Evaluating ChatGPT's Utility in Enhancing Readability of Online English and Spanish Orthopedic Patient Education Materials

Carrie Reaver[1], Daniel Pereira, Elisa Victoria Carrillo, Carolena Rojas Marcos[2], Charles A Goldfarb[3]
[1]Orthopedics, [2]Hospital For Special Surgery, [3]Washington University School of Medicine

INTRODUCTION: The readability of online patient educational materials (OPEMs) in orthopedic surgery has been shown to be above the AMA/NIH recommended reading level of sixth-grade or below, for both English and Spanish. The current project aims to evaluate ChatGPT's performance across English and Spanish orthopedic OPEMs when prompted to rewrite the material at the more accessible sixth-grade reading level.

METHODS:

This cross-sectional study evaluated the readability of the first 10 OPEMs queried online in both English and Spanish for six common orthopedic surgeries. We included OPEMs with more than 100 words, resulting in 57 and 56 for English/Spanish content, respectively. Five distinct, validated readability tests were used to score the OPEMs before and after ChatGPT 4.0 was prompted to rewrite the OPEMs at a sixth-grade reading level (Table 1). We compared the mean averages of each readability test, the cumulative average reading grade level, and the percent of complex words (defined as 3+ syllables) between original content and ChatGPT-rewritten content for both languages using paired t-tests (SPSS 2023).

RESULTS: The mean reading grade level of original English content and original Spanish content was 9.26 ± 1.65 and 9.72 ± 0.65, respectively. ChatGPT successfully rewrote materials at a significantly lower grade level in both English and Spanish, resulting in a mean reading grade level of 7.40 ± 1.73 ($p<0.05$) for English content and 8.525 ± 0.718 ($p<0.05$) for Spanish content (Table 1, Figure 1). An average of approximately 16% and 14% of original English and Spanish content, respectively, consisted of complex words. This proportion rose to about 35% ($p<0.05$) and 34% ($p<0.05$) for ChatGPT-rewritten English and Spanish content, respectively (Figure 2).

DISCUSSION AND CONCLUSION:

Our study demonstrates that ChatGPT can rewrite materials at a more accessible reading grade level in both English and Spanish, although it was able to achieve a lower average reading grade level in English versus Spanish. Additionally, ChatGPT-rewritten materials fail to meet recommended readability, particularly regarding the use of complex words. Further investigation should improve strategies to overcome these limitations in ChatGPT readability editing, particularly as we increasingly rely on artificial intelligence services for addressing health literacy disparities across languages.

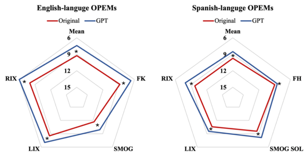

Figure 1: Radar plots depicting the mean readability scores of four individual readability tests for original OPEMs (red) and ChatGPT-rewritten OPEMs (blue) in English (left) and Spanish (right). Asterisks represent significance ($p<0.05$) between original versus ChatGPT readability scores.

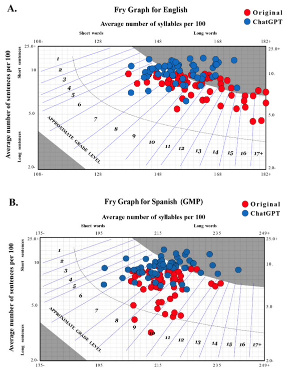

Figure 2. Fry Graph for English content (**A**) and Gilliam Peña Mountain Fry Graph for Spanish content (**B**) demonstrating the reading grade level of OPEMs before (red) and after (blue) ChatGPT intervention. The upper left gray region indicates failed tests as the content contained too many complex words to be classified within a reading grade level.

Table 1. Average reading grade level by readability test for original and ChatGPT-rewritten OPEMs in English and Spanish (Oleander Readability Studio, 2020). Fry and Gilliam Peña Mountain Fry scores were not included in mean readability calculation as several tests failed from the high presence of complex words (Figure 2).

	N	FK / FH	SMOG / SMOG SOL	LIX	RIX	FRY / GMP	Mean Grade Level	p-value
Original English	57	8.78 ± 2.19	11.68 ± 1.70	8.54 ± 1.74	8.02 ± 1.56	failed	9.26 ± 1.65	<0.05
ChatGPT English	57	6.66 ± 1.24	9.91 ± 1.01	7.04 ± 1.09	6.00 ±0.87	failed	7.40 ± 1.73	
Original Spanish	56	61.86 ± 5.71	9.58 ±1.20	10.57 ±1.29	9.73 ±1.42	failed	9.72 ± 0.65	<0.05
ChatGPT Spanish	56	65.43 ±5.11	8.15 ± 0.82	9.54 ±1.09	7.91 ±1.05	failed	8.53 ± 0.72	