Improving Readability of Foot and Ankle Patient-Reported Outcome Measures Using ChatGPT 4.0: A Validation Study of 45 Commonly Used Instruments

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

INTRODUCTION:

Patient-reported outcome measures (PROMs) are integral to foot and ankle surgery, offering critical insights into pain, function, and postoperative recovery from the patient's perspective. Despite their value, the effectiveness of PROMs is compromised when they are written above recommended reading levels. Given that only 12% of U.S. adults demonstrate proficient health literacy, there is concern that many PROMs may be inaccessible, potentially undermining their interpretability and clinical utility.

Recent advances in large language models (LLMs) such as ChatGPT 4.0 offer a novel approach to this problem. With the capacity to simplify complex medical language while preserving clinical meaning, ChatGPT may help revise PROMs to meet the sixth-grade readability standard endorsed by the NIH and AMA.

This study assessed ChatGPT 4.0's ability to revise 45 commonly used PROMs in foot and ankle surgery, aiming to improve readability without compromising the instruments' conceptual integrity or clinical content.

METHODS:

This study involved no patient data and was therefore exempt from IRB review. Forty-five PROMs were selected based on their frequent use in foot and ankle orthopedic research and clinical practice, including the FAOS, AOFAS, MOXFQ, and PROMIS measures. PROMs were uploaded into ChatGPT 4.0 as either image-based or PDF documents.

Each PROM was processed using a standardized prompt:

"Please revise the attached patient-reported outcome measure so that it is written at a sixth-grade reading level or lower. Focus on simplifying medical terms, shortening sentences, and improving clarity for readers without a medical background. Preserve the clinical meaning and intent of the original text."

The revised PROMs were analyzed using Readable.com to compute 12 established readability metrics. A fellowship-trained foot and ankle orthopedic surgeon evaluated each revised PROM for accuracy and integrity. Revisions were categorized as either acceptable or error-containing based on predefined criteria, including: (1) changes in item meaning, (2) alterations to response scales, or (3) omission of key temporal or anatomical details.

Due to non-normality in the distribution of pre- and post-revision scores (confirmed by Shapiro-Wilk test), the Exact Sign Test was employed for statistical analysis, with significance set at $\alpha = 0.05$.

RESULTS:

All 12 readability metrics demonstrated statistically significant improvement following ChatGPT revision ($p < 0.001$). Key improvements included a 22% reduction in average sentence length, a 19% decrease in polysyllabic words, and a 15% drop in overall word count. Revised PROMs reached or surpassed sixth-grade readability thresholds on 8 of 9 grade-based metrics, including SMOG, Flesch-Kincaid, and Fry Graph.

Despite readability gains, 26 of 45 revised PROMs (57.8%) contained at least one clinically significant error. The most common issues were: alteration of validated response scales (29%), oversimplification or removal of key anatomical references (24%), and changes in question intent or meaning (16%). Lexical diversity improved modestly, and grammatical consistency was preserved.

DISCUSSION AND CONCLUSION:

ChatGPT 4.0 substantially improves the readability of PROMs commonly used in foot and ankle surgery, offering a powerful tool to align patient-facing materials with established health literacy guidelines. However, the frequent introduction of content-altering errors highlights the critical need for clinician oversight when implementing AI-revised instruments.

Future studies should explore refinement of prompting techniques, integration of automated validation tools, and the impact of revised PROMs on actual patient comprehension and reporting accuracy in clinical settings.