International Forum of Allergy & Rhinology

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Translation and Cross-Cultural Adaptation of the Chronic Rhinosinusitis Control Test for Global Use

Hye K. Pae[1] | Detong Xia[2] | Hanzhong Sun[1] | Yudi Chen[3] | Kwangoh Yi[4] | Minjeong Song[5] | Eriko Sato[6] | Ali R. Abasi[7] | Jody Ballah[8] | Anna Babarczy[9] | Raymond Bertram[10] | Agnieszka Biernacka[11] | Mable Chan[12] | Teresa Civera[13] | Irina Dubinina[14] | Doğu Erdener[15] | María Isabel Maldonado García[16] | Ali Garib[17] | Tuomo Häikiö[10] | Fuk-chuen HO[18] | Li-Yu Hung[19] | R. Malatesha Joshi[20] | Oksana Kanerva[21] | Kiranpreet Kaur Baath[22] | Björn Köhnlein[23] | Dalibor Kučera[24] | Paula Luegi[25] | Yustinus Calvin Gai Mali[26] | Sivan Medina[27] | Stefan Milosavljević[28] | Amna Mirza[29] | Mohamed Y. Mwamzandi[30] | Fatemeh Nami[31] | Anabella-Gloria Niculescu-Gorpin[32] | Portia Padilla[33] | Georgia Panayiotou[34] | Manuel Perea[35] | Luciano Perondi[36] | Hiển Phạm[37] | Rasmus Puggaard-Rode[38] | Anurag Rimzhim[39] | Sreeparna Sarkar[6] | David L. Share[40] | Gláucia V. Silva[41] | Antônio R. M. Simões[42] | Charlotte Stormbom[43] | Titima Suthiwan[44] | Katsuo Tamaoka[45] | Mila Tasseva-Kurktchieva[46] | Paweł Urbanik[47] | An Van[1] | Katie M. Phillips[48] | Ahmad R. Sedaghat[48]

Correspondence: Hye K. Pae (hye.pae@uc.edu) | Ahmad R. Sedaghat (ahmad.sedaghat@uc.edu)

Received: 11 April 2026 | **Revised:** 29 April 2026 | **Accepted:** 7 May 2026

Keywords: AI | artificial intelligence | ChatGPT | chronic rhinosinusitis | chronic rhinosinusitis control test | claude | copilot | cross-cultural adaptation | perplexity | translation

ABSTRACT

Introduction: The Chronic Rhinosinusitis Control Test (CRCT) is a patient-reported outcome measure (PROM) written in English that is psychometrically validated to measure chronic rhinosinusitis control. Because the availability of translated PROMs is a driver of data equity—collection of data that is fair and generally representative—our objective was to create a library of translated, cross-culturally adapted versions of the CRCT that could ultimately be used for patients worldwide.

Methods: A hybrid approach leveraging generative artificial intelligence (genAI) in collaboration with expert human linguists was employed for translation and cross-cultural adaptation of the CRCT. For each target language, forward translations were performed with three large language models (LLMs) (ChatGPT, Copilot, and Perplexity) after which an expert human linguist provided additional revisions that were used to create a consensus final translation. Backward translations were performed using LLMs (Claude, Copilot, and Perplexity). The accuracy and validity of translations at each step were assessed qualitatively and quantitatively.

Results: The translation and cross-cultural adaptation of the CRCT was achieved into 37 languages: Arabic, Bengali, Brazilian Portuguese, Bulgarian, Cantonese Chinese, Czech, Danish, Dutch, European Portuguese, Filipino, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Mandarin Chinese, Norwegian, Persian, Polish, Punjabi, Romanian, Russian, Serbo-Croatian, Spanish, Swahili, Swedish, Thai, Turkish, Ukrainian, Urdu, and Vietnamese. These translated, cross-culturally adapted versions of the CRCT are made available in this article.

Conclusion: Translated, cross-culturally adapted versions of the CRCT developed in this study promote data equity by serving as a basis for psychometric validation of the CRCT for worldwide use.

1 | Introduction

For chronic diseases that persist and are often uncurable, disease control is an important and desired outcome, which reflects a disease state that is characterized by acceptability of its manifestiations rather than its resolution or eradication [1]. For chronic rhinosinusitis (CRS), control is an important goal of treatment, factor in medical decision making, and outcome measure [1–4]. Despite the long-standing significance of control in the evaluation and management of CRS, the means for measuring CRS control has been historically varied and inconsistent [5].

In order to consolidate support around a broadly accepted, consistent means for assessing CRS control, an international multidisciplinary study identified consensus criteria for CRS disease control [6]. The Chronic Rhinosinusitis Control Test (CRCT) [7], a patient-reported outcome measure (PROM), was subsequently developed based on these consensus criteria with additional input from patients and CRS experts around the world [8], and psychometrically validated to measure the construct of CRS control [7]. Importantly, because the CRCT score can be used to classify CRS control as controlled, partly controlled, and uncontrolled, it can be used to guide CRS treatment decisions.

At present, the CRCT is only validated in the English language. Given the universal importance of assessing, quantifying, and classifying CRS disease control, the psychometric validation of the CRCT in other languages may be of significant utility for CRS patients, healthcare providers, and researchers worldwide. This point is further amplified by the pressing need in biomedical research for data equity—the collection of data that is fair, generally representative, and inclusive of all groups, including marginalized ones—which can be impacted by the availability of translated data collection tools, such as PROMs like the CRCT [9–11]. However, before the CRCT can be psychometrically validated in another language, it must first undergo translation and cross-cultural adaptation, which can be time- and resource-intensive, requiring linguistic expertise that may not be readily available to all clinical researchers [12–14]. To remove this barrier, the objective of this study was to translate and cross-culturally adapt the CRCT into languages that would increase the accessibility of the instrument to CRS patients and their healthcare providers, as well as researchers globally. To achieve our objective, we used an established hybrid methodology that leveraged the growing power of generative artificial intelligence (genAI) with quality control and oversight by expert linguists [15]. Our result is a library of translated, cross-culturally adapted versions of the CRCT in 37 languages that span every populated continent, which we provide herein. Our hope is that the results of this study will reduce the barrier for the CRCT to be psychometrically validated widely for use by the maximum number of patients, healthcare providers, and researchers worldwide.

2 | Methods

2.1 | Study Design

The structure of our study design is shown in Figure 1 and uses a hybrid methodology partnering human linguists with genAI to translate and cross-culturally adapt the CRCT into 37 languages (Table 1). These languages were selected based on the worldwide prevalence of speakers [16, 17] and the potential for the usage of the translated CRCT instrument for CRS research globally. As the first step prior to initiation of the study, we carried out feasibility testing of genAI-supported CRCT translation to ensure the practicality of genAI-supported translation, its applicability to diverse languages, and achievability of successful translation. Chinese, Japanese, and Korean were used for the feasibility testing, given that each language has unique characteristics in both spoken and written forms [18]. The findings of the feasibility testing showed that genAI could be an effective complementary engine for collaboration with human linguists in the translation of the CRCT.

Our protocol followed a previously reported recommendation for a hybrid model for translation and cross-cultural adaptation of text using genAI with humans' linguistic quality check [15]. A step-by-step translation and adaptation process was developed and applied to the CRCT for each target language. As a brief overview, the psychometrically validated English version of the CRCT was forward translated using three genAI large language models (LLMs): ChatGPT [19], Copilot [20], and Perplexity [21]. To ensure the quality of these three forward translations, we invited human experts in linguistics to systematically assess and validate the three genAI translations in the selected languages using a standardized rubric. For each language, based on each linguist's assessments of the three translations, a translation evaluation committee consisting of the linguist and two other authors (H.K.P. and A.R.S.) assembled a consolidated final translation that underwent a genAI-supported backward translation into English using three independent LLMs: Copilot [20], Perplexity [21], and Claude [22]. All three backward translations were then evaluated for changes in meaning or other inaccuracies by the translation evaluation committee for each language. If the backward translation was not satisfactory, this process was performed iteratively to further refine the translation.

2.2 | Forward Translation Using Generative Artificial Intelligence

The psychometrically validated English version of the CRCT was forward translated using three LLMs (ChatGPT [19], Copilot [20], and Perplexity [21]) between November 19, 2025 and December 8, 2025. These LLMs were chosen based on the perceived popularity of the engines in the second half of 2025, on the authors' personal observations about their accuracy, LLMs used in the literature [23, 24], and on the assumption that one translation engine would not be enough. The previously validated English version of the CRCT was uploaded into each platform with the specific instructions: "Please translate the following questionnaire into formal [specified language] appropriate for the general public at a middle-school reading level." No additional post-translation processing was performed.

2.3 | Translation Review Process

For each language, the three different genAI forward translations were given to a linguistic expert in that language to review. Expertise was defined based on post-graduate training,

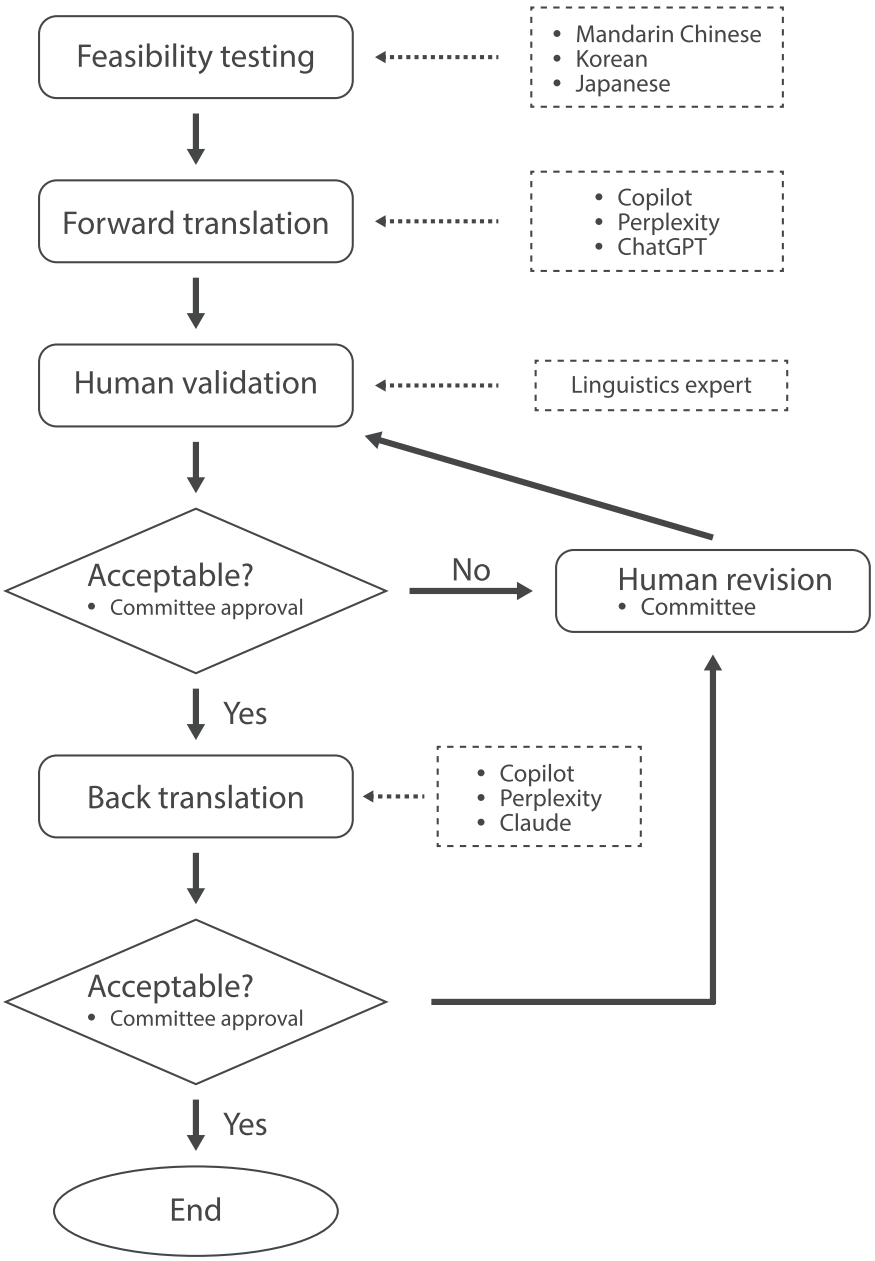

FIGURE 1 | Flow chart for translation process.

experience with translation, and scholarly work in the field. Our final, trusted pool of human experts comprised known authorities and established researchers in theoretical and applied linguistics worldwide. Each linguistic expert was provided with background on the construct of disease control, CRS, and the CRCT. Each linguist was also provided with specific instructions that the target audience for this instrument was the layperson, who would not be expected to have any advanced knowledge of medicine and that the instrument should be understandable to individuals with at least a sixth to eighth grade reading level.

The linguists were asked to assess each of the genAI forward translations using a rubric developed specifically for this study (Figure S1) that scored each of the five domains of conceptual and content equivalence, expressive and semantic accuracy, linguistic accuracy/appropriateness, experiential and cultural equivalence, and measurement equivalence on a continuous scale of 0 – 4 (with anchors of 0 = *inequivalent/unacceptable*, 1 = *major glaring issues*, 2 = *minor issues*, 3 = *mostly acceptable with minor modification*, 4 = *effectively equivalent/acceptable*), with a total possible score—which we refer to as the translation validity score (TVS)—of 0 to 20. The face and content validity of this rubric was established by expert linguists during feasibility testing and also confirmed prior to formal translations. All linguists were asked to offer suggestions for revisions to genAI forward translations to address any TVS domain that was deemed to be less than acceptable or equivalent to (i.e., domain score < 4) the original English version of the CRCT. The translation evaluation committee consolidated the comments and edits to the three genAI forward translations proposed by the linguist into one final version of the translated instrument that had unanimous support from the committee. Through intensive discussions among the committee, the final

TABLE 1 | Languages into which the Chronic Rhinosinusitis Control Test was translated and cross-culturally adapted.

Language
Arabic
Bengali
Brazilian Portuguese
Bulgarian
Cantonese Chinese
Czech
Danish
Dutch
European Portuguese
Filipino
Finnish
French
German
Greek
Hebrew
Hindi
Hungarian
Indonesian
Italian
Japanese
Korean
Mandarin Chinese
Norwegian
Persian
Polish
Punjabi
Romanian
Russian
Serbo-Croatian
Spanish
Swahili
Swedish
Thai
Turkish
Ukrainian
Urdu
Vietnamese

draft was refined to maintain linguistic naturalness and fidelity to the original CRCT. In the structured process of the refinement, multiple genAI engines, including ChatGPT, Copilot, Perplexity AI, Gemini, and Claude, and additional genAI-powered translation tools, such as Google Translate and DeepL Translate, were used. If any questions arose in the consolidation phase that could not be resolved by the translation evaluation committee, a second linguist was recruited to provide additional input. Once a unanimous consensus was achieved around a final version of the forward translation, a final TVS was completed by the linguist(s).

2.4 | Generative Artificial Intelligence-Supported Backward Translation

Each final, forward-translated version of the CRCT was backward-translated into English by entering it into three LLMs (Claude [22], Copilot [20], and Perplexity [21]) with the prompt "Please translate this [specific language] text into English." We used one new LLM (i.e., Claude instead of ChatGPT) for back-translation to examine whether shared LLM tool bias would be observed. In other words, given that using the same LLMs for forward and backward steps might not reveal potential discrepancies that might be overlooked otherwise, we intentionally used a new LLM for back-translation to ensure that the quality check was as independent and unbiased as possible. The backward translations were qualitatively reviewed by the translation evaluation committee for approval. Additionally, each backward translation was quantitatively evaluated and compared to the original English version of the CRCT for the number of words, reading level, and overall match. The reading level of the backward translations was assessed using the Flesch–Kincaid (FK) Grade Level and Flesch Reading Ease (FRE) score [25]. The overall match of each backward translation with the original English version of the CRCT was measured using the Bilingual Evaluation Understudy (BLEU) method [26], implemented in Python. BLEU uses a probabilistic n-gram (a contiguous sequence of n items, such as letters, words, or tokens, from a text) model, estimating the likelihood of a word based on the previous $n-1$ words, up to 4-grams (1-gram, 2-gram, 3-gram, 4-gram) with equal or predefined weights [26]. Because word-level semantic retention (1-gram) and phrase-level consistency (2-gram) were of concern, we used BLEU scores 1 and 2 (reflecting 1-gram and 2-gram match) for this study. As previously described [27], BLEU scores 0.3–0.4 reflected understandable-to-good translations, 0.4–0.5 reflected high-quality translations, 0.5–0.6 reflected very high-quality, adequate, and fluent translations, while scores greater than 0.6 were interpreted as better than human quality translations.

2.5 | Statistical Analysis

All analyses were performed using the statistical software package R version 4.4.0 (www.r-project.org) [28]. Standard descriptive statistics were performed. Bivariate comparisons were performed using the paired Wilcoxon rank sum test. Comparison of TVS (total and domain scores) distributions across the three LLMs was performed using the Kruskal–Wallis rank sum test, with post hoc comparison testing using Dunn's test with Bonferroni p-value adjustment.

3 | Results

3.1 | Forward Translation of the Chronic Rhinosinusitis Control Test by Generative AI

Using three LLMs (ChatGPT, Copilot, and Perplexity), the CRCT first underwent forward translation into the 37 target languages. The TVS for each genAI forward translation is shown in Table 2. The mean TVS and TVS domain scores, stratified by LLM, for the genAI forward translations are shown in Table 3, while the distribution of mean TVS scores for the genAI forward translations is shown in Figure 2. The languages having the lowest mean TVS score were Czech, Hindi, Norwegian, Romanian, Swahili, and Ukrainian (mean TVS < 14 for all), while languages having the highest mean TVS score for genAI forward translations were Arabic, Bengali, Bulgarian, Persian, Japanese, Korean, Mandarin, and Urdu (mean TVS > 18 for all).

The distributions of TVS for forward translations by each of the three different LLMs used are shown in Figure 3. The mean TVS for forward translations generated by ChatGPT was 15.7 (SD: 3.1), by Perplexity was 16.1 (SD: 3.1), and by Copilot was 16.3 (SD: 2.7). Although the distribution of TVS skewed lower for ChatGPT and Perplexity compared to that of Copilot, there was no statistically significant difference (p = 0.635) in TVS across the LLMs. Similarly, there were no statistically significant differences in TVS domain scores between LLMs ($p > 0.400$ for all TVS domain scores). In other words, all LLMs (ChatGPT, Copilot, and Perplexity) produced forward translations of quantitatively comparable quality and validity.

However, we did find significant differences when comparing TVS domain scores within LLMs ($p = 0.016$ for ChatGPT, $p = 0.002$ for Perplexity, and $p = 0.001$ for Copilot). Post hoc comparisons indicated that the experiential and cultural equivalence domain scores were significantly higher than the linguistic accuracy/appropriateness domain scores for all LLMs ($p = 0.020$ for ChatGPT, $p = 0.002$ for Perplexity, and $p = 0.018$ for Copilot). The experiential and cultural equivalence domain scores were also higher than the expressive and semantic accuracy domain scores for Perplexity ($p = 0.043$) and Copilot ($p = 0.026$).

3.2 | Translation and Cross-Cultural Adaptation of the Chronic Rhinosinusitis Control Test

The final translation and cross-cultural adaptation of the CRCT into the 37 languages was completed after input and modification of the genAI forward translations by human linguistics experts in those respective languages; these final translated versions of the CRCT are all provided in the Appendix in the Supporting Information. Common qualitative cultural considerations and adaptations that were discussed and implemented were related to the term "post-nasal drip" not existing in a given language, implications of the term "steroid" as potentially reflective of anabolic steroids rather than corticosteroids, as well as which corticosteroid medication may be most commonly used in a culture and therefore most appropriate to refer to as an example (e.g., the English version of the CRCT refers to prednisone as t example of an oral corticosteroid). Language-specific considerations and adaptations were also discussed and implemented.

The mean TVS of the final CRCT translations was 19.8 (out of a maximum of 20) and 32 translations (out of 37) achieved a maximum TVS of 20. For the final translations that did not reach the maximum TVS of 20 (four [Brazilian Portuguese, Italian, Norwegian, and Vietnamese] had final TVS of 19 and one [Hindi] had final TVS of 18), only minor issues were identified—such as wording issues and nuances lost in the translation—that were ultimately deemed necessary and tolerated in balancing linguistic conventions with fidelity to the original CRCT. In no cases were these minor issues deemed by the linguists to impact the comprehensibility of the translated instrument. As expected, the TVS for the final translations was significantly greater than the TVS for the genAI forward translations by ChatGPT, Perplexity, and Copilot ($p < 0.001$ for all cases).

3.3 | Backward Translation of Translated/Adapted Versions of the Chronic Rhinosinusitis Control Test

Backward translations by Copilot, Perplexity, and Claude LLMs of all final, translated, cross-culturally adapted versions of the CRCT were qualitatively reviewed and unanimously approved as acceptable by their respective translation evaluation committee. The performance of Claude (the LLM newly adopted for backward translation instead of ChatGPT) showed no discrepancies with that of the other two LLMs (Copilot and Perplexity), which provided reassurance for translation independence and quality and indicated no potential shared-tool bias. The quantitative analyses of backward translations are shown in Table 4. In comparison to the original English version of the CRCT, which had 169 total words, an FK grade level of 5.9, and an FRE score of 78.3, the English backward translations of the translated, cross-culturally adapted versions of the CRCT had closely matching quantitative characteristics. Additionally, the BLEU-1 scores were almost universally reflective of high-quality translations or better, with one exception of the BLEU-1 score for the Perplexity backward translation of the Hindi CRCT that had BLEU-1 score of 0.39. BLEU-2 scores were almost entirely reflective of high-quality translations or better, with only Cantonese, Czech, and Thai Copilot backward translations, Cantonese, Hindi, and Japanese Perplexity backward translations, and German, Hebrew, Urdu, and Vietnamese Claude backward translations having BLEU-2 scores in the range of 0.35–0.40, indicating good translation quality. All languages had at least one backward translation that had a BLEU-2 score of high-quality or better.

4 | Discussion

The importance of quantifying CRS control and classifying uncontrolled disease has only increased with time [5, 29–31]. With its international, multidisciplinary, and consensus-derived origins, the CRCT—a psychometrically validated PROM that quantifies and classifies CRS control—is a globally useful instrument, both for patient care and for research [7]. However, before the CRCT, which was developed and validated in the English language, can be psychometrically validated for use in other languages, it must undergo translation and cross-cultural adaptation in those languages [32]. In order to remove this resource-intensive barrier [14, 33], the objective of our study was to create a library of translated and cross-culturally adapted versions of the CRCT that

TABLE 2 | Translation validity scores for CRCT forward translations made by large language models.

	Translation validity score[a]				
Language	**ChatGPT**	**Perplexity**	**Copilot**	**Mean[b]**	**Final translation[c]**
Arabic	19	20	19	19.3	20
Bengali	19	18	18	18.3	20
Brazilian Portuguese	18	13	17	16	19
Bulgarian	19	18	20	19	20
Cantonese Chinese	13	14.5	15.5	14.3	20
Czech	14	13	12	13	20
Danish	16	10	17	14.3	20
Dutch	17	17	17	17	20
European Portuguese	13	15	17	15	20
Filipino	17	12	18	15.7	20
Finnish	14	16	17	15.7	20
French	15	15	15	15	20
German	16	19	19	18	20
Greek	11	17	18	15.3	20
Hebrew	15	16	16	15.7	20
Hindi	11	11	11	11	18
Hungarian	17	20	16	17.7	20
Indonesian	18	15	14	15.7	20
Italian	13	15	15	14.3	19
Japanese	20	19.5	19.5	19.7	20
Korean	18.5	18.5	19	18.7	20
Mandarin Chinese	17	18.5	19	18.2	20
Norwegian	9	14	9	10.7	19
Persian	17	19.5	18.5	18.3	20
Polish	17	19	15	17	20
Punjabi	17	14	16	15.7	20
Romanian	6	15	11	10.7	20
Russian	18	17	18	17.7	20
Serbo-Croatian	14	17	18	16.3	20
Spanish	17	17	17	17	20
Swahili	16	6	13	11.7	20
Swedish	16	19	18	17.7	20
Thai	18	18	18	18	20
Turkish	18	17	16	17	20
Ukrainian	11	15	13	13	20
Urdu	19.5	19.8	19.5	19.6	20
Vietnamese	15.5	18.5	15.5	16.5	19

[a]Maximum score of 20.

[b]Mean of TVS from forward translations created by ChatGPT, Perplexity, and Copilot.

[c]Final forward translation produced by synthesizing forward translations by the three large language models and comments/input from human linguists.

TABLE 3 | Translation Validity total and domain scores[a] for CRCT forward translations made by large language models.

	ChatGPT	Perplexity	Copilot	Final translation[b]
Domain scores (range 0–4)				
Conceptual and content equivalence	3.2 (0.8)	3.3 (0.8)	3.4 (0.7)	4.0 (0.0)
Expressive and semantic accuracy	3.0 (0.8)	3.1 (0.7)	3.0 (0.7)	4.0 (0.2)
Linguistic accuracy	2.8 (0.8)	2.9 (0.8)	3.0 (0.7)	3.9 (0.3)
Experiential and cultural equivalence	3.4 (0.7)	3.5 (0.7)	3.5 (0.7)	4.0 (0.1)
Measurement equivalence	3.2 (0.9)	3.3 (0.8)	3.4 (0.7)	3.9 (0.2)
Total score (range 0–20)	15.7 (3.1)	16.1 (3.1)	16.3 (2.7)	19.8 (0.4)

[a]Mean (standard deviation) shown.

[b]Final forward translation produced by synthesizing forward translations by the three large language models and comments/input from human linguists.

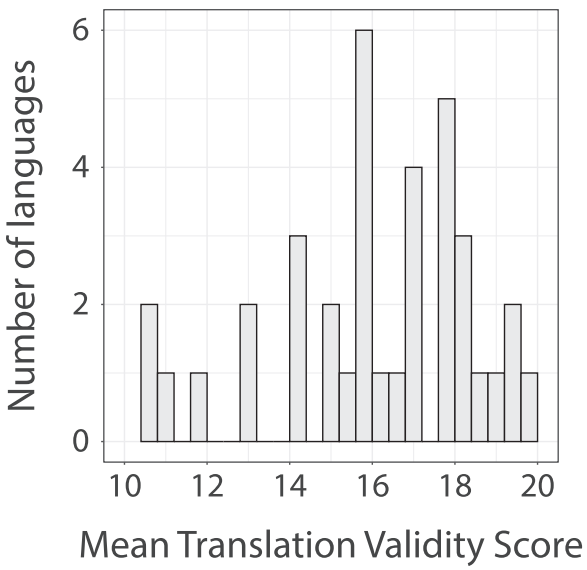

FIGURE 2 | Histogram plot of mean translation validity scores for forward translations of the CRCT created using ChatGPT, Perplexity, and Copilot.

could be used as a starting point for its psychometric validation worldwide. Using genAI-supported methodology in collaboration with expert linguists providing oversight and quality control for each language, we report and provide the translation and cross-cultural adaptation of the CRCT into 37 languages that are spoken across every populated continent.

Our results indicated that no LLM (between ChatGPT, Copilot, or Perplexity) produced significantly higher quality translation than the others, although CRCT translations produced by LLMs alone were not accurate enough for use. While LLMs produced mostly accurate translations, the input of human linguists was necessary to produce maximally accurate, valid translations. Adaptations of the translated CRCT had to balance linguistic validity—including linguistic naturalness—with fidelity to the original CRCT. Although 32 out of 37 CRCT translations reached a maximum TVS, 5 translations (Brazilian Portuguese, Italian, Norwegian, Vietnamese, and Hindi) ultimately retained some possibly mildly uncommon language in the interests of fidelity to the original CRCT, leading to acceptable but not maximal TVS.

In creating and reporting a library of translated, cross-culturally adapted CRCTs, we believe that our results will serve as an important resource for validation and utilization of the CRCT globally.

Our study also makes an important, novel contribution by demonstrating the power of genAI to assist high-quality translation of PROMs. Reporting the largest-scale translation and cross-cultural adaptation of a single PROM using a genAI-supported framework, our work has important implications for PROM accessibility and data equity that extend beyond the CRCT. Data equity is recognized as a major problem by the World Health Organization [34], as clinical trials have historically been disproportionately led by and included participants from high-income countries. The impact of a PROM extends only as far as the individuals who are proficient in its language [35–39], and the present-day deficiency of data equity in clinical trials is driven in part by the accessibility of key outcome measures like PROMs, which are disproportionately developed and translated in English or languages of wealthy, more populous countries [35, 40–42]. With the recognized need for multinational studies that are more globally representative of patients and the implicit constraints placed on PROMs based on language, the critical importance of wide-ranging translation and cross-cultural adaptation of PROMs is clear [43]. However, translation and cross-cultural adaptation demand time, manpower, and financial support [14, 33], all of which present real-world challenges that can be insurmountable barriers to widespread availability of PROMs. Recently, exciting advances in genAI have identified LLMs as an emerging resource that may be used for tasks ranging from creating PROMs [44] to reducing the barriers for translation and cross-cultural adaptation [15, 45].

LLMs have already shown promise in accurately translating clinical patient-facing materials, such as patient instructions or clinical education materials, from English to other languages [46–49]. LLMs have also been demonstrated to accurately translate PROMs [50]. In a study by Lu et al., LLMs were used for forward translation (from English into Arabic, Vietnamese, Italian, Hungarian, Malay, and Dutch) and backward translation of two widely used PROMs for outcomes of reconstructive surgery in comparison to human translators [50]. In comparison to human translators, LLMs generally created good-quality or better

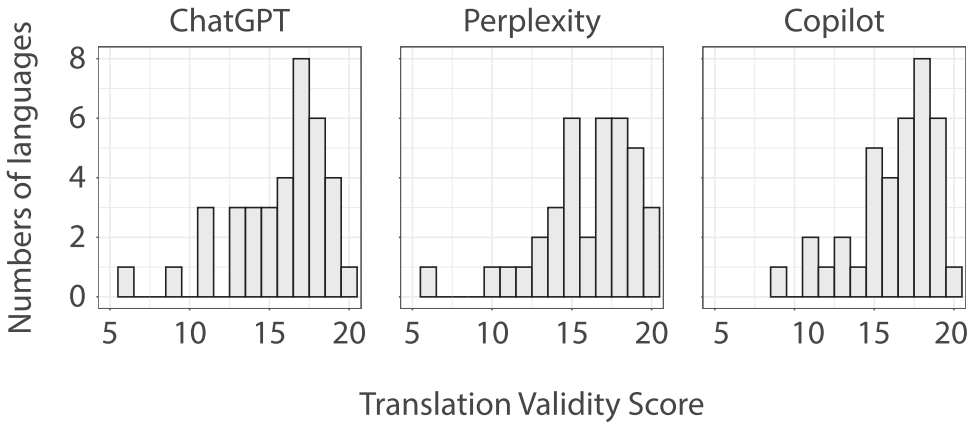

FIGURE 3 | Histogram plots of translation validity scores (TVS) for forward translations of the CRCT created using ChatGPT, Perplexity, and Copilot.

TABLE 4 | Quantitative analysis of backward translations into English of translated/adapted versions of the CRCT.

	Copilot	Perplexity	Claude
Number of words[a]	Mean: 177.8	Mean: 179.1	Mean: 182.9
	SD: 8.5	SD: 10.9	SD: 9.0
	Median: 178	Median: 176	Median: 183
	Range: 164–198	Range: 161–206	Range: 164–199
Flesch–Kincaid Grade Level[b]	Mean: 7.0	Mean: 7.9	Mean: 7.3
	SD: 1.5	SD: 1.3	SD: 0.9
	Median: 6.7	Median: 7.5	Median: 7.2
	Range: 3.7–11.0	Range: 4.8–10.1	Range: 5.3–9.9
Flesch Reading Ease score[c]	Mean: 70.5	Mean: 65.6	Mean: 68.2
	SD: 7.4	SD: 7.0	SD: 5.4
	Median: 71.7	Median: 67.5	Median: 68.3
	Range: 51.1–87.9	Range: 50.6–82.4	Range: 49.6–77.4
BLEU-1 score	Mean: 0.69	Mean: 0.68	Mean: 0.66
	SD: 0.07	SD: 0.09	SD: 0.07
	Median: 0.70	Median: 0.69	Median: 0.67
	Range: 0.55–0.84	Range: 0.39–0.81	Range: 0.46–0.77
BLEU-2 score	Mean: 0.49	Mean: 0.48	Mean: 0.47
	SD: 0.07	SD: 0.06	SD: 0.06
	Median: 0.49	Median: 0.48	Median: 0.47
	Range: 0.38–0.63	Range: 0.36–0.61	Range: 0.35–0.57

[a]The original English version of the CRCT has 169 words.
[b]The original English version of the CRCT has a Flesch-Kincaid Grade Level of 5.9.
[c]The original English version of the CRCT has a Flesch Reading Ease score of 78.3.

translations, with excellent accuracy for rote translation, although human translators were found to be especially important for making sociocultural adaptations that LLMs were unable to make [50], leading to the conclusion that using LLMs with human quality control may lead to the best translation results. In fact, in a randomized double-blind non-inferiority study comparing translation methodologies for a PROM, the combined use of LLMs with quality control by a human translator led to equivalent translation quality to that produced by human translators [24]. Our results support these prior findings as virtually none of the genAI forward translations of the CRCT was found to be acceptable on their own, while the addition of input from human linguists led to appropriate and acceptable translational validity.

Our study should be interpreted within the constraints of its limitations. Although LLMs have been used with human translators for quality control to translate and adapt text [24, 50], this is still an area of active investigation, as LLMs are known

to have limitations in translation and cross-cultural adaptation of text. For example, in many of our translations, the term "post-nasal drip" was translated using descriptive terminology while human linguists pointed out that the terminology itself was not used in their respective languages. Although LLMs are not human, they may still have the same cultural, racial, and gender biases implicit in that human-generated text on which they were trained. Previous studies have found that LLMs may show bias in cultural values toward higher-income, English-speaking, and Protestant European countries [51, 52]. LLMs can also take on cultural biases based on the language that is used to interact with them [51, 53]. As a result, the accuracy of LLMs for translation may be dependent on the target language [50, 54], sometimes struggling with low-resource languages (that have limited digital data that can be used to train LLMs) [55]. In this regard, our results were somewhat mixed, with low-resource and high-resource languages among both the lowest and the highest quality forward translations. The limitations of LLMs reflect the importance of continued oversight and intimate involvement of human experts in the process of translation and cross-cultural adaptation, which was supported by our results. Additionally, we also acknowledge that the translated and cross-culturally adapted versions of the CRCT that we provide should be used as a starting point, and prior to psychometric testing, pilot testing for understandability with real-world patients is needed [33]. Finally, it is of importance to reiterate that the translated and cross-culturally adapted versions of the CRCT that we provide must undergo psychometric validation in each language before they can be used in that respective language.

5 | Conclusion

The library of translated, cross-culturally adapted versions of the CRCT into 37 languages reported here may serve as the foundation for psychometric validation, and ultimately for use by patients, worldwide. Moreover, LLMs appear to function comparably across platforms and consistently across different languages as important resources for the translation of PROMs. However, accompanying input and oversight by human linguists are still needed to produce acceptable and linguistically valid PROM translations/adaptations.

Affiliations

[1]School of Education, University of Cincinnati, Cincinnati, Ohio, USA | [2]Department of Applied Linguistics, Xi'an Jiaotong-Liverpool University, Suzhou, China | [3]College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China | [4]Independent Researcher, Gyeongsan, South Korea | [5]Department of Artificial Intelligence, University of Science and Technology, Daejeon, South Korea | [6]Department of Asian and Asian American Studies, Stony Brook University, Stony Brook, New York, USA | [7]Persian Studies, School of Languages, Literatures, and Cultures, University of Maryland, College Park, Maryland, USA | [8]Department of Language and Culture Studies, University of Cincinnati Blue Ash College, Cincinnati, Ohio, USA | [9]Department of Cognitive Science, Budapest University of Technology and Economics, Budapest, Hungary | [10]Department of Psychology and Speech-Language Pathology, University of Turku, Turku, Finland | [11]Institute of Applied Linguistics, University of Warsaw, Warsaw, Poland | [12]Language Centre, Hong Kong Baptist University, Hong Kong, China | [13]ERI-Lectura and Department of Developmental and Educational Psychology, University of Valencia, Valencia, Spain | [14]Russian Language Program, Brandeis University, Waltham, Massachusetts, USA | [15]Psychology Program, Middle East Technical University, Güzelyurt, Cyprus | [16]Institute of Languages and Linguistics, University of the Punjab, Lahore, Pakistan | [17]School of Natural Sciences, Rice University, Houston, Texas, USA | [18]The Education University of Hong Kong, Hong Kong, China | [19]Department of Special Education, National Taiwan Normal University, Taipei, Taiwan | [20]Department of Educational Psychology, Texas A&M University, College Station, Texas, USA | [21]Department of Languages, University of Helsinki, Helsinki, Finland | [22]Research and Enterprise Directorate, University of Wolverhampton, Wolverhampton, UK | [23]Department of Linguistics, The Ohio State University, Columbus, Ohio, USA | [24]Department of Psychology, Faculty of Education, University of South Bohemia, České Budějovice, Czech Republic | [25]Center For Linguistics, School of Arts and Humanities, University of Lisbon, Lisbon, Portugal | [26]Master's Program in English Language Education, Universitas Kristen Satya Wacana, Salatiga, Indonesia | [27]Department of Learning Disabilities, Faculty of Education, University of Haifa, Haifa, Israel | [28]Department of Slavic Studies, University of Graz, Graz, Austria | [29]Faculty of Education, Mount Saint Vincent University, Halifax, Nova Scotia, Canada | [30]Department of African, African American, and Diaspora Studies, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA | [31]Department of Foreign Languages, Amirkabir University of Technology, Tehran, Iran | [32]Iorgu Iordan – Alexandru Rosetti Institute of Linguistics, Romanian Academy, Bucharest, Romania | [33]College of Education, University of the Philippines Diliman, Quezon City, Philippines | [34]Department of Psychology and Center for Applied Neuroscience, University of Cyprus, Nicosia, Cyprus | [35]Department of Methodology, Faculty of Psychology, University of Valencia, Valencia, Spain | [36]Department of Architecture and Arts, Iuav University of Venice, Venice, Italy | [37]School of Languages and Tourism, Hanoi University of Industry, Hanoi, Vietnam | [38]Faculty of Linguistics, Philology and Phonetics, University of Oxford, Oxford, UK | [39]School of Humanities, Arts and Social Sciences, Stevens Institute of Technology, Hoboken, New Jersey, USA | [40]Department of Learning Disabilities, Faculty of Education, Edmond J. Safra Brain Research Center for the Study of Learning Disabilities, University of Haifa, Haifa, Israel | [41]Department of Portuguese, University of Massachusetts, Dartmouth, Massachusetts, USA | [42]Department of Spanish and Portuguese, University of Kansas, Lawrence, Kansas, USA | [43]Department of Information Technology, Åland University of Applied Sciences, Mariehamn, Finland | [44]Thai Language Programme, Center for Language Studies, National University of Singapore, Singapore | [45]Graduate School of Humanities, Nagoya University, Nagoya, Japan | [46]Linguistics Program, University of South Carolina, Columbia, South Carolina, USA | [47]Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway | [48]Department of Otolaryngology—Head and Neck Surgery, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

Acknowledgments

We extend our thanks to Dimitrios Meletis, who offered insights and constructive feedback to this project.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. A. R. Sedaghat and K. M. Phillips, "Defining 'Control' of Chronic Rhinosinusitis," *Current Opinion in Otolaryngology & Head and Neck Surgery* 31 (2023): 17–23.

2. A. R. Sedaghat and K. M. Phillips, "Chronic Rhinosinusitis Disease Control: A Review of the History and the Evidence," *Expert Review of Clinical Immunology* 19 (2023): 903–910.

3. A. Sedaghat and C. Hopkins, "Chronic Rhinosinusitis Disease Control as a Metric for Guiding Treatment," *Rhinology* 58 (2020): 193.

4. C. Hopkins, R. Hettige, A. Soni-Jaiswal, et al., "CHronic Rhinosinusitis Outcome MEasures (CHROME), Developing a Core Outcome Set for Trials of Interventions in Chronic Rhinosinusitis," *Rhinology* 56 (2018): 22–32.

5. A. Ali, D. R. Fakunle, V. Yu, et al., "Heterogeneity in the Definition of Chronic Rhinosinusitis Disease Control: A Systematic Review of the Scientific Literature," *European Archives of Oto-Rhino-Laryngology* 280 (2023): 5345–5352.

6. A. R. Sedaghat, W. J. Fokkens, V. J. Lund, et al., "Consensus Criteria for Chronic Rhinosinusitis Disease Control: An International Delphi Study," *Rhinology* 61 (2023): 519–530.

7. R. A. Cotter, C. W. Lee, K. Wilson, et al., "Development and Psychometric Validation of the Chronic Rhinosinusitis Control Test," *Rhinology* 64 (2026): 38–50.

8. R. A. Cotter, F. A. Houssein, R. K. Reinert, K. M. Philips, and A. R. Sedaghat, "Patient Perspectives on International Multidisciplinary Consensus Criteria for Chronic Rhinosinusitis Disease Control," *Laryngoscope Investigative Otolaryngology* 9 (2024): e70005.

9. Y. Wang, A. E. Boyd, L. Rountree, et al., "Ten Core Concepts for Ensuring Data Equity in Public Health," *JAMA Health Forum* 7 (2026): e256031.

10. A. Chong, J. Claydon, S. Chhina, M. Sadarangani, and Q. Doan, "Strategies to Reduce Language Barriers in Clinical Research: A National Survey of Pediatric Health Researchers," *Ethics, Medicine and Public Health* 33 (2025): 101122.

11. J. M. Kahn, D. M. Gray, J. M. Oliveri, C. M. Washington, C. R. DeGraffinreid, and E. D. Paskett, "Strategies to Improve Diversity, Equity, and Inclusion in Clinical Trials," *Cancer* 128 (2022): 216–221.

12. V. Nittas, P. Daniore, S. J. Chavez, and T. B. Wray, "Challenges in Implementing Cultural Adaptations of Digital Health Interventions," *Communications Medicine* 4 (2024): 7.

13. A. D. Sperber, "Translation and Validation of Study Instruments for Cross-Cultural Research," *Gastroenterology* 126 (2004): S124–128.

14. D. E. Beaton, C. Bombardier, F. Guillemin, and M. B. Ferraz, "Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures," *Spine* 25 (2000): 3186–3191.

15. J. R. Kunst and K. Bierwiaczonek, "Utilizing AI Questionnaire Translations in Cross-Cultural and Intercultural Research: Insights and Recommendations," *International Journal of Intercultural Relations* 97 (2023): 101888.

16. "List of Languages by Number of Native Speakers," Wikipedia, accessed November 23, 2025, https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers.

17. "What Are the Top 200 Most Spoken Languages?" Ethnologue, accessed October 1, 2025, https://www.ethnologue.com/insights/ethnologue200/.

18. H. K. Pae, ed. "Chinese, Japanese, and Korean writing system: All East-Asian but different scripts," in *Script Effects as the Hidden Drive of the Mind, Cognition, and Culture* (Springer, 2020), 71–105.

19. "ChatGPT 5.1 [Large Language Model]," OpenAI, accessed December 8, 2025. https://chat.openai.com/chat.

20. "Copilot (GPT-4) [Large Language Model]," Microsoft, accessed February 21, 2026. https://copilot.microsoft.com/.

21. "Perplexity AI [Large Language Model]," Perplexity AI, accessed February 21, 2026. https://www.perplexity.ai/.

22. "Claude (Version 4.5 Sonnet)," Anthropic, accessed February 21, 2026. https://claude.ai/.

23. J. M. Kowal, K. H. Bryant, D. Segall, and T. Kantrowitz, "Harnessing Generative AI for Assessment Item Development: Comparing AI-Generated and Human-Authored Items," *International Journal of Selection and Assessment* 33 (2025): e70021.

24. C. B. Sorensen, A. Gram-Hanssen, J. Rosenberg, and J. J. Baker, "Comparing ChatGPT-4 and Human Translation of an Outcome Questionnaire: A Randomized, Double-Blinded Non-Inferiority Study," *Cureus* 17 (2025): e82525.

25. J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Chief of Naval Technical Training, Naval Air Station Memphis*, (1975): 8–75.

26. K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, (2002): 311–318.

27. C. L. Chen, Y. Dong, C. Castillo-Zambrano, et al., "A Systematic Multimodal Assessment of AI Machine Translation Tools for Enhancing Access to Critical Care Education Internationally," *BMC Medical Education* 25 (2025): 1022.

28. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2011).

29. W. J. Fokkens, A. S. Viskens, V. Backer, et al., "EPOS/EUFOREA Update on Indication and Evaluation of Biologics in Chronic Rhinosinusitis With Nasal Polyps 2023," *Rhinology* 61 (2023): 194–202.

30. W. J. Fokkens, V. J. Lund, C. Hopkins, et al., "European Position Paper on Rhinosinusitis and Nasal Polyps 2020," *Rhinology* 58 (2020): 1–464.

31. J. Chiu, J. Fastenberg, C. Tong, B. Navetta-Modrov, and S. Marcus, "Biologics for Chronic Rhinosinusitis with Nasal Polyposis: Current Landscape and Future Directions," *Laryngoscope Investigative Otolaryngology* 10 (2025): e70282.

32. P. Cruchinho, M. D. Lopez-Franco, M. L. Capelas, et al., "Translation, Cross-Cultural Adaptation, and Validation of Measurement Instruments: A Practical Guideline for Novice Researchers," *Journal of Multidisciplinary Healthcare* 17 (2024): 2701–2728.

33. D. Wild, A. Grove, M. Martin, et al., "Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation," *Value in Health* 8 (2005): 94–104.

34. *Guidance for Best Practices for Clinical Trials* (World Health Organization, 2024).

35. B. G. Allar, C. N. Eruchalu, S. Rahman, et al., "Lost in Translation: A Qualitative Analysis of Facilitators and Barriers to Collecting Patient Reported Outcome Measures for Surgical Patients With Limited English Proficiency," *American Journal of Surgery* 224 (2022): 514–521.

36. J. A. Konopka, D. A. Bloom, K. W. Lawrence, J. F. Oeding, R. Schwarzkopf, and C. M. Lajam, "Non-English Speakers and Socioeconomic Minorities Are Significantly Less Likely to Complete Patient-Reported Outcome Measures for Total Hip and Knee Arthroplasty: Analysis of 16,119 Cases," *Journal of Arthroplasty* 38 (2023): S69–S77.

37. J. R. M. Ngwayi, J. Tan, N. Liang, K. U. Obie, and D. E. Porter, "Exploring the Impact of Patient Reported Outcome Measures (PROMs) Among Orthopaedic Surgeons in Mainland China: Systematic Review and Survey-Based Study on Hip and Knee Instruments," *BMC Musculoskeletal Disorders* 22 (2021): 566.

38. L.-M. Alpers and I. Hanssen, "Ethnic Minority Patients' Experiences of Filling in Translated Patient-Reported Outcome Measures," *Journal of Public Health Research* 14 (2025): 22799036251390945.

39. J. Nikolovski, B. Kim, R. L. Morton, et al., "Strategies to Promote the Completion of Patient-Reported Outcome Measures by Culturally and Linguistically Diverse and Indigenous Peoples in Clinical Care Settings: A Systematic Review," *Quality of Life Research* 34 (2025): 1541–1551.

40. Y. Zhang, J. Ren, Y. Zang, W. Guo, A. Disantis, and R. L. Martin, "Cross-Culturally Adapted Versions of Patient Reported Outcome Measures for the Lower Extremity," *International Journal of Sports Physical Therapy* 18 (2023): 653–686.

41. J. M. Hirsh, "The Challenge and Opportunity of Capturing Patient Reported Measures of Rheumatoid Arthritis Disease Activity in Vulnerable Populations With Limited Health Literacy and Limited English Proficiency," *Rheumatic Diseases Clinics of North America* 42 (2016): 347–362.

42. V. E. Kwok, A. Diaz, E. Grellinger, and I. Swarup, "Validating Patient-Reported Outcomes and Surveys in Other Languages," *Current Review in Musculoskeletal Medicine* 19 (2025): 1.

43. I. Coskun Benlidayi and L. Gupta, "Translation and Cross-Cultural Adaptation: A Critical Step in Multi-National Survey Studies," *Journal of Korean Medical Science* 39 (2024): e336.

44. L. Boyer, S. Fernandes, P. Auquier, B. Falissard, and T. Panch, "Reimagining Patient-Reported Outcomes in the Age of Generative AI," *NPJ Digital Medicine* 8 (2025): 624.

45. H. Williams, "Harmonising Linguistic Validation With AI: Precision, Efficiency, and the Human Touch in Patient-Reported Outcome Translation," *Medical Writing* 33 (2024): 66–69.

46. M. Ray, D. J. Kats, J. Moorkens, et al., "Evaluating a Large Language Model in Translating Patient Instructions to Spanish Using a Standardized Framework," *JAMA Pediatrics* 179 (2025): 1026–1033.

47. A. AlSammarraie and M. Househ, "The Use of Large Language Models in Generating Patient Education Materials: A Scoping Review," *Acta Informatica Medica* 33 (2025): 4–10.

48. F. Dzuali, K. Seiger, R. Novoa, et al., "ChatGPT May Improve Access to Language-Concordant Care for Patients with Non-English Language Preferences," *JMIR Medical Education* 10 (2024): e51435.

49. S. Andalib, A. Spina, B. Picton, S. S. Solomon, J. A. Scolaro, and A. M. Nelson, "Using AI to Translate and Simplify Spanish Orthopedic Medical Text: Instrument Validation Study," *JMIR AI* 4 (2025): e70222.

50. S. C. Lu, C. Xu, M. Kaur, M. O. Edelen, A. Pusic, and C. Gibbons, "Can Machine Translation Match Human Expertise? Quantifying the Performance of Large Language Models in the Translation of Patient-Reported Outcome Measures (PROMs)," *Journal of Patient-Reported Outcomes* 9 (2025): 94.

51. Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural Bias and Cultural Alignment of Large Language Models," *PNAS Nexus* 3 (2024): 346.

52. L. Vargas-Parada, "Large Language Models Are Biased—local Initiatives Are Fighting for Change," *Nature* (2025), https://doi.org/10.1038/d41586-025-03891-y.

53. J. G. Lu, L. L. Song, and L. D. Zhang, "Cultural Tendencies in Generative AI," *Nature Human Behaviour* 9 (2025): 2360–2369.

54. M. Martos, B. Fields, S. G. Finlayson, et al., "Accuracy of Artificial Intelligence vs Professionally Translated Discharge Instructions," *JAMA Network Open* 8 (2025): e2532312.

55. N. Robinson, P. Ogayo, D. R. Mortensen, G. Neubig, and M. T. ChatGPT, *Competitive for High- (but Not Low-) Resource Languages* (Association for Computational Linguistics, 2023), 392–418.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting File 1: alr70194-sup-0001-Appendix.pdf. **Figure S1:** Rubric used by human linguists to determine translation validity score for forward translations of the CRCT.