Can AI Replace Patient Reported Outcome Measures? A Pilot Study

Arianne T. Salunga, Austin Stoner[1], Jonathan Wen, Hiba Naz, Amishi Jobanputra[2], Dilan de Silva, Kali R Tileston, John Schoeneman Vorhies
[1]School of Medicine, [2]Department of Orthopaedic Surgery

INTRODUCTION: Patient reported outcome measures (PROMs) are valuable but can be time consuming and limited in scope. Artificial intelligence (AI)-powered scribe large language models (LLMs) can analyze speech during standard clinical encounters and generate clinical notes. Here, we assess the potential of LLMs to predict PROMs through analysis of transcripts of clinic interactions.

METHODS: We analyzed 47 clinical visits from a pediatric orthopedic spine clinic, where patients completed Patient-Reported Outcomes Measurement Information System (PROMIS) surveys. Clinic encounters were recorded and transcribed using a secure LLM, which we trained to estimate patient responses to PROMIS short form questions. Patients were stratified into new and follow-up visits, and paired t-tests compared AI predictions to actual scores.

RESULTS: Among new patient visits, AI estimated PROMIS scores were not significantly different ($p>.05$) from the patient reported scores in the domains of peer relationships, physical stress, physical activity, pain behavior, mobility, and pain interference. AI estimated significantly better PROMIS scores in the domains of anxiety and depressive symptoms. Among follow-up patient visits, AI estimated PROMIS scores were not significantly different ($p>.05$) from the patient reported scores in the domains of physical stress, anxiety, and pain behavior. AI estimated significantly worse scores in the domains of peer relationships, physical activity, mobility, and significantly better scores the domains of depressive symptoms and pain interference.

DISCUSSION AND CONCLUSION: AI demonstrated greater accuracy in estimating T-scores in the new versus follow-up patient group, though it underestimated disability related to depression and anxiety. LLMs show promise in quantifying patient outcomes, potentially streamlining PROM collection.

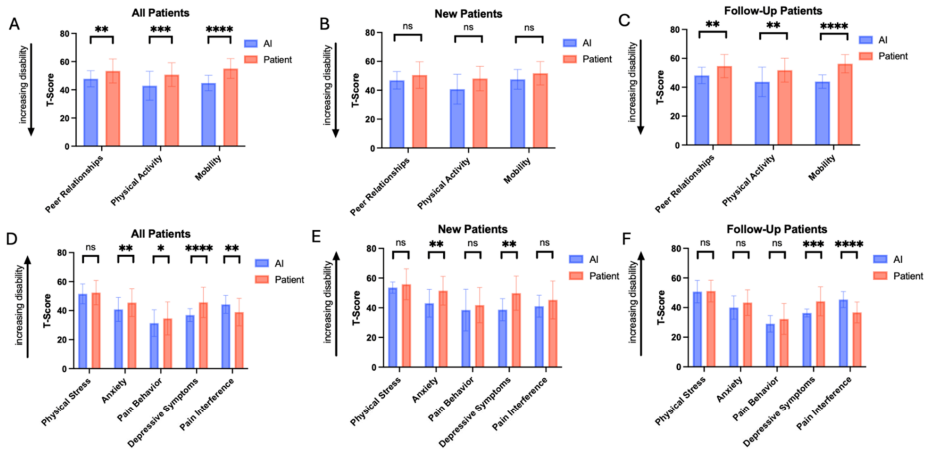

Figure 1. Comparison of AI- Interpolated and Patient-Reported PROMIS Measures. Higher t-scores are associated with the patient possessing "more" of that domain". Arrows indicate the direction of t-score corresponding to disability. Data are presented as mean +/ - SD. Statistical comparisons designated as paired t-tests between AI-interpolated and patient-reported t-scores. **p=0.01, ***p<0.001, ****p <0.001. ns, not significant.