Final Report.

Title: A clinical trial to validate an automated online language interpreting tool with Hispanic patients who have limited English proficiency

PI: Yellowlees PM

Team Members: Burke Parish M, Iosif A-M, Gonzalez A, Fisher A, Chan S, Martini J, Sciolla A. Chun R, Tougas H, Shahrvini T.

University of California Davis
Department of Psychiatry and Behavioral Sciences.

Project Dates Sept 2016 – June 2022

Federal Project Officers: Derrick Wyatt, Sheena Patel.

This grant was fully funded by AHRQ whose support is gratefully acknowledged

AHRQ. R01 HS024949

Abstract (248 words)

Purpose: Patients with Limited English Proficiency frequently receive substandard healthcare. Asynchronous Telepsychiatry (ATP) has been established as a clinically valid method for psychiatric assessments. The addition of automated speech recognition and automated machine translation technologies to ATP may be a viable artificial intelligence language interpretation option.

Scope: This project involved measuring the accuracy of translation of simple language formats and sophisticated figurative language devices (FLDs) and a clinical trial to test patient satisfaction with the ATP app and provider diagnostic reliability.

Methods: The ATP app was built. 114 patients with chronic psychiatric or physical conditions underwent two assessments, once by an English-speaking psychiatrist through a Spanish-speaking human interpreter and once in Spanish by a trained mental health interviewer-researcher with AI-interpretation on the ATP app. The accuracy of language translation engines and their capacity to assess FLD's were compared as was patient satisfaction and diagnostic inter-rater reliability across providers.

Results: Google was more accurate than Microsoft at basic translation. Both human and AI-interpreted FLDs were frequently translated inaccurately, while human-interpreted interviews were found to have a significant reduction in the use of FLDs and patient word count per minute; FLD translation was more accurate on videoconferencing. Patients preferred being interviewed in their own language on the ATP app. Provider diagnostic inter-rater reliability is still being statistically assessed. Using videoconferencing for human interpreting may be more accurate than in-person interpreting but automated translation of sophisticated language, as commonly used in psychiatric interviews, is not yet accurate enough for clinical purposes.

Key Words: Limited English Proficiency, Asynchronous Telepsychiatry, Speech Recognition, Machine Translation, Automated Intelligence, Mental Health, Interpretation, Figurative Language Devices.

A. Purpose of Study

The specific aims of the study were:

- *Aim 1:* To iteratively evaluate and refine the automated asynchronous interpretation tool already developed in phase one. (Completed – screen shots shown below)
- *Aim 2:* To compare patient satisfaction of Method A vs. Method B. (Still undergoing statistical evaluation although some results presented below)
- *Aim 3:* To compare the diagnostic accuracy and psychiatrist inter-rater reliability of Method A vs. Method B and demonstrate psychiatrist inter-rater reliability for Method B. (Still undergoing statistical evaluation)
- *Aim 4:* To compare the interview and language interpretation quality and accuracy of Method A vs. Method B. (Completed – 2 papers written, 1 of which is published)

B. Scope

There is a pressing national need to provide higher quality, more effectively accessible language interpretation services to improve the health outcomes of the 4.7% of the US population who have limited English proficiency (LEP) and who currently, as a result, have increased rates of hospital admissions, misdiagnosis, improper treatment and poorer health comprehension and outcomes. This project addresses a critical component of this problem: the need to improve access to high quality, mental health services for diverse populations by improving the flow of clinical work across care settings (primary care and specialty care) through the use of online asynchronous methods of communicating. In prior studies we created and demonstrated an efficient, provider compatible, administratively simple health IT solution: Asynchronous Telepsychiatry (ATP). The next stage in this study was to build an automated language translation process into our ATP consultation software to allow clinical evaluations to occur across languages without the use of human interpreters.

This is particularly important in California where 28.6% of the population speak Spanish as a first language at home, and 39% of the population in 2020 reported being of either Hispanic or Latino ethnicity, the largest single ethnicity group in the state. Unfortunately, the current proportion of physicians in California does not reflect this population and, while medical and nursing schools are trying to recruit more medical and nursing students with Hispanic heritage, there is still a major deficit of bilingual healthcare providers. All health systems are mandated to provide interpreters for any patient who has LEP but as this is an unfunded mandate it is common for health providers to be unable to access them. This is especially the case if the LEP patients are being seen in small (often rural) healthcare clinics or environments that are not able to internally fund significant numbers of interpreters. Some parts of California are well known for their extremely diverse populations and in Sacramento, where this study is occurring, it is estimated that over 90 languages are spoken by the population.

C. Methods

The first stage involved the building and development of our asynchronous telepsychiatry tool, the ATP App, with iterative improvements occurring during years 1-4. The second stage (years 2 through 6) of a 6-year project funded by AHRQ comprised of studies on the accuracy of the machine transcription and translation, to see if it could be used for clinical

purposes, and a Randomized Clinical Trial and analysis of video recorded data from recruitment of patients.

This grant therefore involved three separate studies:

1. Assessment of accuracy of simple words transcription and translation mostly sufficient for short medical interviews
2. Assessment of accuracy of sophisticated language transcription and translation (figurative language such as metaphors and similes) required for psychiatric interviews
3. The overall RCT described below and the focus of most of the outcomes

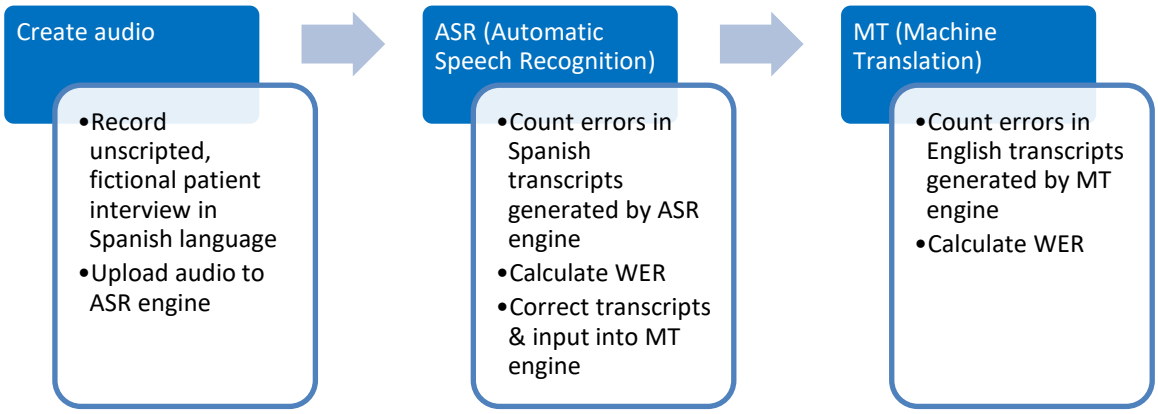
Study 1. Assessment of accuracy of simple words transcription and translation mostly sufficient for short medical interviews

We recorded two brief semi-structured fictional patient interviews in Spanish, focused on two common mental health disorders: video 1 captured a mock clinical encounter discussing anxiety symptoms, and video 2 covered a discussion of depressive symptoms. These Spanish-language interviews were recorded to video (including audio) files using a laptop. To ensure high quality audio, we added an extra lapel microphone for each speaker. We created these fictional mock interviews to allow us to test the Google and Microsoft ASR and MT engines without compromising real patient information.

We compared Microsoft Translator and Google Translate as MT engines widely and freely available to the public as desktop versions and mobile apps. These engines are considered to be two of the world's leading translation applications although they differ in strengths and limitations. As of October 2020, Google supported 108 languages and Microsoft supported 76 languages. Both MT's use a system called Neural Machine Translation (NMT). NMT uses deep learning algorithms to translate whole sentences at a time, which has been shown to be more accurate than translating individual words. In addition, Microsoft supports text and voice translations within a group of people rather than just two people at a time. Both have a mobile app-based interface, a web version, and offline downloadable apps that operate without a connection to the Internet.

After the clinical interview recordings, we uploaded the combined video and audio files to the two web apps to compare the engines. YouTube Studio was used to access Google's engine. ATP App, a cloud-based application developed by the UC Davis team in conjunction with the commercial app developer, Appstem, for the larger clinical trial, was used as the Microsoft engine. Both apps output English-language transcripts from the Spanish-language audio in almost real time, as well as audio and video files.

Process 1: Process from recording to transcript to translation



Step 1: Transcription

A bilingual research assistant analyzed raw un-translated Spanish-language transcripts generated by either Google Translate (in YouTube Studio) or Microsoft Translator (through our ATP App). The initial text edit involved using a media player to listen to the audio and then comparing the audio with the ASR output to complete a first edit for language inconsistencies between the audio and the ASR generated transcripts. Words were added, subtracted, or substituted as needed to correct discrepancies. Once the first edit of the ASR text was completed, the research assistant conducted a second more in-depth correction of the transcripts. This step included creating sentence boundaries, necessary because of the occasional overlap in audio between the interviewer and the interviewee on the transcribed audio. Correction of capitalization, correction of grammar, and insertion of accent marks were also necessary in this second correction of the output.

Step 2: Translation

After the in-depth correction of the ASR output, the corrected output text for both interviews was copied to each of the MT engines to generate language transcription output files. A bilingual research assistant evaluated the MT output for errors to determine Word Error Rate (WER), following the protocol established by Glen Flores et al. The research assistant counted the number of words *omitted* —the machine did not interpret a word/phrase uttered — and *added*, the machine added a word not uttered by the subject. *Semantic changes* were also noted and were marked in one of three ways: no differences in meaning, preserved meaning with unusual syntax, and meaning with significant differences from the original translation.

The two transcripts from each MT engine were compared and the Word Error Rate (WER) was calculated to assess for transcription errors; accuracy rate was used to assess for translation errors.
WER and accuracy were calculated using the following formulas:
$WER = (I + D + S)/ N$
$Accuracy = (N-D-S)/ N$
In these formulas, I represented inserted words, D for deleted words, S for drastic words that substituted words and changed the meaning of the sentence, and N for the total of words spoken.

Study 2. Assessment of accuracy of sophisticated language transcription and translation (figurative language such as metaphors and similes) required for psychiatric interviews

Six patients with psychiatric disorders were randomly selected from the original study of 114 patients. The first three patients were recruited prior to the COVID-19 pandemic and the second three patients were recruited after the start of the pandemic. This allowed us to assess if the transition to a virtual, Zoom platform would impact AI-interpretation.

Transcription and Translation

Transcripts for both methods were generated from the video/audio-recording of each interview. These transcripts were initially generated automatically and were then subsequently verified for accuracy and edited by two bilingual researchers. The verification process was a labor-intensive process, requiring each reviewer to replay the file multiple times to add, remove, and replace words. The process of transcript verification required approximately four minutes of editing per one minute of the interview. Instances of use of Figurative Language Devices (FLDs) spoken by the patient were then separately marked by two bilingual researchers. There are a wide variety of FLDs — such as similes, metaphors, irony, idiomatic expressions, and euphemisms — all of which apply language in a non-literal manner to add connotation. Table 1 presents examples for some common types of FLDs. FLDs used by the interviewers were excluded from analysis to control for natural variation in the style of speech used by the interviewers.

Accuracy of transcription and translation of each FLD was independently determined by two bilingual researchers. If an FLD was categorized as an inaccurate transcription, the FLD was marked as "transcript inaccurate" and no subsequent analysis of translation was made, as translation is dependent on accurate transcription. If an FLD was categorized as an *accurate transcription*, the FLD was then subdivided into either an accurate or an inaccurate translation.

To analyze the quantity of patient speech, separate sub-transcripts were created of only the patients' speech to obtain a patient word count. This word count was then divided by the minutes of the interview, to control for varying lengths of interviews. The number of instances of FLDs was divided by the number of minutes of the interview. This was done in order to control for the varying lengths of patient interviews, and the limitation on word count that is seen in Method A, as the time taken for the interpreter to translate necessarily reduced the amount of time that the patient can be speaking and using FLDs.

Statistical analyses for FLD frequency and patient word count per minute were performed using Excel with paired-sample t-tests between Method A and Method B for each patient. Statistical analyses of aggregate translational accuracy of FLDs for in-person, pre-COVID vs Zoom, post-COVID groups were performed using Excel with independent samples t-tests. $P < 0.05$ was used to determine significance for all analyses.

Study 3. The overall RCT described below was the focus of most of the outcomes of the grant

Patients were recruited from local primary care clinics in Sacramento. This proved to be a very difficult process. We attempted initially to engage all Spanish speaking primary care providers and have them refer patients, but only one such physician referred any of his patients, with the rest saying it took too much time. We then set up research assistants in the clinics so that they could directly approach Spanish speaking patients identified by the clinic staff, but this proved to be very slow and inefficient, leading to few actual referrals. Finally, we gained agreement to go through clinic lists to identify all patients who were Spanish speaking and send them letters about the study, followed by a phone call for screening. This did not work originally but when we employed a research assistant with specific experience in working in call centers doing cold calls, we finally had success and she was able to recruit many more patients who gave initial verbal consent on the phone than we had previously.

In this randomized cross-over study with patient recruitment occurring during years 2 through 4, 114 Spanish speaking patients were recruited from several outpatient clinics in Sacramento, California and received two psychiatric interviews described below.

In addition to these interviews, the patient completed a battery of questionnaires and a Spanish version of the Structured Clinical Interview for DSM (SCID). Both clinical interviews were digitally video recorded for comparative analysis of language and translation accuracy, as well as for repeat diagnostic and inter-rater reliability assessments.

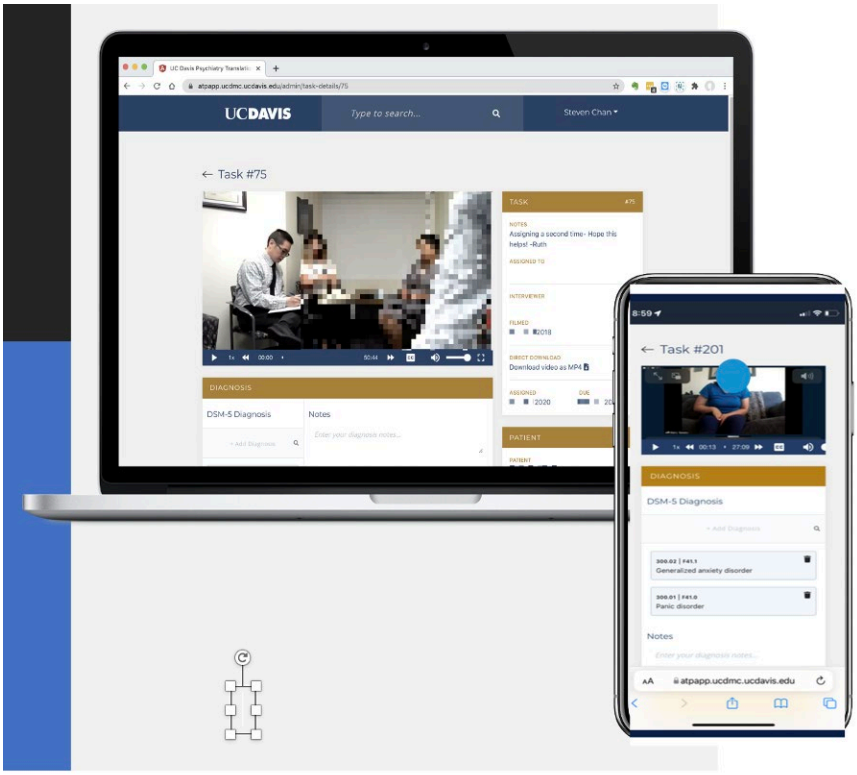
We compared two methods of cross-language psychiatric assessment:

- Method A (current gold standard of in-person/virtual real-time interpreting practice). A Spanish-speaking patient is diagnostically assessed in-person or virtually by an English-speaking psychiatrist through a Spanish-speaking interpreter.

- Method B (comparative practice – ATP). A Spanish-speaking patient is interviewed in Spanish by a trained mental health interviewer. The interview is recorded in real time, translated into English with sub-titles added to the video file, and sent to an English-speaking psychiatrist to asynchronously — that is, at a later time — review the video and make a diagnosis.

We had to halt our study recruitment for about eighteen months due to COVID-19, but when we were allowed to recommence, we modified our IRB protocol to allow us to see all our subjects via video. We had originally seen 91 patients in-person and instead of finishing our study at 100 patients as originally planned, we decided to expand our group to try and examine the impact of online care compared with in-person care. We then recruited another 23 participants for a total of 114 participants making us able to compare the 91 in person patients seen by both methods with the 23 virtual patient subsample also seen by both methods, on the primary outcome measures. This was an expansion of our research because of Covid-19.

We had study information in English and Spanish, had a bilingual physician (Dr. Odor – now deceased) as our project co-investigator, a bilingual psychiatrist (Dr. Sciolla), and a bilingual project coordinator to work in the clinics and manage day-to-day study issues.



Screenshot 1. The ATP platform built for this study during year 1

Inclusion and Exclusion Criteria

We recruited Hispanic individuals with significant LEP — a critically underserved population — comprising 5% of the local Hispanic population at the Sacramento County Adult Psychiatric Support Services (APSS) Clinic and other Northern California mental health and primary care clinics. The clinics are well served with bilingual staff, interpreters and documentation. As described above we received relatively few referrals initially using conventional referral approaches within the clinics, but once we gave patients information and followed up with Spanish language phone calls from a trained call center operator we were much more successful.

Participants were adults with significant LEP such that they prefer being interviewed in Spanish, aged 18 or older and referred by County clinical providers or those who self-referred via the flyers we distributed, or defined as being primarily Spanish speaking by clinic staff. We recruited two types of LEP Spanish-speaking patients:
1) patients who have a non-urgent psychiatric issue — a mood disorder, psychotic disorder, anxiety disorder, or substance or alcohol use disorder(s) — and/or
2) patients who have a chronic medical condition such as diabetes, cardiac or renal disease.

At the phone screening patients were asked whether they were primarily concerned about a behavioral/psychiatric issue or a chronic medical condition issue. Many patients had comorbid conditions and multiple diagnoses. These diagnoses were noted as would happen under routine clinical circumstances, and patients with multiple diagnoses were not excluded.

We excluded any patients less than 18 years of age, patients with imminent suicidal ideation and/or plans, patients who had immediate violent intentions or plans, patients who had significant cognitive deficits and any patient whose primary care provider or psychiatrist did not recommend participating. When we eventually moved to recruiting via clinic lists reviewed these lists with the patient's PCPs to ensure that we did not approach any patients that they felt would not meet our inclusion criteria.

Study Timelines

Year 1 was spent iteratively building and testing our software tool. We then continued to improve this over the next several years with the last technical changes being made in year 3 as some of the translation software was difficult to incorporate in a HIPAA compliant manner.

Patient recruitment occurred during years 2 through 6. It was very slow during year 2 and ceased completely for about 18 months when Covid-19 occurred in years 3-4, and then finished off in years 5 and the beginning of year 6. The second half of year 6 was spent exclusively on data analysis.

Study Endpoints

The entire RCT only involved one clinic visit of typically 3-5 hours at the end of which each patient was given a credit card valued at $75 for their compensation, as well as reimbursement of any travel costs. Given that the IRB evaluated the study as low risk and it only took a short time with no intervention being undertaken there was no need for formal study endpoints to be evaluated, and no reason to have a data safety monitoring board. We had only 3 patients fail to complete the study protocol once they had arrived post-screening. One changed her mind after consenting and left, one decided to leave after the first interview was completed and one was unable to undergo the study because of a technical problem with the laptop we used to record subjects. All 3 subjects were compensated as they had started the study although we treated them as study dropouts and their sparse results were not included in our data analysis. We did have several patients who became quite distressed during the interviews, usually in relation to past traumas that they had repressed, and these individuals were all assisted by our study psychiatrists although none of them was suicidal. In these instances, our Spanish speaking research assistant telephoned them the following day to check-in, and if they had a PCP, a copy of their study assessment was sent to that individual with a plan of management.

Outcomes

At baseline, all patients were evaluated in person using the SCID by our trained research physician. This is the "gold standard" diagnostic tool for psychiatric disorders and is widely used in psychiatric research.

Clinical outcome measurements (all presented in Spanish) focused on disorders likely to be seen in primary care, notably anxiety, depression and substance abuse, and include:

- The SF-12, which is a widely validated and used self-report health survey consisting of 12 questions that produces a functional health, well-being, physical and mental health summary;
- The PHQ-9 is a multipurpose instrument that is widely used for screening, diagnosing, monitoring and measuring the severity of depression (Scores greater than 9 have sensitivity and specificity of 88% for major depression);
- The GAD-7 is a widely used screening tool and severity measure for generalized anxiety disorder with high sensitivity and specificity;
- The Alcohol Use Disorders Identification Test (AUDIT) was developed for the World Health Organization to identify persons whose alcohol consumption has become hazardous or harmful to their health and has been widely used in many studies. The AUDIT takes under 2 minutes to administer and is commonly used in primary care. Based on our previous research we have found a great deal of substance abuse comorbid with other disorders.
- The Clinical Global Impression – Severity scale (CGI-S) is a 7-point scale that requires the clinician to rate the severity of the patient's illness at the time of assessment, relative to the clinician's past experience with patients who have the same diagnosis. The analogue scale that is routinely used by UCD psychiatrists as an outcome measure ranges from 1 (Normal) to 7 (Extremely ill, and likely to be an inpatient).

The satisfaction and Interview comparison measures were:

- Patient Telepsychiatry Satisfaction Questionnaire. The provider questionnaire was used in our preliminary studies. The patient questionnaire is a modified version of the Parent Telemedicine Satisfaction Survey.
- Patients reported their perceived quality of the interpretation using a modified version of a UC San Francisco-developed tool measuring satisfaction of patient-centered tools through of quality of communication and visit satisfaction.
- Patients were asked to quantify on a number of visual analogue rating scales of 0-10 the two interview methods in terms of their likeability and effectiveness, using questions developed using the STAR technique (Situation, Task, Action, Result) which assess the effectiveness of the interview process in the recruiting sector.

The satisfaction, quality of interpretation and interview comparison questions were integrated into a single document which all patients completed in 3 components after each of the interviews separately, and then after the second interview to enable forced choice comparison of the two interview methods. The questionnaires, self-report measures and interviews were conducted in person or electronically by our trained clinic provider in Spanish. The clinical functioning CGI was completed by the interviewing psychiatrists.

From a statistical perspective we will assess/did assess:

- Patient satisfaction and preference for the two methods;
- provider diagnostic accuracy, inter-rater reliability and other psychiatrist related clinical outcome measures using Kappa as in our previous research
- total encounter time, comprising of time for the recorded patient-clinician encounter and time for the consulting psychiatrist to view, evaluate, and write assessments for the patient

- language accuracy and cultural competence: All 228 interviews in both methods were video-recorded and a subset of these interviews was randomly selected to assess the accuracy of language translation, first at a basic level, (See summary below in a paper which is unpublished), and secondly in a more sophisticated manner examining higher level language accuracy, such as the use of synonyms and metaphors (See JMIR published abstract below).

D. Results

Study 1

Results

As noted above we examined the two steps of the translation process for all videos – both transcription and translation. Obviously if the initial transcription process is inaccurate, then these inaccuracies become magnified in the translation stage. The results were as follows:

Step 1: Transcription

Video 1: the anxiety video at 6 minutes and 45 seconds duration contained a total of 928 words, yielded a total of 162 errors when run by the Microsoft engine, and 138 errors when run by the Google engine.

Video 2: the depression video at 8 minutes and 45 seconds duration contained a total of 1122 words, yielding 503 errors when run by the Microsoft engine and 141 errors when run by the Google engine.

To obtain the transcription WER and Accuracy for each engine, we combined the data from both the anxiety and depression videos, adding up to a total of 1526 words with Microsoft engine and 1875 with Google engine and calculated the following rates:

$$\text{Microsoft WER} = \frac{48+579+61}{1526} \rightarrow 0.45 \qquad \text{Microsoft Accuracy} = \frac{1526-579-61}{1526} \rightarrow 0.58$$

$$\text{Google WER} = \frac{56+198+102}{1875} \rightarrow 0.19 \qquad \text{Google Accuracy} = \frac{1875-198-102}{1875} \rightarrow 0.84$$

Step 2: Translation

Video 1: the anxiety video contained 147 errors when run by the Microsoft translator and 94 errors with Google Translate.

Video 2: the depression video contained a total of 222 errors in Microsoft Translator and 127 with Google Translate. Data from both videos was combined to calculate the accuracy of each engine as follows:

$$\text{Microsoft WER} = \frac{39+84+178}{2057} \rightarrow 0.15 \qquad \text{Microsoft Accuracy} = \frac{2057-84-178}{2057} \rightarrow 0.87$$

$$\text{Google WER} = \frac{37+50+68}{2157} \rightarrow 0.07 \qquad \text{Google Accuracy} = \frac{2157-50-68}{2157} \rightarrow 0.95$$

Table 1: Transcript (ASR) and Translation (MT) assessment

A.

	Transcript (ASR)	
Error Type:	**Microsoft**	**Google**
N=	**1526**	**1875**
Deleted Words (D)	579	198
Added Words (I)	48	56
Drastic Words (S)	61	102
Unusual Words	9	6
Punctuation	18	1
Fillers	13	5
Overlapped Sentences	20	4
Total Error Count	748	**3**72
WER	***0.45***	**0.19**
Accuracy	***0.58***	***0.84***

B.

	Translation (MT)	
	Microsoft	**Google**
N=	**2057**	**2157**
Error Type:		
Deleted Words (D)	84	50
Added Words (I)	39	37
Drastic Words (S)	178	68
Unusual Words	49	51
Punctuation	15	11
Fillers	4	4
Total Error Count	***369***	**221**
WER	***0.15***	***0.07***
Accuracy	***0.87***	**0.95**

Study 2

Participants included 4 females and 2 males, with a range of 42-71 and an average of 53 years. 4 participants were born in Mexico, 1 in Costa Rica and 1 in Guatemala [3] .

Figure 1 shows the number of FLDs per minute, and the mean for each method. Figure 2 shows the patient word count per minute, and the mean for each method. There was a significant increase in the per minute frequency of FLDs using AI-interpretation *(M = 0.61, SD = 0.26)* compared to using the human interpreter *(M = 0.2, SD = 0.10),* (t(5) = -4.59, p = < .05). There was a significant increase in the per minute patient word count using AI-interpretation *(M = 86.9, SD = 29.49)* compared to using the human interpreter *(M = 48.8, SD = 17.53)*, (t(5) = -3.31, p = .02).

Figure 1. Figurative Language Devices per minute

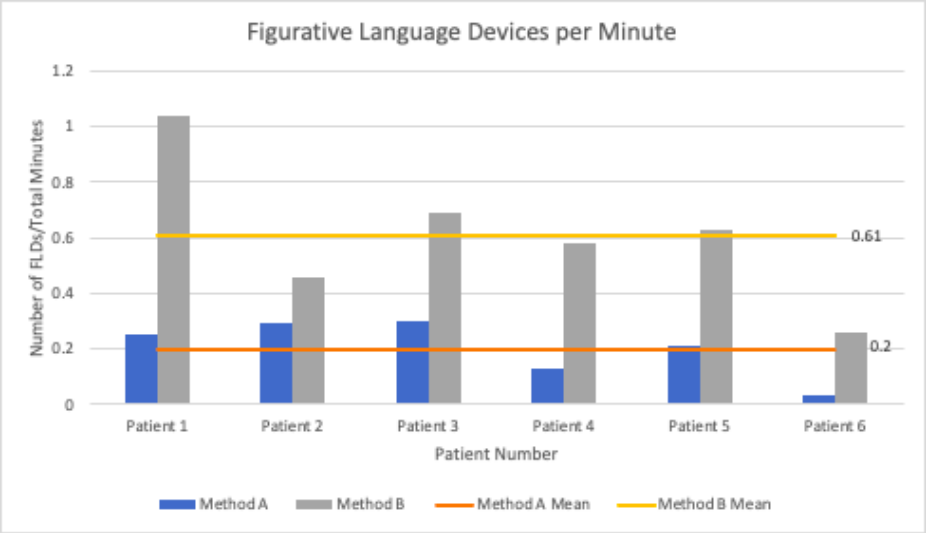


Figure 2. Patient word count per minute

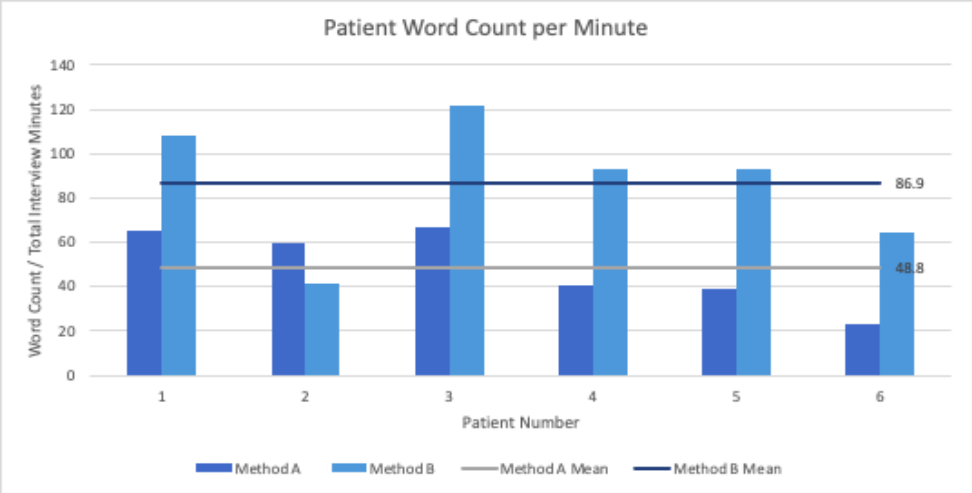


Figure 3. Percentage Accurate Translation

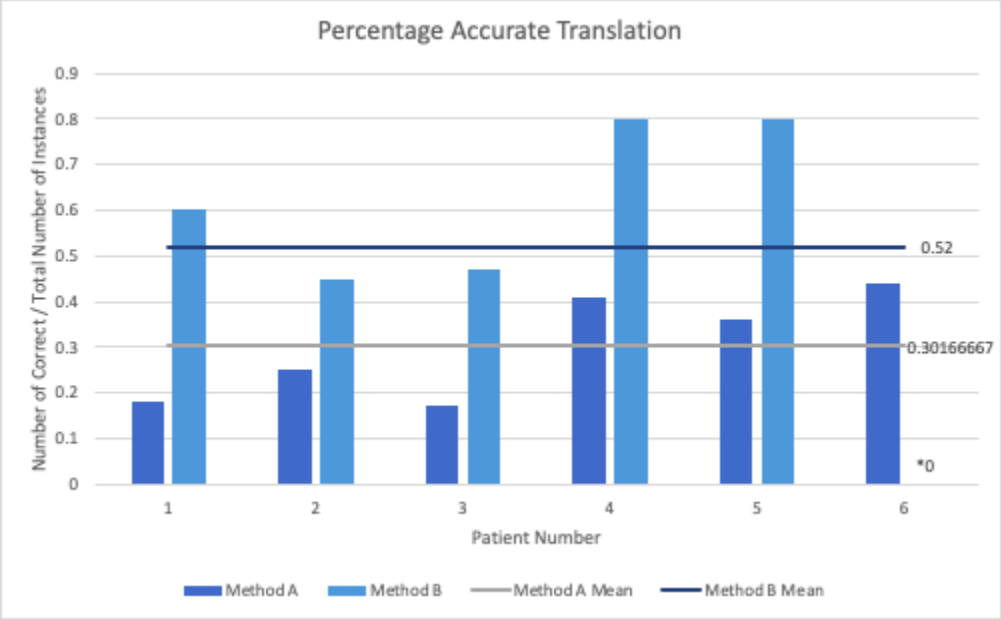


Figure 3 demonstrates the results of the percentage of accurate translation of FLDs, taken as the number of accurate translation instances over the total number of instances, and the mean for each interview, over all participants [5].

There was an insignificant decrease in the mean percentage of accurate translation of FLDs using AI-interpretation *(M = 0.28, SD = 0.09)* compared to using human interpretation *(M = 0.52, SD = 0.29),* (t(5) = -1.6, p = 0.17).

Figure 4. Aggregate Percentage Correct

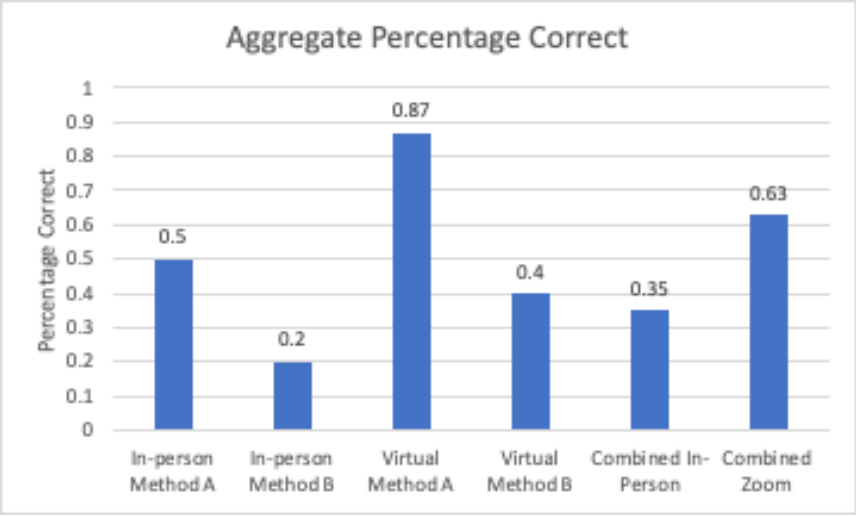


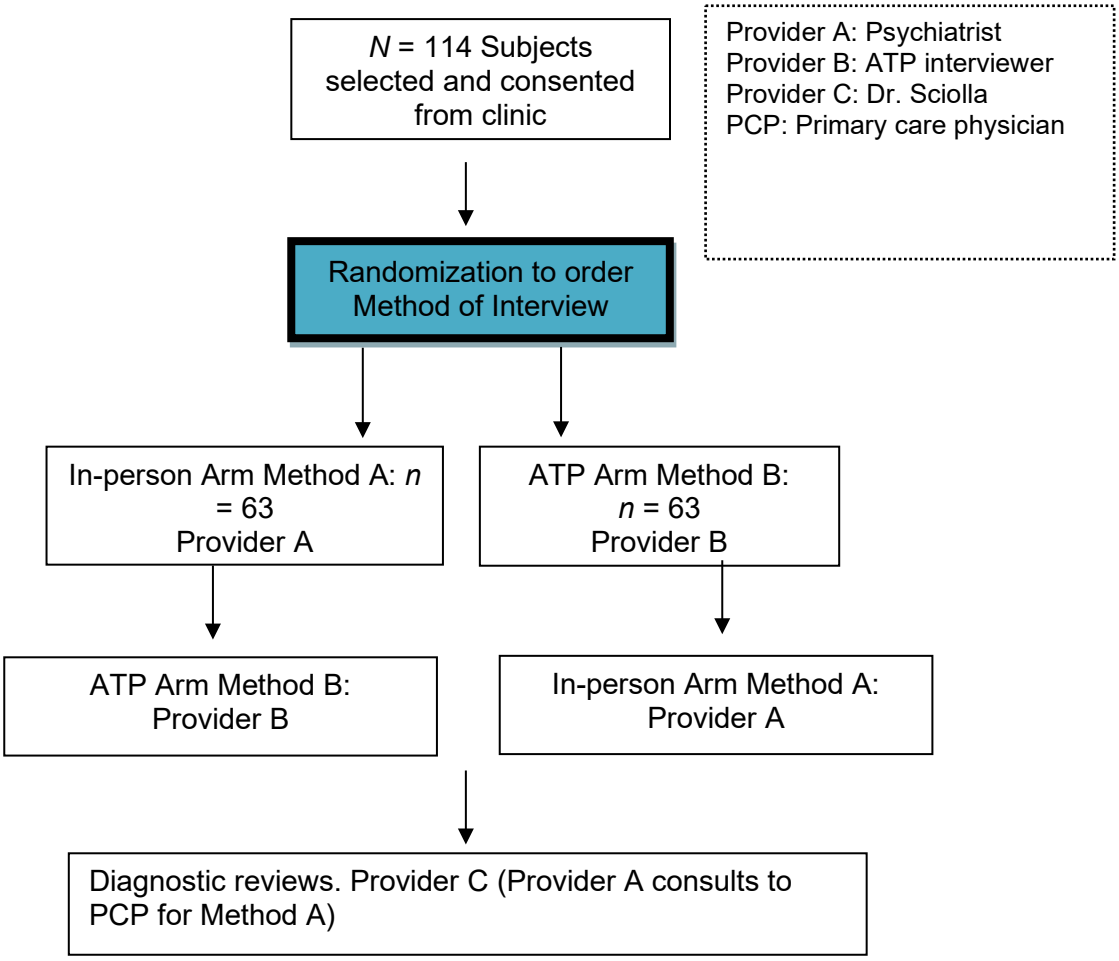
Figure 4 demonstrates the aggregate percentage of accurate translation for patients 1-3 in the in-person format, and for patients 4-6 on Zoom, and as grouped by method.

There was a significant increase in the mean percentage of accurate translation of FLDs across both methods in the virtual Zoom format (*M = 0.63, SD = 0.27)* compared to the in-person format prior to the COVID-19 pandemic *(M = 0.35, SD = 0.18),* (t(10) = -2.15, p = .03).

Study 3

The RCT results to this point – analysis is incomplete and ongoing. Details of the subject flow are shown in the consort diagram.

Consort Diagram. Outcome Comparison of 2 Interview Methods Study Design



The only analysis we have performed so far is a quick review of the patients global forced choice preference for ATP or Interpreter interviews by asking them 3 questions at the end of their study session.

1. Which of the two interviews did you prefer overall?
   Result. ATP 30%; Interpreter 13%; No preference 57%

2. In which of the two interviews today do you think that your medical or psychiatric problems were better understood?
   Result. ATP 28%; Interpreter 20%; No preference 52%

3. In future, if you had to choose between these two interviews, which would you prefer?
   Result. ATP 50%; Interpreter 22%; No preference 28%

Discussion

Our comments below are early and non-conclusive with respect to the RCT as much of the data analysis on this part of the study has still to be completed. We can however draw several conclusions from the first two studies, and in particular from the second one which included data taken using both in-person and online (zoom) interviews.

We found that the two language engines were likely accurate enough to be used for simple medical interviews (study 1) but were unable to translate figurative language devices often used in psychiatric disorders with a degree of accuracy that would make them clinically accurate at this time, although they will undoubtably be accurate enough in future as the artificial intelligence behind them improves.

We found that patients' speech differs significantly: Method A in the presence of a human interpreter showed fewer instances of FLDs, compared with Method B with language-concordant interviews augmented with AI-interpretation. Additionally, in Method A patients spoke with a lower word count per minute — a more than 50% reduction in rate — compared to Method B. There was no significant change in these results when using videoconferencing, compared to in-person consultations, although the interpreting accuracy over videoconferencing was improved for both methods.

Our findings aligned with our expectation that patient speech becomes simplified and truncated when using a human interpreter. This simplification aligns with many published guidelines and articles that detail best practices for use of human interpreting services, which often encourage a reduction in the use of idiomatic speech, as well as a simplification of sentence structure. Within the specialty of psychiatry, diagnosis and treatment decisions are heavily reliant on the verbal history conveyed to the provider. Our results suggest that the history provided through use of a human interpreter will likely differ and could represent a less comprehensive picture of the patient's psychopathology. Of note, human interpreting services guidelines are generally geared towards providers rather than patients, and the patients included in our study would likely not have read these guidelines prior to the study. Instead, we propose that there is an innate tendency for the patients to simplify their speech when having to pause between sentences to allow for translation. The use of a human interpreter has additionally previously been associated with a

reduced number of follow-up appointments, reduced patient and provider satisfaction, and an increased likelihood of not asking the questions that the patient wanted to ask.

The results of our study additionally demonstrate that the use of an in-person human interpreter (Method A) is, at this time, more accurate than AI-interpretation (Method B) with regard to the translation of FLDs. The aggregate translational accuracy for human interpreters was 52% versus 28% for AI-interpretation (p-value > 0.05), suggesting that both methods lend themselves to high degrees of inaccuracy when translating FLDs. Of note, a sizable contribution to the inaccuracy of translation by the AMT starts from an inaccurate transcription of the conversation, suggesting that improvements in audio-recording and transcription would increase the translational accuracy of the AI-interpretation.

Finally, our results demonstrate that the transition of interviews from in-person to the virtual Zoom format in response to the COVID-19 pandemic led to a higher percentage of translational accuracy of FLDs. The aggregate translational accuracy of Method A is 50% in-person vs 75% over Zoom, and the aggregate translational accuracy of Method B is 20% in-person vs 39% over Zoom [10]. This difference primarily stems from an improvement in *transcriptional* accuracy on the Zoom format, likely seen because interview participants took longer pauses after speaking and spoke in shorter phrases over the Zoom format.

There are a number of limitations that we have identified in this study, primarily the fact that we are still analyzing most of the results of the RCT. First, the results discussed above are limited due to the small panel of patient interviews that are included. The decision to analyze a limited subset of 12 patient interviews from the initial cohort of 228 patient interviews was made due to the significant time required to both generate transcriptions for the in-person Method A, and to verify the machine generated transcripts for accuracy for Method B. Expanding the sample size of included patient interviews is possible in the future using our database of recorded interviews but will be time consuming. This study is additionally limited by the wide variety of types of FLDs used in the interview discourse. Some devices, such as idioms and metaphors, are clear to delineate from non-figurative speech. For example, the following patient statement, 'estoy viendo una luz al final del túnel,' or '*I am seeing a light at the end of the tunnel'* is clear to recognize as a figurative language device; it is well understood that the patient is not actually seeing a light, but rather that they are using an idiom that is in common use in both the English and the Spanish languages, to describe feeling a sense of hope after a period without such hope. By contrast, some of the types of devices that are used less frequently — such as personification and euphemism — are more subtle. For example, the following patient statement, 'la enfermedad me hizo traermelo para acá,' or 'the sickness made me bring him too' is less obvious to recognize as figurative language, whereby her depression — "the sickness" — is personified to have forced the patient to do something.

This study was characterized by a number of practical difficulties and barriers which resulted in considerable "lessons learned." These include the following:

A. Recruitment was difficult. We approached multiple clinics and conventional recruitment approaches through MD referrals and study staff attending the clinics in person failed. We finally were successful by going through the entire clinic patient lists and selecting out patients whose providers agreed to their involvement, who we then sent letters of

introduction to and followed up with cold phone calls in Spanish to discuss the study and do initial verbal screening and consents. We were enormously helped by employing a Spanish speaking research assistant with call center experience.

B. Many of our Hispanic patients, especially those who were undocumented and who had fled to the USA often 20-30 years previously, had horrendous trauma and abuse histories in their childhood that they had never discussed with anyone, and for which they had never received therapy. This meant that a number of the interviews were much more difficult and intense than we had expected. These interviews had an especially chilling effect on some of our interpreters, several of whom ended up in tears and had to take breaks during the interviews.
C. Covid was a challenge that was ultimately helpful although we had to stop recruitment for 18 months. When we finally returned to recruiting, we deliberately over-recruited from our original number of 100 patients to create a reasonable number of patients seen online (23) and compare these with the in-person interviewees (91).
D. Patients with chronic medical problems often had unrecognized psychiatric disorders, especially those with chronic renal disease who often exhibited untreated symptoms of depression, and who were advised to go to see their primary care physicians who we sent consultation notes to.
E. Performing cross language studies is time consuming and much more expensive and we were certainly not properly funded for this study, with several of the study team working many more hours than were paid for. We were ultimately quite severely under-funded for this study, not just because we had to manage Covid-19 and deliberately over-recruited, but because of the enormous difficulties of recruitment of Spanish speaking patients (often undocumented, traumatized, impoverished and untrusting). Running a study where staffing and documentation are all in Spanish is simply more costly than otherwise and being fully aware of the patients cultural issues is essential for both recruitment and the running of the study data collection process.

Going forward, technological improvements of AI-interpretation from both the transcription component and the translation component will be required for ATP interviews to be conducted in languages other than English. The field of AI-interpretation has made substantial progress within the past decade with the transition from statistical machine translation to neural machine translation; we expect that AI-interpretation will continue to expand and improve in the coming years and to eventually be at least as accurate as professional interpreters, allowing it to be introduced into regular clinical use. As our patient population in the US continues to diversify, it will be important to further develop novel technological approaches to allow LEP patients to engage with their providers without the inequality of time and attention to detail, as well as language simplification, that human interpretation currently entails. Further studies of the accuracy of interpretation over videoconferencing compared with in-person interpreting are required.

We have arrived at several conclusions at this stage of our data analysis which have important implications going forward:

1. ATP and automated translation is feasible with Spanish speaking patients
2. Google Translate is more accurate than Microsoft Translator (85-90% literal accuracy) which is probably sufficient for simple interviews

3. When using medical interpreters for long interviews, patients speak in “pidgin” language, often encouraged by interpreters, using less words and less figurative language devices than when speaking to a natural language interviewer
4. Interpreters are slightly more accurate at figurative language translation than automated systems, and are more accurate on virtual consults than they are in-person.
5. Patients prefer ATP consultations in their own language than the use of interpreters.

E. List of publications and products.

One paper is published:

Hailee Tougas , Steven Chan , Tara Shahrvini , Alvaro Gonzalez , Ruth Chun Reyes , Michelle Burke Parish , Peter Yellowlees. The Use of Automated Machine Translation to Translate Figurative Language in a Clinical Setting: Analysis of a Convenience Sample of Patients Drawn From a Randomized Controlled Trial. JMIR Ment Health 2022 Sep 6;9(9):e39556. doi: 10.2196/39556.

One paper is under review:
Steven Chan, Ruth Chun, Alvaro Gonzalez, Michelle Burke Parish, Peter Yellowlees
Automated Machine Translation systems may be sufficiently accurate to translate psychiatric interviews