Enhancing the Readability of Sports Medicine Patient-Reported Outcome Measures Using ChatGPT 4.0: A Quality Assessment of 26 Common Instruments

Harjot Uppal, Nikhil Sahai, Kumar Gautam Sinha, Ki S Hwang, Andrew McGinniss, Arash Emami, Anthony James Scillia

INTRODUCTION:

Patient-reported outcome measures (PROMs) are essential in sports medicine, offering insight into pain, function, and recovery. However, many PROMs exceed the average U.S. adult reading level, potentially limiting their utility—particularly among underserved populations. With only 12% of Americans demonstrating proficient health literacy, simplifying PROMs is a critical step toward improving patient comprehension and data accuracy.

Large language models like ChatGPT 4.0 have shown promise in generating health information at a sixth-grade reading level while preserving meaning. While this has been demonstrated in educational materials, its impact on commonly used sports medicine PROMs remains unknown. This study evaluated whether ChatGPT 4.0 could improve PROM readability while preserving interpretability, structure, and clinical value.

METHODS:

A retrospective analysis of 26 sports medicine PROMs—identified through citation frequency in PubMed—was conducted. Instruments included the IKDC, KOOS, FAAM, and DASH. Each PROM was digitized and submitted to ChatGPT 4.0 using the standardized prompt:

"Please revise this outcome measure so it is written at or below a 6th-grade reading level. Simplify medical terms, shorten sentences, and enhance clarity for patients with limited health literacy. Do not change the original intent or structure."

Revised PROMs were assessed using 12 validated readability metrics via Readable.com. A board-certified sports medicine physician independently reviewed each output for content fidelity. Errors were categorized as (1) response scale alterations, (2) timeframe distortions, or (3) semantic changes to item intent. Readability scores were compared using the Wilcoxon signed-rank test ($\alpha = 0.05$).

RESULTS:

ChatGPT 4.0 significantly improved PROM readability across all metrics ($p < 0.001$), reducing median grade level from 9.1 to 5.7. Sentence length decreased by 21%, polysyllabic word usage declined by 27%, and word count was reduced by 15%. Readability improvements were most pronounced in Flesch-Kincaid, SMOG, and Gunning Fog scores.

However, 15 of 26 PROMs (57.7%) contained at least one clinically significant error. The most frequent issues were changes to validated response scales (27%), omission of time qualifiers (19%), and alterations to item intent (11%). These modifications may compromise psychometric validity and clinical applicability.

DISCUSSION AND CONCLUSION:

ChatGPT 4.0 substantially enhances the readability of PROMs in sports medicine, aligning many with NIH/AMA guidelines for sixth-grade comprehension. While promising as a first-pass tool for simplifying health communication, over half of the revised PROMs introduced content-altering errors. Expert review remains essential before clinical adoption.

Future studies should refine prompt strategies, develop automated validation tools, and assess how AI-revised PROMs perform in real-world patient populations in terms of comprehension, completion, and outcome reliability.