with patients, which renders a different set of priorities.

2.2 Measuring Reliability in AI Systems

A major barrier to the responsible application of machine learning in real-world domains is a lack of clear guidelines for how to ensure safety and reliability. This problem is especially pronounced at the intersection of ML and healthcare, where the stakes of failure are high and evaluation standards are misaligned across the two fields. As a result, it can be difficult for end-users of ML systems to judge when and how they can use those systems reliably.

One way to promote users' trust in ML systems is to subject them to rigorous evaluations and communicate the results to users [32]. However, it is not yet clear how this should be done. A growing body of work has demonstrated the shortcomings of standard evaluation practices in machine learning, particularly with respect to measuring how those systems will perform in the real world. For example, a model's performance may be inflated if it was evaluated on a test set that is not representative of the population with which the system will be deployed [7], or without attending to the human and contextual factors that shape how it is used in practice [3]. These issues put standard ML evaluation practices in stark contrast with the more established and rigorous traditions for evaluating medical devices. To bridge this gap, researchers have begun to develop specific guidelines for evaluating AI systems for use in healthcare, particularly for diagnostic algorithms [13, 32, 35].

While communicating overall system performance may increase adoption and trust, even very high performing ML systems will occasionally make inappropriate or incorrect predictions. Overreliance on ML models could therefore lead to harm in high-stakes settings like medicine [19]. This is one motivation for research in the field of explainable AI (XAI), which promotes a set of techniques that can explain predictions made by AI systems in an interpretable, intuitive way [14]. Related approaches include implementing human oversight of ML-driven decision-making or designing for human-AI collaboration to leverage human expertise to compensate for a model's limitations [1, 8]. Despite substantial effort in these areas, it remains difficult to design systems that actually protect users from erroneous predictions. There is an assumption that the end-user has some domain knowledge or expertise that they can rely on, possibly assisted by some explainability intervention, to assess the quality of predictions and come up with an alternative when necessary [19, 51]. However, machine translation is a case where the user often does not have the relevant expertise, i.e. language ability, to do so.

2.3 Machine Translation Systems in Medicine

There is growing evidence that clinicians and other healthcare workers use free online services like Google Translate as a last resort when no other language services are available [54]. In 2013, Turner et al. surveyed local health departments and found that almost a third had used MT to translate written materials because they lacked the budget for professional translations [53]. Clinicians have also successfully used MT to communicate directly with patients, particularly in urgent situations when they had exhausted other alternatives [26, 33, 38, 60].

Unfortunately, clinicians are faced with vague and conflicting guidance about how to mitigate these risks. Clinicians have been warned to exercise caution due to the risk of miscommunication [37, 49], but it is unclear how they can do so effectively besides avoiding the tools altogether. Researchers have suggested that clinicians be skilled in cross-cultural communication and rely on non-verbal cues to avoid miscommunication [43]. One study found that using simpler language and checking for spelling and grammar mistakes can improve translation quality [27]. Due to the risk of

mistranslations with open-ended MT systems, some healthcare professionals and patients have expressed a preference for phrase-based translation applications, which limit what users can communicate, but are more reliable because all of the available phrases have been professionally translated [40, 46, 52].

These issues point to an opportunity for language support systems designed with the specific needs and challenges of healthcare communication in mind. Some prior research has proposed specialized systems for medical translation [17, 44], but few of these systems have moved beyond pilot studies [16]. Our goal in this work is to understand the major challenges that clinicians face when communicating with patients across language barriers, how they overcome these challenges, including whether and how they use MT tools, and where those strategies fall short. This need-finding work starts from an understanding of clinicians' existing practices and strategies to identify paths forward for language support tools that are safe, reliable, and useful.

3 METHODS

Our goal in this research was to understand how clinicians approach language barriers with patients, with a focus on how, if at all, clinicians use machine translation in their current workflow. Towards this goal we conducted semi-structured interviews with 20 healthcare clinicians across 7 specialties.

3.1 Recruitment & Interview Procedure

We conducted in-depth interviews with clinicians across the United States to better understand if and how they are using machine translation tools in medical settings. We ran a pilot with a few physicians to pre-test our questions and identify potentially ambiguous wording as well as opportunities to add more questions. We found providing an example of a machine translation system, specifically Google Translate, was helpful so clinicians understood the scope of the interview study and could reflect how they interacted with this tool.

We recruited 20 clinicians using snowball sampling and conducted semi-structured interviews via a video conferencing tool. The interview questionnaire consisted of two background questions inquiring about the medical specialization of the clinician and the number of years the clinician has been in practice (post-residency for physicians), five open-ended questions around how, if at all, the clinician has interacted with an MT tool, and four demographic questions.

During the interview, we asked about experiences where they faced a language barrier with a patient and how they navigated the situation, when, if at all, they had used a machine translation tool and why, and how they assessed the quality of a translation from a medical interpreter versus a machine translation tool. We also asked questions around the challenges of providing care to vulnerable populations, how technology is currently integrated into their workflows, and how language barriers interfere with both verbal and written communication. The interviews lasted between 20-45 minutes. All interviews were recorded and transcribed. This study was approved by our Institutional Review Board (IRB) and all clinician participants were compensated for taking the time to participate in our study during the ongoing COVID-19 pandemic.

Our sample includes physicians, surgeons, nurses, and midwives in the United States across the following specialties: cardiology (1), orthopedic surgery (1), nephrology (1), family medicine (8), obstetrics and gynecology (7) , trauma surgery (1), and emergency medicine (1). Clinicians came from a range of settings and institutions including private practice, county hospitals, community hospitals, and academic institutions.

3.2 Data Analysis

We transcribed our interview recordings and used interpretative qualitative coding. Our open-ended inductive analysis drew on elements from grounded theory methodology [12]. The first author conducted interpretative qualitative coding, and all authors discussed emerging themes regularly. The first author then used axial coding and identified high-level themes across the codes. Some examples can be found below:

- Language Barriers: How do language barriers influence verbal and written patient instructions, and ultimately, affect the quality of patient care?
- Measuring quality of translation: What methods do clinicians employ to measure patient understanding in the presence of language barriers?
- Challenges: What challenges do clinicians face with collaborative care approaches in vulnerable patient populations?
- Machine Translation Tools: How, if at all, are clinicians leveraging machine translation tools and how do they assess the quality of such translations?
- Medical Settings: How do clinicians adjust how they navigate language barriers with patients in low- versus high-stakes medical settings?

4 RESULTS

We found that clinicians faced communication challenges stemming from lack of time and resources as well as cultural barriers and medical literacy. They were also concerned with the accuracy of translations and accountability measures. When faced with a language barrier, clinicians relied on a combination of medical interpreters, Machine Translation, and their own knowledge of the other language. In response to communication challenges, clinicians had developed strategies to evaluate cross-language communication which include: back-translation, non-verbal communication, and testing patient understanding. These strategies can offer insight and paths forward to developing reliable MT for medical settings which we discuss in the next section.

4.1 Communication Challenges Across Languages

In this section we describe the challenges that clinicians face in cross-lingual communication with patients.

4.1.1 Limited time and resources. We found that constraints on time and resources interfere with interlingual patient-clinician communication and adversely affect the quality of care a clinician is able to provide to a patient. These challenges are exacerbated in low-resource medical settings.

Medical interpreters are certified translators that clinicians solicit when a language barrier with a patient arises. While this is the gold-standard for language support, the process of calling an interpreter can be incredibly time-consuming. The scarcity of time often prevents clinicians from calling a language translation service:

> *You just have one or two questions and [they're] not too lengthy, and it's not worth calling, or [taking] the time to call the interpreter.* (P2, Nephrologist)

In situations where medical interpreters were necessary, many clinicians expressed how an interpreter could make patient visits twice as long due to mistranslations. Due to back to back patient consults, clinicians are often short on time and as a result have to prioritize what they ask their patients. The act of tracking down an on-site interpreter or calling a language translation line took time, as did facilitating communication with the patient via an interpreter. The

time of patient-clinician interactions drastically increased in the presence of a language barrier, and clinicians described how they had to prioritize what questions to ask after taking a patient's history.

> *These are patients who, unfortunately, aren't given the amount of time that they deserve. They are maybe given a 20 minute appointment with me when they really should have double that time.* (P8, Obstetrician-Gynecologist)

Clinicians described that they often cannot call an interpreter due to resource-constrained settings, the time of day, or because the patient speaks a specific dialect or less widely spoken language. In such cases, clinicians have to reschedule appointments with those patients or settle for imperfect means of communicating with patients.

There was widespread recognition among clinicians about the flawed nature of communication with patients who did not speak English. These patients were often either scheduled on different days when there was more clinician availability or were given standard time that never proved long enough for a thorough consult.

4.1.2 Cultural barriers and medical literacy. We found that patient-clinician communication was also adversely affected in settings where patients had low medical literacy rates, particularly among non-English-speaking populations, and in instances where there were cultural barriers between patients and clinicians. Many clinicians noted that navigating low medical literacy was a challenge in and of itself that was exasperated by language barriers:

> *Medical literacy is so low, so patients come back all the time and say, "Oh, I didn't understand what you had said." or, "Nobody told me this." When clearly I told them that even when there's no language barriers [...] they really just don't understand instructions regardless of the language they're given in.* (P3, Obstetrician-Gynecologist)

When low medical literacy rates were compounded by cultural differences in the context of a language divide, the quality of care a patient receives drastically decreased. Low medical literacy rates among English-speaking populations typically resulted in patients missing medication refills, or returning to medical settings with follow-up questions. In non-English-speaking populations low medical literacy posed additional challenges as clinicians had trouble deciphering if the miscommunication was attributed to a mistranslation or a lack of knowledge about medical jargon in the patient's native language.

Furthermore, we found that cultural barriers between clinicians and patients also interfered with methods clinicians used to test patient understanding. The pervasive 'yes' culture is an example of one these barriers that has negative downstream effects on patient-clinician care. Clinicians observed this phenomenon in patients regardless of the presence of a language barrier. However, the effects were more pronounced with patients from non-English-speaking populations who had a tendency to tell physicians they understood their patient plan moving forward but would then express doubt during interactions with nurses. The authoritative presence of a physician played a possible role in the patient feeling uncomfortable asking questions.

> *I think that's especially true culturally with power dynamics with hierarchy. And I think that if someone present is either an immigrant or undocumented there's so many factors there. But it's also true for English speaking patients too, they say, "Yes, yes, yes, yes."* (P7, Certified Nurse-Midwife)

The culture of signaling understanding can be especially pronounced when it manifests in vulnerable populations [24, 29]. Although physicians, and clinicians more generally, worked to foster a safe space for patients to ask questions, many patients wanted to be mindful of clinician time or did not know how to communicate their concerns with clinicians, particularly if they could not speak English.

Some clinics have sought to address these limitations by employing cultural navigators. Distinct from medical interpreters, these individuals provide context on the culture the patient comes from and hold credibility because they usually share the patient's background. Typically, these navigators belong to the same communities as their patients, and often times already know them. As one clinician described:

> *[Cultural navigators] know the backgrounds of the patients or at least their cultural beliefs and where the patient's coming from or what they believe in. And that helps us bridge that gap of number one, language barrier and number two, just cultural differences that we may perceive differently.* (P3, Obstetrician-Gynecologist)

Understanding how certain health topics are stigmatized in some communities is an important part of delivering high quality care. A physician working at a county hospital described how a patient with limited medical knowledge declined using a medical interpreter due to cultural taboos around the reason of her visit:

> *We're talking like Orthodox Christianity here and trying to talk about things like contraception or take a sexual history is just nightmare. I had one situation where [. . .] this Russian speaking patient was diagnosed with HIV in pregnancy and she didn't want anybody to know. But that's something that I have to have a conversation with her [about]. She's okay with me knowing obviously. I diagnosed it, but she didn't want anyone [else to know] Her husband had to know, but she didn't want the rest of her family knowing. So I couldn't even explain some of the things that I was doing during her delivery. Like I couldn't explain that I was giving her antiretrovirals, that I'm going to take the baby away and give them antiretrovirals, that I'm going to have to do a bunch of blood draws on the baby. Just because there was no interpretation.* (P6, Family Medicine Physician)

Navigating cultural barriers proved to be difficult for many clinicians in our study who often had to operate with limited information in less than ideal circumstances. Trust played a crucial role in what was communicated to clinicians.

> *Women's health, sexual health, significant disability, memory issues. People are very comfortable universally with musculoskeletal complaints and things like that. But a lot of the things people see me for are things that their families might not even know. And so that's really hard to get out of people.* (P6, Family Medicine Physician)

4.1.3 Accuracy and accountability. We found that when faced with language barriers, clinicians preferred to rely on certified medical interpreters when they had access to one, had enough time, and for higher stakes conversations. But they tended to resort to MT when that was not feasible or seemed too costly given the circumstances. In both cases, clinicians worried about the accuracy of the translation, particularly when they didn't know the other language. Many expressed concern and frustration with the variance in quality of medical interpreters:

> *A lot of our patients speak Spanish and I understand a lot of Spanish [. . .] But when they're doing my translations, there are times that I have to correct them [. . .] that does worry me sometimes, because languages that I don't understand, I don't know what the patients are being told.* (P3, Obstetrician-Gynecologist)

In addition to the accuracy of translations, clinicians worried about accountability. In high-stakes medical settings, defined by clinicians as instances where patient consent was needed, calling an interpreter was viewed as mandatory. Though there was no legal requirement to do so, having an interpreter translate and verify patient consent provided a protective layer of accountability that an MT system would not.

> *If she had to actually sign any papers, then you need to have a medical interpreter. They need to have it written and be able to fully understand everything [. . .] And then unfortunately we also had to call her husband who*

> *wasn't able to come to the clinic because of COVID restrictions. So he was on her cell phone in like a three-way call with me, her, and him. And then we also had the interpreter. So that was very challenging, because you can imagine that would take a really long time to go through the surgery [consent] with that type of barrier.* (P1, Obstetric and Gynecological Surgeon)

The question of accountability in the event of a mistranslation from an MT also gave some clinicians pause. Generally, MT were a last-resort option.

> *Usually, when I'd use it [it] is out of sheer desperation. So often it was more rare languages where there was no interpreter, we'd be on hold for 15 minutes, realized we were probably not going to get someone at 6:00 AM, while bedside rounding, and just used Google Translate to do the best we could to try to communicate in that setting.* (P19, Family Medicine Physician)

The recurring tradeoff from our interviews was the time it would take to call the translator versus how urgent of a medical situation the patient was in. The decision to call a translator was largely contingent on whether: a) they had baseline familiarity of the language the patient spoke, and b) how high-stakes the situation was.

Many (19 out of 20) clinicians resorted to Google Translate when they had familiarity with the language the patient spoke but were not medically certified in it. Thus, they used Google Translate to verify specific medical words.

> *I've looked at Google Translate to look up specific words, where I'm not totally sure that I'm using the exact correct words, for example, in Spanish, which is the language that I otherwise speak. But just to make sure that the vocabulary that I'm using is the most correct, but I have not used it to translate large swaths of speech or writing in a language that I otherwise don't speak at all.* (P4, Obstetrician-Gynecologist)

In low-stakes medical settings, clinicians typically settled for piecemeal conversation or Google Translate to navigate language barriers. Such situations included taking a patient's history, which has significantly less legal liability than obtaining patient consent for a surgery. Generally, medical environments that centered around simple questions did not warrant a medical interpreter:

> *I did use it (Google Translate) in residency rounding just to speak with patients when I couldn't access an interpreter for basic things like, are you hurting? Are you comfortable? Do you have nausea? Simple questions like that where there isn't quite as much legal risk if something's misunderstood.* (P19, Family Medicine Physician)

Clinicians recognized that Google Translate is an all-purpose technology and wasn't built to be deployed in medical settings. Thus, in high-stakes medical settings, many clinicians expressed that they knew they would be accountable for a mistranslation when using an MT system.

> *If it was very complex, where I needed patient's consent, probably would've used a translator, because I knew they would not be supported by Google. Nobody would back me up for the Google Translate.* (P2, Nephrologist)

A handful of clinicians had used translation systems, specifically Google Translate, before. While most clinicians preferred in-person interpreters to language lines, there were instances where no translation services were available.

> *I have used it in the past to communicate with a patient. It was actually a Russian speaking patient and there were no translators available. It was a Sunday in the hospital and I just couldn't get a hold of anyone on the weekend and I was trying to explain to him what was going on so I ended up just pulling out Google Translate and typing what I was trying to say to him, showing it to him and then he would try to say something back or type something back.* (P14, Family Medicine Physician)

Overall, we found that for clinicians, communication across language barriers was made extremely difficult by a lack of time and resources as well as cultural barriers and medical literacy issues. Clinicians also worried about the accuracy of translations and the legal liability mistranslations imposed. In the next section we discuss the strategies that clinicians used to communicate with patients under these conditions.

4.2 Clinician's Strategies for Communicating Across Languages

We found that clinicians used four main tactics to confirm the reliability of the communication between them and their patients: rephrasing and simplifying medical terms, back-translation, non-verbal communication, and testing patient knowledge. In this section we describe each in more detail.

4.2.1 Rephrasing and simplifying medical terms. A recurrent theme that we found among clinicians was using simple, non-medical terms when communicating with patients. Clinicians rephrased terms to improve translation accuracy. They did this with both medical interpreters, who may not have knowledge of complex medical jargon, and with machine translation systems, since clinicians were aware that tools like Google Translate were not trained using medical vocabulary.

> *Learning to just change your language, avoiding a lot of jargon, just saying things in plain English, [and not using] any medical terms, [but] explaining things in the easiest way possible.* (P14, Family Medicine Physician)

Therefore, learning when and how to simplify and rephrase language was a skill that clinicians developed to communicate better with patients across language barriers.

4.2.2 Back-translation. Some clinicians used back-translation methods where they asked the interpreter to repeat back what they had translated to the patient:

> *My personal practice is if I'm dealing with a high acuity situation where there is a lot of explanation and a deep amount of informed consent to be done with the patient, I actually ask the interpreter what they're actually telling the patient. Like I really grill the interpreter into how are you translating that? Can you tell it back to me? It's almost like confirm back as to what you did explain and how much of it did the patient get? So it's trying to do a closed looped communication after every two or three sentences. So summarizing as you go along rather than after 30 minutes.* (P5, Obstetrician-Gynecologist)

The use of back-translation for validation has become a standard in many multilingual medical settings [30], but has been critiqued for failing to consider issues of cultural adaptation [39].

4.2.3 Gestures, drawings, and non-verbal communication. Consistent with prior research [28], we found that clinicians used a variety of methods to improve communication with non-English-speaking patients including pen and paper drawings, visual aids, and gestures. Particularly when medical interpreters were not available, clinicians relied on writing to patients, providing visual aids, miming, and drawing pictures. One physician described the challenge of working with a deaf patient that used American Sign language:

> *We actually just wrote to each other the entire visit because our interpreter iPad was glitching and it was only working through audio, and the video wasn't working. So we're like, "This is besides the point." So we conducted her entire visit in our chicken scratch handwriting to each other.* (P7, Certified Nurse-Midwife)

Another physician described a more innovative tactic when the clinic she worked at did not have translation services for Marshallese. She used teach-back methods where she had the patient mime back what she was planning on doing:

> *So there have been visits where I have mimed or drawn pictures and then I make them mime back and draw a picture of their understanding just because we haven't been able to find an interpreter [which] is nuts.* (P6, Family Medicine Physician)

4.2.4 Social cues and teach-back methods. Communication is often thought of as a one-way transfer of information. However, the clinicians we spoke to viewed communication as more of a mutual production of meaning, and part of that includes testing understanding [21]. We found that clinicians relied on social cues and teach-back methods, asking a patient to repeat information back, to estimate the extent to which patients understood what was said to them.

Clinicians in our study described how they often relied on mental heuristics to determine if the interpreter was translating accurately including visual and social cues from the patient:

> *The ways to judge that [translation quality] for me is usually if I'm telling something that I'm expecting a certain reaction [to] and I don't get that reaction from the patient, then I wonder if they were actually told what I was trying to tell them. For example, if it's sad news or something and I want them to understand the level of seriousness of it.* (P3, Obstetrician-Gynecologist)

This same physician also described how she measured the interval of talking time between her and the interpreter and the interpreter and the patient to assess translation quality:

> *But the other way sometimes is if I say a very short sentence and the translation is going on for two minutes, then you're like, "Well, clearly this person said a lot more than I did."* (P3, Obstetrician-Gynecologist)

Some also described using body language and facial expressions of patients as signals of comprehension.

Clinicians recognized the need to combine simple phrases with teach-back methods to test patient understanding.

> *At the end of all my interactions, I'll ask my patient, "Do you understand what I'm going to be doing? Can you explain to me what we talked about?" And if there's any gap there, then I either re-explain it or I look for another person to help me.* (P9, Orthopedic Surgeon)

Thus, clinicians relied on both social cues including patient's body language and facial expressions as well as teach-back methods as a means to communicate and verify if patients understood their interactions.

5 DISCUSSION AND FUTURE WORK

Clinicians in our sample faced challenges providing quality care to patients with limited English proficiency, largely due to time constraints and the limited availability of skilled medical interpreters. Our participants have devised innovative strategies for coping with these challenges including relying on back-translations, non-verbal communication, and testing patient understanding. Most of our participants were excited about the potential for MT-mediated communication with their patients if it was designed appropriately and rigorously evaluated for use in clinical settings. While general purpose MT systems fall short of meeting these needs, our findings demonstrate potential for MT systems to be designed specifically for safe and reliable use in healthcare. In this section, we discuss the implications of our findings for the design of MT for medical settings, and identify paths forward for future work.

5.1 Accurate translation of medical language

When they needed automatic translation, clinicians typically turned to general purpose, commercial machine translation tools, like Google Translate. However, they recognized that these tools were not built for use in medical settings, raising concerns that they would not accurately translate domain-specific medical language. Accurate translation is especially important in high-stakes settings like healthcare, where mistakes can be life-threatening [27]. While neural MT systems always provide the most statistically likely translation of a phrase, in high-stakes contexts, not providing a translation is often better than providing an incorrect one. MT systems for use in healthcare must be specifically designed to prioritize accurate translation of medical language, possibly through domain adaptation methods. Another important factor that clinicians identified for future model development is ensuring support for different dialects of languages. Existing MT tools do not only lack support for different dialects, but do not even indicate which dialect of a language they *do* support, or whether the MT produces language in a consistent dialect at all.

A complementary approach is to combine neural machine translation with professionally translated phrases. Prior studies of patient and clinician attitudes towards MT in healthcare suggests a preference for phrase-based applications, which limit the range of things a user can communicate, but guarantees accurate translation of those phrases [40, 52]. In many areas of medicine, there are common and relatively standardized topics of conversation (e.g., medical history, current symptoms), which lend themselves to fixed phrase-based applications more easily. Together, these findings suggest the potential for tools that combine phrase-based translation with neural MT to provide some guarantee of accuracy while also enabling more flexible communication when necessary. Clinicians and patients may be more confident using MT-enabled tools if they can see when a translation is mostly or entirely professionally translated.

5.2 Rigorous evaluation and endorsement

Clinicians repeatedly commented on the importance of a machine translation tool having been vetted and endorsed by a governing body, such as a hospital board or medical society. Such validation typically signals that a tool was tested via a randomized control trial, and that the findings were published in a peer-reviewed medical journal. While randomized controlled trials are the standard practice for drugs and medical devices, the practice has lagged behind in ML for health [13, 20, 32]. Collaborations across healthcare and machine learning will be critical to developing rigorous, consistent standards for evaluating new ML systems for clinical use.

Evaluation in MT is notoriously difficult, due to the nuanced and subjective nature of human language translation [10, 18]. Prior evaluations of MT for health have adopted a range of evaluation criteria, from directly assessing translation accuracy or adequacy [15, 41, 49], to judging translation errors based on their potential to cause clinical harm [27], to testing whether a clinician is able to reach a correct diagnosis in a role-play scenario using an MT tool [46]. Several studies also assessed user satisfaction or the overall usability of a tool [40, 46, 52]. Future work is necessary to understand how best to measure the risks and benefits of MT in a clinical setting. Our findings suggest that what is most important to clinicians is not translation quality for its own sake, but how a translation tool is able to support more effective patient-clinician communication, and thus improve patient outcomes. While MT tools should initially be evaluated for acceptable levels of translation accuracy, they must also be evaluated using RCTs to ensure that they have a beneficial impact on patients' health.

5.3 Beyond one-to-one translation

The core functionality of existing machine translation systems is to take an input text and produce a single best translation of that input into another language. Our approach in this research strives to reorient the goals of MT systems for clinical use from seemingly objectively optimizing translation quality between two texts, to designing for the overall quality of cross-lingual patient-clinician communication. This mirrors theories in communication that differentiate between the monological model of communication with the dialogical model [56]. In the monological model of communication, communication is seen as a transfer of intent from the speaker. In this model, meaning is a result of the speakers intentions only and the speaker is thought of as in a social vacuum. The dialogical model of communication on the other hand, views meaning as a joint product of the speaker and listener. In this model, sense is made in and by a joint activity and there is reciprocity in both communication and miscommunication. Therefore, building on the dialogical model of communication can offer insights into what reliable MT-mediated communication might look like. To that end, our interviews with clinicians offered insight into their broader communication challenges and tactics that suggest paths forward for technological support that goes beyond one-to-one mapping translation tools.

5.3.1 Consider varying language proficiency. Existing MT systems do not account for users' language proficiency in the source or target language. Often, people use machine translation to communicate because they do not share a common language. Our findings, however, highlight that cross-lingual communication where both people have *no* knowledge of the other's language is not the only use case for MT. For example, clinicians with some proficiency in a language, but who are not medically certified to use that language in clinical practice, may use MT to search for unfamiliar terms or double-check their own translations. Similarly, patients with limited English proficiency may understand plain English but struggle with medical jargon. Considering a wider range of user language proficiency opens up new design opportunities for MT systems and language support more broadly. For instance, integrating MT support into patient portals could allow patients to only translate parts of a text they don't understand, or clinicians to easily view suggestions or translations as they write. In the other direction, patients and clinicians with some bilingual knowledge could offer useful feedback to an MT system when a translation is incorrect or confusing.

5.3.2 Support for checking understanding. Clinicians emphasized that cross-lingual communication was not only about conveying information to a patient, but also checking their understanding and offering them an opportunity to clarify as needed. MT systems increasingly support two-way conversation, for instance, with speech-to-speech conversation modes or easily switching source and target languages, but these features are not explicitly designed with teach-back methods or clarifying questions in mind. Research has found that it is particularly difficult for users to ask clarifying questions during MT-mediated conversation, because translations are not guaranteed to be symmetric, making it difficult to clearly refer back to part of an earlier translation [59]. Future work could design specific features that support clinicians' existing practices or introduce new strategies for checking patient understanding.

5.3.3 Literacy challenges. Finally, for MT systems to promote more equitable outcomes in healthcare, we must consider how to support patients with various levels of literacy in their primary language, as well as patients who speak languages with no written form. Literacy challenges include both the ability to read written instructions and information as well as medical literacy. While clinicians touched on this challenge in the study, particularly with regard to communicating with patients with limited medical literacy, future work is needed to more deeply understand these challenges. One direction for design would be to consider multimodal support, including visual and audio communication [40]. One physician in this study who had served in a number of international service missions had sent patients audio recordings

via WhatsApp to instruct them on what to do after being discharged; speech-to-speech translation with an option to export translated audio could hold potential for extending this practice to cross-lingual communication.

5.4 Setting transparent safety standards and recommended practices

Although clinicians were aware of some of the limitations of MT systems, such as their unreliable performance on medical terminology, this awareness was in spite of limited transparent information from available MT systems. For example, MT performance varies widely across language pairs, based on the quantity of available training data and the investment that has been made in developing models for each language, but this information is not conveyed to end users of commercial systems. Designers developing MT tools for healthcare must consider ways to clearly convey the system's limitations to its users. Cai et al. (2019) identified system strengths and limitations, design objectives, and subjective perspective as key information needs for clinicians adopting new ML tools [9]. In the MT context, further work is needed to develop onboarding materials and concrete guidelines to help clinicians use MT safely. Given clinicians' limited time, a system would ideally offer timely reminders when a user appears to be violating those guidelines, with interactive suggestions for how to resolve the issue and improve translation quality.

5.5 Limitations and ethical considerations

While we argue that there is substantial potential to improve on MT tools to support cross-lingual communication in clinical settings, this is not without risks and limitations. We conclude by enumerating some of these issues and offering possible paths forward.

Additional language support is most urgently needed in low-resource clinics and for low-resource languages, where a human interpreter is less likely to be available. However, appropriating new technologies into work practices will always involve substantial resources and human effort [36]. There is a risk that introducing new technology, even with the best of intentions, will further drain resources and attention from patients with limited English proficiency, exacerbating, rather than ameliorating, disparities in quality of care. Further, any errors that an MT system makes will disproportionately impact these patients in lower-resourced settings and/or who speak lower-resourced languages. This risk underscores the importance of rigorous, iterative, and multi-faceted evaluation practices. Any MT system for healthcare must be evaluated for reliability and accuracy as well as benefit to patient health outcomes. While it is important to evaluate these systems in the low-resource settings where they have the most potential to both help and harm, these evaluations must be conducted with caution to avoid undue burden or risk for the patients and clinicians involved.

Language translation is also not the only challenge that clinicians face in cross-lingual communication with patients. A prominent theme in our interviews was that clinicians face difficulties providing culturally appropriate care and building trust with patients across cultural differences. These difficulties are exacerbated by language barriers but not necessarily resolved by interpreter support alone. There may be some ways to incorporate cultural awareness into MT systems, for instance, ensuring that translations are evaluated for appropriateness and that the evaluators understand the cultural context, but this is likely to be limited in scope. Developers of MT support for healthcare have a responsibility to clearly communicate the limitations of MT systems, and ensure that they are not viewed as a solution for bridging cultural differences or a replacement for investment in resources like cultural navigators or continuing education about providing culturally appropriate care.

Finally, in this study we engaged clinicians to understand how they manage cross-lingual communication, whether and how they include MT in their practice, and how MT tools could better support them. While the insights from our

sample offer valuable directions for future work developing MT for healthcare, patient perspectives are also extremely important to inform this work. Unfortunately, much of the existing literature on MT in healthcare settings has privileged clinician perspectives over patient perspectives. Future work must meaningfully engage with patients to understand how, if at all, they want MT to play a role in their care.

6 CONCLUSION

In this paper, we conducted a qualitative interview study of clinicians–physicians, surgeons, nurses, and midwives–across seven medical specialties to understand how to design appropriate machine translation systems for medical settings. Cross-lingual communication between clinicians and patients can greatly affect quality of care. Machine translation systems are a low-cost solution that can address challenges clinicians have when interacting with patients in the presence of language barriers. Such systems, when accounting for reliability and transparency, can improve the quality of patient care in under-resourced settings that typically are limited to access to medical interpreters. Even in high-resourced medical settings, machine translation systems can complement existing translation services by working with medical interpreters who many not have knowledge of complex medical jargon. Ultimately, we recommend that machine translation systems: 1) account for dialect differences, 2) combine neural machine translation with professionally translated phrases, 3) move beyond one-to-one translations by varying language proficiency levels, 4) encourage a feedback system to check for understanding, 5) consider literacy levels of patients, and 6) set transparent safety standards and interactive suggestions for clinicians. We discuss how and when machine translation systems should be deployed in high-stake medical settings with the hope to better serve diverse patient populations.

ACKNOWLEDGMENTS

We would like to thank the participants for their time and for sharing their experiences, and the anonymous reviewers for their thoughtful feedback.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[2] Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783* (2019).

[3] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[5] Yotam Berger. 2017. Israel arrests Palestinian because Facebook translated'good morning'to'attack them'. *Ha'aretz* 22 (2017).

[6] Andrew Bernard, Misty Whitaker, Myrna Ray, Anna Rockich, Marietta Barton-Baxter, Stephen L Barnes, Bernard Boulanger, Betty Tsuei, and Paul Kearney. 2006. Impact of language barrier on acute care medical professionals is dependent upon role. *Journal of Professional Nursing* 22, 6 (2006), 355–358.

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[8] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[9] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[10] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic, 136–158. https://aclanthology.org/W07-0718

[11] Olveen Carrasquillo, E John Orav, Troyen A Brennan, and Helen R Burstin. 1999. Impact of language barriers on patient satisfaction in an emergency department. *Journal of general internal medicine* 14, 2 (1999), 82–87.

[12] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[13] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie J. Calvert, Ara Darzi, Christopher Holmes, Christopher Yau, David Moher, Hutan Ashrafian, Jonathan J. Deeks, Lavinia Ferrante di Ruffano, Livia Faes, Pearse A. Keane, Sebastian J. Vollmer, Aaron Y. Lee, Adrian Jonas, Andre Esteva, Andrew L. Beam, Maria Beatrice Panico, Cecilia S. Lee, Charlotte Haug, Christophe J. Kelly, Christopher Yau, Cynthia Mulrow, Cyrus Espinoza, John Fletcher, David Moher, Dina Paltoo, Elaine Manna, Gary Price, Gary S. Collins, Hugh Harvey, James Matcham, Joao Monteiro, M. Khair ElZarrad, Lavinia Ferrante di Ruffano, Luke Oakden-Rayner, Melissa McCradden, Pearse A. Keane, Richard Savage, Robert Golub, Rupa Sarkar, Samuel Rowley, The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, and SPIRIT-AI and CONSORT-AI Consensus Group. 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine* 26, 9 (Sept. 2020), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

[14] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *CoRR* abs/2006.11371 (2020). arXiv:2006.11371 https://arxiv.org/abs/2006.11371

[15] Prithwijit Das, Anna Kuznetsova, Meng'ou Zhu, and Ruth Milanaik. 2019. Dangers of Machine Translation: The Need for Professionally Translated Anticipatory Guidance Resources for Limited English Proficiency Caregivers. *Clinical Pediatrics (Phila)* 58, 2 (Feb 2019), 247–249. https://doi.org/10.1177/0009922818809494

[16] Kristin N Dew, Anne M Turner, Yong K Choi, Alyssa Bosold, and Katrin Kirchhoff. 2018. Development of machine translation technology for assisting health communication: A systematic review. *Journal of biomedical informatics* 85 (2018), 56–67.

[17] Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. Improving statistical machine translation in the medical domain using the Unified Medical Language System. In *Coling 2004: Proceedings of the 20th international conference on Computational Linguistics*. 792–798.

[18] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics* 9 (12 2021), 1460–1474. https://doi.org/10.1162/tacl_a_00437 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00437/1979261/tacl_a_00437.pdf

[19] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4, 31 (2021). https://doi.org/10.1038/s41746-021-00385-9

[20] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Viewpoint* 3 (2021). Issue 11.

[21] Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill, 41–58.

[22] Louis C Hampers, Susie Cha, David J Gutglass, Helen J Binns, and Steven E Krug. 1999. Language barriers and resource utilization in a pediatric emergency department. *Pediatrics* 103, 6 (1999), 1253–1256.

[23] Kotaro Hara and Shamsi T Iqbal. 2015. Effect of machine translation in interlingual conversation: Lessons from a formative study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3473–3482.

[24] Rachel E Hommes, Amy I Borash, Kari Hartwig, and Donna DeGracia. 2018. American Sign Language interpreters perceptions of barriers to healthcare communication in deaf and hard of hearing patients. *Journal of community health* 43, 5 (2018), 956–961.

[25] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813* (2020).

[26] Feroze Kaliyadan and Sreekanth Gopinathan Pillai. [n. d.]. The use of Google language tools as an interpretation aid in cross-cultural doctor-patient interaction: A pilot study. *Informatics in Primary Care* 18, 2 ([n. d.]), 141–143.

[27] Elaine C Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. 2019. Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA internal medicine* 179, 4 (2019), 580–582.

[28] Maria Kletečka-Pulker, Sabine Parrag, Klara Doppler, Sabine Völkl-Kernstock, Michael Wagner, and Thomas Wenzel. 2021. Enhancing patient safety through the quality assured use of a low-tech video interpreting system to overcome language barriers in healthcare settings. *Wiener klinische Wochenschrift* 133, 11 (2021), 610–619.

[29] Janis Kritzinger, Marguerite Schneider, Leslie Swartz, and Stine Hellum Braathen. 2014. "I just answer 'yes' to everything they say": Access to health care for deaf people in Worcester, South Africa and the politics of exclusion. *Patient education and counseling* 94, 3 (2014), 379–383.

[30] Dagmara Kuliś, Cheryl Whittaker, Eva Greimel, Andrew Bottomley, Michael Koller, and EORTC Quality of Life Group. 2017. Reviewing back translation reports of questionnaires: the EORTC conceptual framework and experience. *Expert review of pharmacoeconomics & outcomes research* 17, 6 (2017), 523–530.

[31] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337* (2019).

[32] David B. Larson, Hugh Harvey, Daniel L. Rubin, Neville Irani, Justin R. Tse, and Curtis P. Langlotz. 2021. Regulatory Frameworks for Development and Evaluation of Artificial Intelligence–Based Diagnostic Imaging Algorithms: Summary and Recommendations. *Journal of the American College of*

Radiology 18, 3 (2021), 413–424. https://doi.org/10.1016/j.jacr.2020.09.060

[33] Fernando Ochoa Leite, Catarina Cochat, Henrique Salgado, Mariana Pinto da Costa, Marta Queirós, Olga Campos, and Paulo Carvalho. 2016. Using Google Translate in the hospital: A case report. *Technology and Health Care* 24, 6 (2016), 965–968.

[34] Daniel J Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet Needs and Opportunities for Mobile Translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[35] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* 1, 6 (2019). https://doi.org/10.1016/S2589-7500(19)30123-2

[36] Alexandra Mateescu and Madeleine Clare Elish. 2019. AI in context: the labor of integrating new technologies. *Data & Society* 30 (2019).

[37] Tom Moberly. 2018. Doctors are cautioned against using Google Translate in consultations. *BMJ* 363 (2018). https://doi.org/10.1136/bmj.k4546 arXiv:https://www.bmj.com/content/363/bmj.k4546.full.pdf

[38] Tom Moberly. 2018. Doctors choose Google Translate to communicate with patients because of easy access. *BMJ* 362 (2018). https://doi.org/10.1136/bmj.k3974 arXiv:https://www.bmj.com/content/362/bmj.k3974.full.pdf

[39] Uldis Ozolins, Sandra Hale, Xiang Cheng, Amelia Hyatt, and Penelope Schofield. 2020. Translation and back-translation methodology in health research–a critique. *Expert review of pharmacoeconomics & outcomes research* 20, 1 (2020), 69–77.

[40] Anita Panayiotou, Kerry Hwang, Sue Williams, Terence W H Chong, Dina LoGiudice, Betty Haralambous, Xiaoping Lin, Emiliano Zucchi, Monita Mascitti-Meuter, Anita M Y Goh, Emily You, and Frances Batchelor. 2020. The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. *Journal of Clinical Nursing* 29, 17-18 (Sep 2020), 3516–3526. https://doi.org/10.1111/jocn.15390

[41] Sumant Patil and Patrick Davies. 2014. Use of Google Translate in medical communication: evaluation of accuracy. *Bmj* 349 (2014).

[42] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* 32, 10 (2020), 6363–6381.

[43] Gurdeeshpal Randhawa, Mariella Ferreyra, Rukhsana Ahmed, Omar Ezzat, and Kevin Pottie. 2013. Using machine translation in clinical practice. *Canadian Family Physician* 59, 4 (2013), 382–383.

[44] Alejandro Renato, José Castano, Maria del Pilar Avila Williams, Hernán Berinsky, Maria Laura Gambarte, Hee Joon Park, David Pérez-Rey, Carlos Otero, and Daniel R Luna. 2018. A Machine Translation Approach for Medical Terms.. In *HEALTHINF*. 369–378.

[45] Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*. 97–98.

[46] Hervé Spechbach, Ismahene Sonia Halimi Mallem, Johanna Gerlach, Nikolaos Tsourakis, and Pierrette Bouillon. 2017. Comparison of the quality of two speech translators in emergency settings : A case study with standardized Arabic speaking patients with abdominal pain. In *Proceedings of European Congress of Emergency Medicine (EUSEM 2017)*. Athens, Greece. https://archive-ouverte.unige.ch/unige:100812

[47] Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591* (2019).

[48] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[49] Breena R. Taira, Vanessa Kreger, Aristides Orue, and Lisa C. Diamond. 2021. A Pragmatic Assessment of Google Translate for Emergency Department Instructions. *Journal of General Internal Medicine* 36 (2021), 3361–3365. https://doi.org/10.1007/s11606-021-06666-z

[50] Tsuyoshi Tatemoto, Masahiko Mukaino, Nobuhiro Kumazawa, Shigeo Tanabe, Koji Mizutani, Masaki Katoh, Eiichi Saitoh, and Yohei Otaka. 2021. Overcoming language barriers to provide telerehabilitation for COVID-19 patients: a two-case report. *Disability and Rehabilitation: Assistive Technology* (2021), 1–8.

[51] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Proceedings of the 4th Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 106)*, Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (Eds.). PMLR, 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html

[52] Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study. *JMIR Public Health Surveill* 5, 1 (Jan 2019). https://doi.org/10.2196/11171

[53] Anne M Turner, Hannah Mandel, and Daniel Capurro. 2013. Local health department translation processes: potential of machine translation technologies to help meet needs. In *AMIA annual symposium proceedings*, Vol. 2013. American Medical Informatics Association, 1378.

[54] Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2020. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society* 0, 0 (2020), 1–18. https://doi.org/10.1080/1369118X.2020.1776370 arXiv:https://doi.org/10.1080/1369118X.2020.1776370

[55] Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society* 24, 11 (2021), 1515–1532.

[56] Cecilia Wadensjo. 2014. *Interpreting as interaction*. Routledge.

[57] Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 679–688.

[58] Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. 515–524.

[59] Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06*. ACM Press, Banff, Alberta, Canada, 515. https://doi.org/10.1145/1180875.1180955

[60] Emre Şentürk, Mukadder Orhan-Sungur, and Tülay Özkan Seyhan. 2021. Google Translate: Can It Be a Solution for Language Barrier in Neuraxial Anaesthesia? *Turkish journal of anaesthesiology and reanimation* 49, 2 (2021), 181–182. https://doi.org/10.5152/TJAR.2021.101