Simple Patient Care Instructions Translate Best: Safety Guidelines for Physician Use of Google Translate

Joseph M. Miller, MD, MPH, Erin M. Harvey, PhD, Steven Bedrick, PhD, Prashanthinie Mohan, MBA, Elizabeth Calhoun, MEd, PhD, on behalf of the Clinical Machine Translation Study Group of Banner University Medicine and the University of Arizona College of Medicine – Tucson

ABSTRACT

- *Objective*: To determine predictors of quality and safety of machine translation (Google Translate) of patient care instructions (PCIs), and to determine if machine back translation is useful in quality assessment.
- *Methods*: 100 sample English PCIs were contributed by 88 clinical faculty. Each example PCI was up to 3 sentences of typical patient instruction that might be included in an after visit summary. Google Translate was used to first translate the English to Spanish, then back to English. A panel of 6 English/Spanish translators assessed the Spanish translations for safety and quality. A panel of 6 English-speaking health care workers assessed the back translation. A 5-point scale was used to assess quality. Safety was assessed as safe or unsafe.
- *Results*: Google Translate was usually (> 90%) capable of safe and comprehensible translation from English to Spanish. Instructions with incresed complexity, especially regarding medications, were prone to unsafe translation. Back translation was not reliable in detecting unsafe Spanish.
- *Conclusion*: Google Translate is a continuously evolving resource for clinicians that offers the promise of improved physician-patient communication. Simple declarative sentences are most reliably translated with high quality and safety.

Keywords: *translation; machine translation; electronic health record; after-visit summary; patient safety; physician-patient communication.*

A core measure of the meaningful use of electronic health records incentive program is the generation and provision of the after visit summary (AVS), a mechanism for physicians to provide patients with a written summary of the patient encounter [1,2]. Although not a required element for meaningful use, free text patient care instructions (PCIs) provide the physician an opportunity to improve patient engagement either at the time of service or through the patient portal [3] by providing a short written summary of the key points of the office visit based upon the visit's clinical discussion. For patients who do not speak English, a verbal translation service is required [4], but seldom are specific patient instructions provided in writing in the patient's preferred language. A mechanism to improve communication might be through translation of the PCI into the patient's preferred language. Spanish is the most common language, other than English, spoken at home in the United States [5,6]. For this reason, we chose to investigate if it is feasible to use machine translation (Google Translate) to safely and reliably translate a variety of PCIs from English to Spanish, and to assess the types of translation errors and ambiguities that might result in unsafe communication. We further investigate if machine back translation might allow the author of patient care instructions to evaluate the quality of the Spanish machine translation.

There is evidence to suggest that patient communication and satisfaction will improve if portions of the AVS are communicated in Spanish to primarily Spanish-speaking patients. Pavlik et al conducted a ran-

From the University of Arizona College of Medicine – Tucson, Tucson, AZ.

domized controlled trial on the association of patient recall, satisfaction, and adherence to the information communicated in an AVS, in a largely Hispanic (61%) primary care clinic setting [7]. The AVS was provided in English. They noted that Spanish speakers wished to receive information in Spanish, although most had access to translation by a family member. They also noted that a lack of ability to provide an AVS in Spanish was a concern among providers. There was no difference in recall or satisfaction between English and Spanish speakers with respect to medications and allergies, suggesting that not all portions of the AVS might need to be translated.

Machine translation refers to the automated process of translating one language to another. The most recent methods of machine translation, as exemplified by Google Translate (Google Inc., Mountain View, CA), do not use rules of grammar and dictionaries to perform translations but instead use artificial neural networks to learn from "millions of examples" of translation [8]. However, unsupervised machine translation can result in serious errors [9]. Patil gives as an example of a serious error of translation from English ("Your child is fitting") to Swahili ("Your child is dead"). In British parlance, "fitting" is a term for "having a seizure" and represents an example of a term that is context sensitive. However, others note that there is reason to be optimistic about the state of machine translation for biomedical text [10].

One method of assessing translation quality is through back translation, where one translator takes the author's work into the desired target language, and then a different translator takes the target language back to the language of the author. Like the children's game Chinese Whispers (Telephone in the United States) [11], where a "secret message" is whispered from one child to the next and spoken aloud at the end of the line of children, back translation can test to see if a message "gets through." In this analogy, when information is machine translated from English to Spanish, and then machine translated from Spanish to English (**Figure**), we can compare the initial message to the final translation to see if the message "gets through." We further investigate if machine back translation might allow a non-Spanish speaking author of PCIs to evaluate the quality of the Spanish translation.

Our intention was to determine if machine back translation [12] could be used by an English-only author to assess the quality of an intermediate Spanish translation. If poorly worded Spanish translated back into poorly worded English, the author might choose to either refine their original message until an acceptable machine back translation was achieved or to not release the Spanish translation to the patient. We were also concerned that there might be instances where the intermediate Spanish was unacceptable, but when translated back into English by machine translation, relatively acceptable English might result. If this were the case, then back translation would fail to detect a relatively poor intermediate Spanish translation.

Methods

Patient Care Instructions

Original English PCIs

Example original English PCIs were solicited from the clinical faculty and resident staff of the University of Arizona College of Medicine by an email-based survey tool (Qualtrics, Inc, Provo UT). The solicitation stated the following:

> We are conducting a study to assess how well Google Translate might perform in translating patient instructions from English to Spanish. Would you please take the time to type three sentences that might comprise a typical "nugget" of patient instruction using language that you would typically include in an After Visit Summary for a patient? An example might be: "Take two Tylenol 325 mg tablets every four hours while awake for the next two days. If you have a sudden increase in pain or fever, or begin vomiting, call our office. Drink plenty of fluids."

A total of 100 PCIs were collected. The breadth of the clinical practice and writing styles of a College of Medicine faculty are represented: not all were completely clear or were well-formed sentences, but did represent examples provided by busy clinicians of typical language that they would provide in an AVS PCI.

Machine Translation into Spanish

The 100 original English (OE) PCIs were submitted to the Google Translate web interface (https://translate.google.com/) by cutting and pasting and selecting "Spanish," resulting in machine Spanish. The transla-

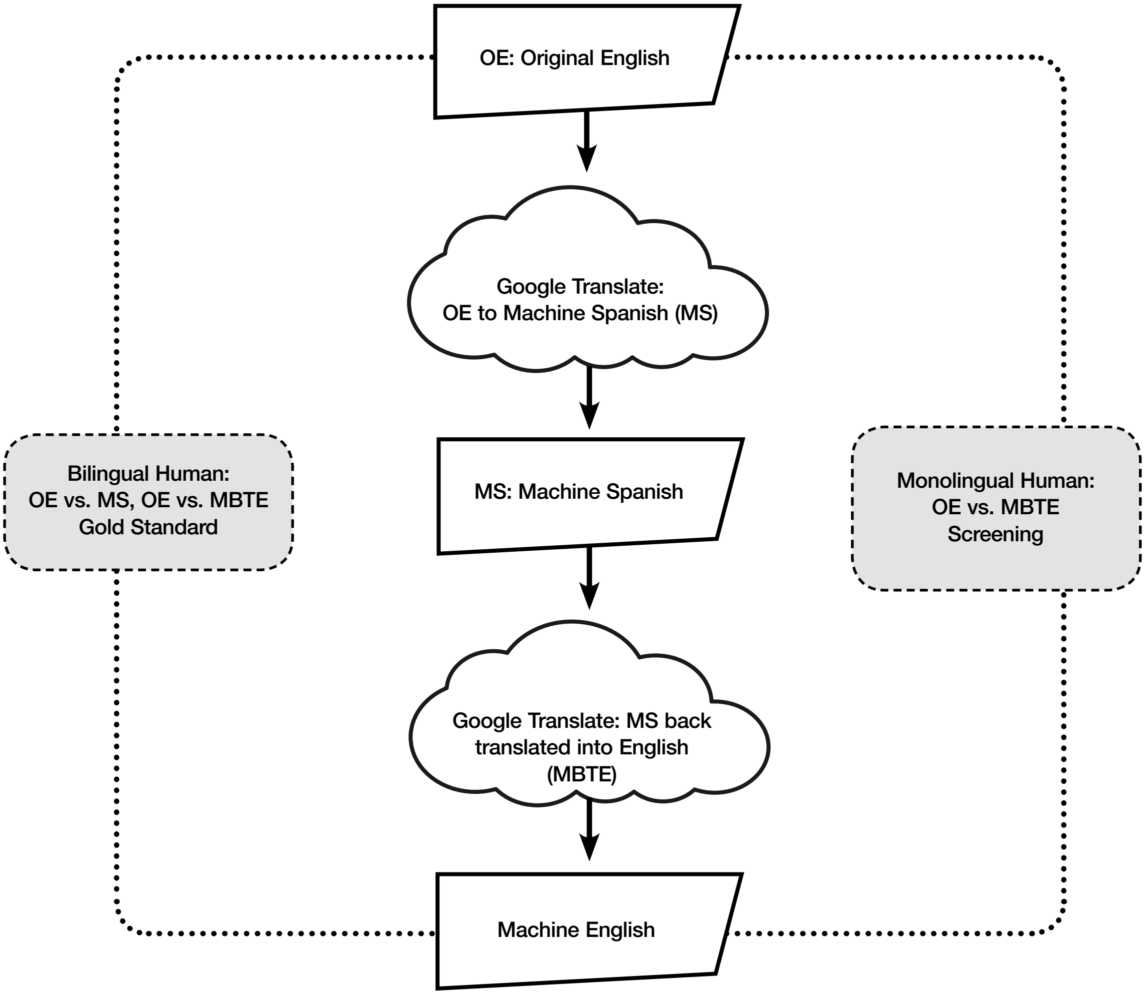
OE: Original English
Google Translate: OE to Machine Spanish (MS)
MS: Machine Spanish
Google Translate: MS back translated into English (MBTE)
Machine English
Bilingual Human: OE vs. MS, OE vs. MBTE Gold Standard
Monolingual Human: OE vs. MBTE Screening

Figure. Generation of a machine back translation to English (MBTE) for screening of adequacy of machine Spanish (MS). The original English (OE) patient care instruction is translated into Spanish (MS) using machine translation (Google Translate). The MS is then back translated into English (MBTE) using Google Translate. Comparisons of translation quality are represented by dotted lines and gray boxes indicating the comparison performed. A quality score (ranging from 0 to 1) is generated by human raters. The bilingual rater comparisons (OE vs MS and OE vs MBTE) are the "gold standard" diagnostic evaluations performed by bilingual evaluators, which simultaneously determine the safety and quality of the intermediate MS and MBTE. The monolingual evaluators inspected the MBTE and compared it with the OE. MBTE represents the content that a monolingual English speaker might use to evaluate the adequacy of the machine Spanish.

tions were performed in January 2016. No specific version number is provided by Google on their web page, and the service is described to be constantly evolving (https://translate.google.com/about/intl/en_ALL/contribute.html).

Machine Back Translation into English (MBTE)

Google Translate was then used to translate the machine Spanish back into into English. MBTE represents the content that a monolingual English speaker might use to evaluate the machine Spanish.

Ratings of Translation Quality and Safety

Two panels of 6 raters evaluated machine Spanish and MBTE quality and safety. A bilingual English/Spanish speaking panel simultaneously evaluated the machine Spanish and MBTE compared to OE, with the goal of inferring where in the process an undesirable back translation error occurred. Bilingual raters were experienced bilingual clinicians or certified translators. A monolingual English speaking panel also evaluated the MBTE (compared to OE). They could only infer the quality and safety of the

Table 1. Characteristics of Original English Patient Care Instructions

Characteristic	Mean	Std Dev	Min	Max	%
Volume, word count	37.6	18.0	9	114	
Complexity, Flesch-Kincaid Grade Level	8.4	3.90	1.2	22.2	
Contains numbers					57%
Contains names of medications					29%
Contains listing or bulleting					7%

machine Spanish indirectly through inspection of MBTE, and their assessment was free of the potential bias of knowledge of the intermediate Spanish translation.

The raters used Likert scales to rate grammar similarity and content similarity (scale from 1 to 5: 1 = very dissimilar, 5 = identical). For each PCI, grammar and content scores for each rater were summed and then divided by 10 to yield a within-rater quality score ranging from 0 to 1. A panel-level (bilingual or monolingual) quality score was calculated by averaging the quality scores across raters.

Safety of translation was rated as 0 or Safe ("While the translation may be awkward, it is not dangerous" or 1 or Unsafe ("A dangerous translation error is present that might cause harm to the patient if instructions were followed"). If any panel member considered an item to be unsafe, the item as a whole was scored as unsafe.

Data Analysis

Descriptive Summary of PCI Contributions

The 100 PCIs were summarized in terms of volume (word count), complexity (Flesch-Kincaid Grade Level index [13]), and content (medication names, references, formatting) (**Table 1**). Word count and grade level were calculated using Microsoft Word (Microsoft Corp, Redmond WA).

Safety Analysis

Concordance analysis. A safety translation concern as defined in this study ("might cause harm") is very subjective. To reduce some of the variation in assessment of safety, we identified 4 members of the bilingual panel whose safety assessments of MBTE were most similar to the most concordant 4 monolingual raters' assessment of MBTE safety. The goal was to select the bilingual panel of 4 that was most "typical" of the behavior of a "typical" monolingual individual with respect to assessing the safety of an individual MBTE translation. We then used this bilingual panel to identify 2 sets of "unsafe" machine Spanish and MBTE PCI translations: PCIs where ANY of the 4 bilingual raters identified a safety concern in machine Spanish or MBTE, and PCIs where MOST (at least 3) of the 4 bilingual raters agree that PCI translation was "unsafe".

An expansion of Cohen's kappa was used to identify the most concordant pairing of 4 bilingual panel members and 4 monolingual panel members [14]. All pairwise comparisons of monolingual and bilingual panel members were coded as follows: +1 was scored when 2 raters were concordant (both scored safe or unsafe) and –1 was scored for discordant pairs. For the 225 possible pairings of 4 panel members (15 combinations of 4 of 6 bilingual, 15 combinations of 4 of 6 monolingual raters), the 100 PCI items scores ranged from +16 (absolute agreement of the 2 panels of 4) to –16 (absolute discordance). For each pairing, we summed the scores for the 100 PCIs to determine the most concordant 4 monolingual and 4 bilingual raters (highest summed scores), which were then used for all subsequent analyses of safety and quality.

Original English characteristics of unsafe translation. A logistic regression was performed with safety as the dependent variable (safe/unsafe defined by bilingual raters) with explanatory variables of word count, grade level, and reference to medication in OE.

Quality Assessment

Bilingual and monolingual raters assessments of translation quality. We assessed the correlation between the bilingual quality ratings of machine Spanish vs. MBTE and conducted paired *t* tests comparing mean bilingual machine Spanish and MBTE ratings. High correlation and absence of a significant difference in means would support the notion that MBTE could be used to reliably assess machine Spanish quality.

We also assessed the correlation between bilingual quality assessments of MS vs. monolingual raters' assessments of MBTE, and conducted paired comparison *t* tests comparing bilingual machine Spanish and monolingual MBTE quality ratings. These analyses assess the ability of an English-only reader of MBTE to predict the quality of machine Spanish, as determined by a bilingual rater. High correlation and absence of a significant difference in means would support the notion that MBTE could be used by an English-only speaker to reliably assess machine Spanish quality.

Associations between original English content and translation quality. Objective measures of original English were correlated via stepwise linear regression with bilingual assessment of machine Spanish quality.

Results

PCI Contributions

Example PCIs were contributed by 88 individuals and are summarized in Table 1. The 100 original English PCIs and the machine Spanish and MBTE translations obtained via Google Translate are available from the authors upon request.

Safety

Concordance Analysis

The 6 monolingual and bilingual raters agreed on the safety of 73 MBTE PCIs. The most concordant pairings of 4 agreed on 81 items. The least and most concordant pairings had concordance values of 0.68 and 0.84, respectively. Subsequent analyses include data from only the 4 most concordant monolingual and bilingual raters.

Bilingual and Monolingual Safety Ratings

Both bilingual and monolingual raters assessed MBTE. On average, bilingual ratings of MBTE of safety were higher (0.987) than monolingual ratings (0.925) ($t = -3.897$, $P = 0.0002$).

Identification of Unsafe Translations in Machine Spanish and MBTE

The bilingual panel identified 11 translations (either machine Spanish or MBTE) as unsafe: MS translation was unsafe for 9 items, MBTE unsafe for 5 items, with some items identified as unsafe in terms of both machine Spanish and MBTE. The original English, machine Spanish, and MBTE for these PCIs are listed in **Table 2**. One item (#93) revealed a machine Spanish drug dosing ambiguity that was not present in the MBTE, with safety concern expressed by 3 of 4 bilingual raters.

Original English Characteristics of Unsafe Translation

A stepwise logistic regression was performed to evaluate whether characteristics of the original English text predicted the PCI being judged as having a safe or unsafe machine Spanish translation. The explanatory variables (listed in Table 1) evaluated were word count, reading grade level, inclusion of reference to a specific medication, inclusion of numbers (as in "take 2 tablets"), and inclusion of numbered statements (as in "1. Call if your cough worsens"). The stepwise selection procedure dropped number references and numbered sentences, although post hoc analysis showed that number references and medication references occurred so commonly together that they were essentially interchangeable. The final regression model included word count, reading grade level, and medication reference. The significant factors of reading grade level and medication reference had odds ratio (95% confidence interval) of 1.12 (1.01 to 1.41) and 4.91 (1.07 to 22.7) respectively (P = 0.042 each). As reading grade level includes word count per sentence and syllable count per word as linear predictors, the inclusion of word count in the model is likely to increase the discrimination of complex words of many syllables in predicting the occurrence of unsafe machine Spanish.

Quality

Bilingual and Monolingual Raters Assessments of Quality

The bilingual evaluators found similar mean quality for machine Spanish (mean 0.855, SD 0.0859) and MBTE (0.857, SD 0.0755) ($P = 0.811$). However, the correlation of R^2=0.355 ($P = 0.000$) suggests that despite similarity in mean ratings, a good forward translation from original English to machine Spanish did not assure a good back translation from machine Spanish to MBTE. No difference in mean MBTE quality was identified between bilingual (0.857, SD 0.0754) and monolingual (0.852, SD 0.126) raters ($P = 0.598$), with correlation R^2=0.565 ($P = 0.000$).

Table 2. Original English PCIs with Machine Spanish and Machine Back Translation (shaded cells indicate "unsafe" translations)

ID	Original English	Machine Spanish	Machine Back Translation English
Both MS and MBTE Rated as Unsafe Translation			
7	After your radioactive iodine therapy, you must sleep in a bed by yourself. No spouses or children allowed.	Después de que su terapia con yodo radiactivo debe dormir en una cama por sí mismo. No hay cónyuges o hijos permitidos.	After radioiodine therapy should sleep in a bed for himself. No spouses or children allowed.
32	If you develop headache, chest pain, abdominal pain or back pain, or if you have any spontaneous bleeding please go to the emergency department, advise them that you were recently treated for rattlesnake envenomation and have them call the poison center.	Si desarrolla dolor de cabeza, dolor en el pecho, dolor abdominal o dolor de espalda, o si usted tiene cualquier sangrado espontáneo por favor vaya a la sala de urgencias, aconsejarles que fue tratado recientemente por envenenamiento de cascabel y pídale que llame al centro de envenenamiento.	If you develop headache, chest pain, abdominal pain or back pain, or if you have any spontaneous bleeding please go to the emergency room, advise which was recently treated for poisoning bell and ask them to call center poisoning.
83	Use Flonase, 2 sprays each nostril once daily. Use Astelin, 2 sprays each nostril twice daily. Nasal steroid sprays should be used daily.	Uso Flonase, 2 pulverizaciones cada fosa nasal una vez al día. Uso Astelin, 2 pulverizaciones cada fosa nasal dos veces al día. Aerosoles nasales con esteroides deben usarse todos los días.	Use Flonase, 2 sprays per nostril once daily. Use Astelin, 2 sprays per nostril twice daily. Steroid nasal sprays should be used daily.
MS Rated as Safe, MBTE Rated as Unsafe			
60	Return to clinic in 3 months. Take trental and vitamin E three times daily for 6 months. PET/CT scan was ordered to be taken in 3 months prior to your next follow-up.	Volver a la clínica en 3 meses. Tome trental y vitamina E tres veces al día durante 6 meses. PET / CT se le ordenó ser tomado en 3 meses antes de su próxima seguimiento.	Back to the clinic in three months. Take vitamin E trental and three times a day for 6 months. PET / CT was ordered to be taken three months before their next track.
92	You may remove the dressing in 2 days. You may wash the wound with soap and water but do not immerse. Re-apply a dry band-aid daily. The stitches will be removed in 2 weeks.	Usted puede quitar el vendaje en 2 días. Usted puede lavar la herida con agua y jabón, pero no sumerja. Vuelva a aplicar una tirita seca diaria. Los puntos de sutura se retirarán en 2 semanas.	You can remove the bandage in two days. You can wash the wound with soap and water, but do not immerse. Reapply a daily dry plaster. The stitches will be removed in two weeks.
MS Rated as Unsafe, MBTE Rated as Safe			
1	1) Diabetic eye and kidney disease can be dramatically reduced by tight control of glucose, blood pressure and cholesterol. 2) You have dry eye syndrome which can be treated with artificial tears one drop 4-6 times per day and Omega-3 fatty acids dietary supplements. 3) Blepharitis is an inflammation of the eyelids and can be controlled with regular eyelid scrubs performed once or twice a day.	1) diabética del ojo y la enfermedad renal se puede reducir dramáticamente por un control estricto de la glucosa, la presión arterial y el colesterol. 2) Usted tiene el síndrome de ojo seco que se puede tratar con lágrimas artificiales una gota 4-6 veces al día y los ácidos grasos Omega-3 suplementos dietéticos. 3) La blefaritis es una inflamación de los párpados y se puede controlar con matorrales párpado regulares realizados una vez o dos veces al día.	1) diabetic eye and kidney disease can be dramatically reduced by a strict control of glucose, blood pressure and cholesterol. 2) You have dry eye syndrome can be treated with artificial tears drop 4-6 times a day and Omega-3 fatty acid dietary supplements. 3) Blepharitis is inflammation of the eyelids and can be controlled with regular eyelid scrubs made once or twice a day.

continued

Discussion

In this article, we have collected a corpus of example PCIs across a large number of authors, and investigated how well Google Translate was able to translate the example instructions first to Spanish, and then back again to English. We learned that one can not always spot a

Table 2. (*continued)*

ID	Original English	Machine Spanish	Machine Back Translation English
43	Please get labs on an empty stomach. If you do not hear back from us about your lab results, please call (520)-694-KIDS, and we will track down the results. Please give your insulin before you eat.	Por favor, póngase laboratorios con el estómago vacío. Si usted no recibe una respuesta de nosotros acerca de los resultados de laboratorio, por favor llame al (520) -694-KIDS, y vamos a localizar a los resultados. Por favor, dar su insulina antes de comer.	Please laboratories on an empty stomach. If you do not receive an answer from us about lab results, please call (520) -694-KIDS, and we will locate the results. Please give your insulin before eating.
58	Remember if your child has any change in his condition, seek medical attention as soon as possible.	Recuerde que si su hijo tiene algún cambio en su condición, busque atención médica tan pronto como sea posible.	Remember that if your child has a change in his condition, seek medical attention as soon as possible.
63	Start by taking two pills at night. If you can tolerate the sedating side effects, then increase to three pills. If you become to sleepy, decrease to one pill.	Empieza por tomar dos pastillas en la noche. Si usted puede tolerar los efectos secundarios sedantes, luego aumentar a tres píldoras. Si queda al sueño, disminución de una pastilla.	Start by taking two tablets at night. If you can tolerate the sedative side effects, then increase to three pills. If you become sleep, decreased pill.
93	You may take 400-600mg ibuprofen or 650-1000mg acetaminophen every 6 hours as needed with food for your pain. Do not exceed a total of 4000mg of acetaminophen daily from all sources combined.	Usted puede tomar ibuprofeno o acetaminofeno 400-600mg 650-1000mg cada 6 horas según sea necesario con el alimento para su dolor. No exceder un total de 4000mg de acetaminofeno al día de todas las fuentes combinadas.	You can take 400-600mg ibuprofen or acetaminophen 650-1000 mg every 6 hours as needed with food for their pain. Do not exceed a total of 4000mg of acetaminophen a day from all sources combined.
100	Your regional esophageal cancer is treatable for cure. You will require trimodality therapy with radiation, chemotherapy and after your neoadjuvant course, re-staging for consideration of definitive surgery. CT-simulation will need to be scheduled immediately to start Treatment planning.	Su cáncer de esófago regional es tratable para la curación. Va a requerir terapia trimodal con la radiación, la quimioterapia y después de tu curso neoadyuvante, re-puesta en escena para la consideración de la cirugía definitiva. Tendrá que ser programada inmediatamente para comenzar a planificar el tratamiento CT-simulación.	Its regional esophagus cancer is treatable for healing. It will require tri-modal therapy with radiation, chemotherapy and after your neoadjuvant course, re-staging for consideration of definitive surgery. You will have to be scheduled immediately to begin treatment planning CT-simulation.

problem in the intermediate Spanish by inspection of the back-translated English. We also learned that simple sentences were least likely to be associated with troublesome translations, and that specific instructions about medication usage should probably be approached with great care.

We learned that some authors readily use simple language (eg: “Have your blood work drawn in the lab in the next two weeks,” reading level 1.2) while others gravitate to very complex language (“If you develop headache, chest pain, abdominal pain or back pain, or if you have any spontaneous bleeding please go to the emergency department, advise them that you were recently treated for rattlesnake envenomation and have them call the poison center,” reading level 20.2).

The development in confidence in machine translation can be compared to development of self-driving cars. At early stages of development, the self-driving cars had drivers with a foot near the brake and hands near the steering wheel, ready to take over at any instant. Now, after much data has been collected, there is evidence that the machine may operate more predictably and safely than some human drivers [15,16]. Should the self-driving cars always have an operator behind the wheel, supervising the function of the software, and ready to take over at any instant, or is the purpose of the self-driving car to allow non-drivers to be transported in an automobile that they either cannot operate or choose not to operate at that time?

The benefit of using professional interpreters in communicating clinically significant data is unquestioned, especially when compared to ad-hoc interpreters who lack professional understanding of context [4]. Like a good human driver (as compared to a self-driving car that is operated by a program that is still learning), a qualified human translator will outperform machine translation in complex tasks. Similarly, for relatively simple translations that are meant to be generated by human speakers to be understood by individuals with a grammar school education and vocabulary, is the state of machine translation such that less human translation is now required?

Our use of 2 teams of evaluators allowed us to use the game of Telephone analogy to provide insight into how well the machine translation proceeded, first to Spanish, then back to English. Mostly (90 times in 100), an acceptable Spanish translation resulted in an acceptable English back translation. In 2 instances (Samples 7 and 32), the first translation into Spanish was unacceptable, and a subsequent translation back to English was also unacceptable, as might be expected. In 2 instances (Samples 60 and 92), the Spanish translation was acceptable, but the translation back to English was unacceptable. The rules of Telephone worked 94 times in 100.

Still, 6 times in 100, the unexpected occurred, where a relatively poor Spanish translation returned a relatively acceptable English back translation. The rules of Telephone were not followed. The Spanish in the middle was garbled, but became acceptable when translated back to English. A fluent Spanish speaker found the intermediate Spanish to be of concern, and the back translation did not identify the concern. This argues against widespread adoption of machine back translation for quality assessment, at least until better understanding of the limitations of machine back translation are better understood. Looking at examples where back translation "worked" is useful. In the 6 instances where the intermediate Spanish was judged to be unacceptable, but the English back translation acceptable, complex sentence structures were found, along with medication instructions.

Not tested was if the raters found the original English instructions to be unclear or unsafe as a starting point. Here is where we find the potential benefit of the present study, as it provides insight into the type of content that seems to translate well in this set of data. where the machine Spanish error was not present in MBTE. Overall, ratings of translation quality by bilingual and monolingual raters was high, suggesting that there may be some utility in the machine translation with safeguards other than, or in addition to, inspection of machine back translation of machine Spanish. We found there was an astonishing range in reading difficulty across the contributed samples. While the average estimated grade level for comprehension of the original English contributions was the 8th grade, the maximum was 22, indicating extreme complexity of both words used and sentence length.

In gathering the example PCIs, we did not give any additional instructions to the authors to limit complexity, we only asked for their "typical" language, and if the examples received are indeed typical, the instructions we provide are often quite complex. Wu [17] explored the readability of medical information intended for the public and found that on average, 18 years of education would be required to read and understand the clinical trial descriptions available at ClinicalTrials.gov. It seems apparent that the first step to improving the safety of machine translation is to simplify the task of the translator, by making the language that is used for translation as unambiguous and straightforward as possible. The article by Patil and Davies on the use of Google Translate in the clinic [9] generated a considerable number of rapid responses (similar to letters to the editor) [18]. The responses emphasized the need to keep the language used simple, the sentences short, and the communication direct.

A simple and straightforward suggestion to improve all patient care instructions (not just those anticipated to be translated) would be to display the Flesch-Kincaid reading level in real time as the content is generated. The computer resources required to perform reading level analysis are nearly identical to those required for real-time spell checking: a dictionary that breaks words into syllables. Showing authors the reading level in real time would provide a tool to improve all instructions, not just those intended for translation. Limiting the dictionary to specifically exclude potentially dangerous, complex, or confusing words as well as forbidden abbreviations would further identify troublesome language to the author, and would improve communication overall. Implementing such real-time feedback to authors of patient instructions is a logical next step in adding utility to the electronic health record.

It is important that culture and contextual understanding is taken into consideration while organizations use interpretation services. In the United States, federal law requires that language interpreters employed by health care organization receiving federal funds are not only bilingual but also bicultural [16]. We did not find examples of dangerous synonyms being misapplied in translation, but we cannot rule out the possibility that such errors can occur. This is beyond the scope of typical machine translation software.

Our data suggest that use of medication names and dosing frequencies should not be repeated in the PCI where confusion can arise from imprecise language translation. Translation ambiguities that generate safety concerns in PCI might be mitigated by moving such content into structured areas of the AVS.

Conclusion

This study suggests that 9 times out of 10, the quality of machine translation using Google Translate is acceptable in terms of quality and safety. Currently, machine back translation may fail to reveal a relatively poor translation from English to Spanish. This study showed that increasing sentence complexity, as measured by the reading level index, was associated with a significant ($P < 0.05$) increase in unsafe machine translation. Similarly, including medication instructions in machine translations were associated with increased risk ($P < 0.05$) of machine translation safety error in this study.

A simple way to improve communication now would be to display the reading level to authors of patient communication content in real time, and limit the dictionary of acceptable words to forbid the use of known ambiguous terms or forbidden abbreviations. This would teach authors to use simple language, and increase the chance that translation (either human or machine) would be effective. This preliminary study suggests that keeping medication dosing instructions in a structured format is advisable, as is keeping sentences simple. As with spoken language [4], starting with clear, simple to understand English instructions provides the best machine translations into Spanish.

The Clinical Machine Translation Study Group: Todd W. Altenbernd, Steven Bedrick, Mark D. Berg, Nerida Berrios, Mark A. Brown, Colleen K. Cagno, Charles B. Cairns, Elizabeth Calhoun, Raymond Carmody, Tara F. Carr, Clara Choo, Melissa L. Cox, Janiel Cragun, Rachel E.M. Cramton, Paola Davis, Archita Desai, Sarah M. Desoky, Sean Elliot, Mindi J. Fain, Albert Fiorello, Hillary Franke, Kimberly Gerhart, Victor Jose Gonzalez, Aaron John Goshinska, Lynn M. Gries, Erin M. Harvey, Karen Herbst, Elizabeth Juneman, Lauren Marie Imbornoni, Anita Koshy, Lisa Laughlin, Christina M. Laukaitis, Kwan Lee, Hong Lei, Joseph M. Miller, Prashanthinie Mohan, Wayne J. Morgan, Jarrod Mosier, Leigh A. Neumayer, Valentine Nfonsam, Vivienne Ng, Terence O'Keeffe, Merri Pendergrass, Jessie M. Pettit, John Leander Po, Claudia Marie Prospero Ponce, Sydney Rice, Marie Anoushka Ricker, Arielle E. Rubin, Robert J. Segal, Aurora A.G. Selpides, Whitney A. Smith, Jordana M. Smith, William Stevenson, Amy N. Sussman, Ole J. Thienhaus, Patrick Tsai, J. Daniel Twelker, Richard Wahl, Jillian Wang, Mingwu Wang, Samuel C. Werner, Mark D. Wheeler, Jason Wild, Sun Kun Yi, Karl Andrew Yousef, Le Yu.

Corresponding author: Joseph M. Miller, MD, MPH, Department of Ophthalmology and Vision Science, University of Arizona, 655 North Alvernon Way, Suite 108, Tucson AZ 85711, jmiller@eyes.arizona.edu.

Financial disclosures: None.

References

1. Hummel J, Evans P. Providing clinical summaries to patients after each office visit: a technical guide. Qualis Health 2012. Accessed 14 Mar 2016 at http://hit.qualishealth.org/sites/default/files/hit.qualishealth.org/Providing-Clinical-Summaries-0712.pdf.
2. Neuberger M, Dontje K, Holzman G, et al. Examination of office visit patient preferences for the after-visit summary (AVS). Persp Health Infor Manage 2014;11:1d.
3. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015;17:e44.
4. Schoonover, K. Using a medical interpreter with persons of limited English proficiency. J Clin Outcomes Manage 2016;23:567–75.
5. Shin HB, Bruno R. Language use and English-speaking ability: 2000. Census 2000 Brief. Accessed 9 Nov 2017 at https://census.gov/content/dam/Census/library/publications/2013/acs/acs-22.pdf.
6. Lewis MP, Simons GF, Fennig CD, editors. Ethnologue: languages of the Americas and the Pacific. 19th ed. Dallas: Sil International; 2016.
7. Pavlik V, Brown AE, Nash S, et al. Association of patient recall, satisfaction, and adherence to content of an electronic health record (EHR)-generated after visit summary: a randomized clinical trial. J Am Board Fam Med 2014;27:209–18.
8. Johnson M, Schuster M, Le QV, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. Accessed 9 Nov 2017 at https://arxiv.org/pdf/1611.04558.pdf.
9. Patil S, Davies P. Use of Google Translate in medical communication: evaluation of accuracy. BMJ 2014;349:g7392.
10. Kaliyadan F, Gopinathan Pillai S. The use of Google language tools as an interpretation aid in cross-cultural doc-

tor-patient interaction: a pilot study. Inform Prim Care 2010;18:141–3. 11. Zhang Y, Zhou S, Zhang Z, et al. Rumor evolution in social networks. Physical Review E 2013;87.

12. Shingenobu T. Evaluation and usability of back translation for intercultural communication. In: N. Aykin, editor. Usability and internationalization. Global and local user interfaces. UI-HCII 2007, Lecture Notes in Computer Science, vol 4560. Springer, Berlin, Heidelberg.

13. Kincaid JP, Fishburne Jr RP, Rogers RL, et al. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch. 1975. Accessed 7 May 2016 at http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf.

14. Kwiecien R, Kopp-Schneider A, Blettner M. Concordance analysis—part 16 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011;108:515–21.

15. Goodall N. Ethical decision making during automated vehicle crashes. Transportation Research Record: Journal of the Transportation Research Board 2014;2424:58–65.

16. Kalra N, Groves D. The enemy of good: estimating the cost of waiting for nearly perfect automated vehicles. Santa Monica, CA: RAND Corporation, 2017.

17. Wu DT, Hanauer DA., Mei Q, et al. Assessing the readability of ClinicalTrials.gov. J Am Med Inform Assoc 2016;23:269–75.

18. Responses to: Use of Google Translate in medical communication: evaluation of accuracy. BMJ 2014;349:g7392 Accessed 10 Dec 2017 at www.bmj.com/content/349/bmj.g7392/rapid-responses.

18. Nápoles AM, Santoyo-Olsson J, Karliner LS, et al. Inaccurate language interpretation and its clinical significance in the medical encounters of Spanish-speaking Latinos. Med Care 2015;53:940–7.