Evaluation of ChatGPT in Assessing Postoperative Outcomes Following Flexor Tendon Repair (FTR)

Kismat A Touhid, Firdavs Kurbanov, Amy Phan, Constantinos Ketonis

INTRODUCTION: Large language models (LLMs) like ChatGPT have the potential to ameliorate the administrative burden hand surgeons face when querying the electronic medical record (EMR) for patient data. Our pilot case series evaluated the reliability of ChatGPT-4 in synthesizing rehabilitation data from postoperative EMR notes following FTR to assess its utility and limitations.

METHODS: Postoperative progress notes for 5 FTR patients were anonymized and input into ChatGPT-4 Turbo with memory disabled. Using a standardized multi-prompt framework, we directed the model to: (1) generate qualitative summary of patient recovery, (2) extract and tabulate quantitative metrics from postoperative notes, and (3) graph the data for easy visualization. The output was then cross-referenced against original notes within the EMR for accuracy and omissions.

RESULTS: Analysis of 95 clinical notes (29,000 words) across postoperative office, physical therapy (PT), and occupational therapy (OT) documentation revealed that GPT-4 achieved sufficient coverage of postoperative data without generating hallucinated claims. However, qualitative summaries demonstrated contextual omissions (e.g., social determinants, injury etiology) in 4/5 patients (Table 1, Figure 1). Quantitative metrics, including joint-specific ranges of motion (ROM), grip strength, and pain scores, were tabulated accurately and with precision; however, graphical outputs exhibited inconsistencies: omitted extension ROM trajectories in 2/5 patients, omitted total active motion (TAM) in 1/5 patients, omitted non-injured limb grip strength in 2/5 patients, and misrepresentation of patient-reported pain ranges in 2/5 patients (Table 2).

DISCUSSION AND CONCLUSION:

Integrating structured prompting frameworks allows ChatGPT-4 to reliably aggregate postoperative rehabilitation data with high fidelity. However, its limitations in contextual granularity and graphical representation necessitate clinician-guided oversight to mitigate the risk of oversimplification in complex recovery trajectories.

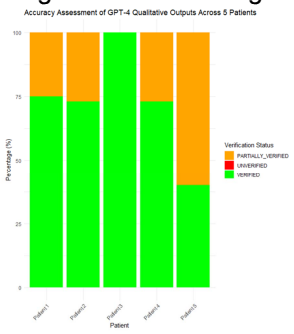

Table 1. Accuracy Assessment of GPT-4 Outputs: Verification Rates for Qualitative Data Across Five Flexor Tendon Repair Patients

Patient ID *(Table reference)*	Total # of Qualitative Claims based on Clinical Record Prompting	% of Qualitative Claims Verified against Clinical Record (VERIFIED)	% of Qualitative Claims with Potential Contextual Omissions (PARTIALLY VERIFIED)	% of Qualitative Claims Unverified against Clinical Record (NOT VERIFIED)
Patient 1 *(Table A1)*	12	75% (9/12)	25% (3/12)	0% (0/12)
Patient 2 *(Table A3)*	11	73% (8/11)	27% (3/11)	0% (0/11)
Patient 3 *(Table A5)*	12	100% (12/12)	0% (0/12)	0% (0/12)
Patient 4 *(Table A7)*	11	73% (8/11)	27% (3/11)	0% (0/11)
Patient 5 *(Table A9)*	10	40% (4/10)	60% (6/10)	0% (0/10)

Table 2. Accuracy Assessment of GPT-4 Outputs: Verification Rates for Quantitative Data Across Five Flexor Tendon Repair Patients

Patient ID	Accuracy of Table Metrics	Graphical Issues (Omissions/Inaccuracies)
Patient 1 *(Table A2; Figure 1)*	MCP Extension/Flexion: 18/18 PIP Extension/Flexion: 18/18 DIP Extension/Flexion: 18/18 Pain Score: 18/18 Grip Strength: 0/3 Dominant Injured Limb 0/3 Non-dominant Non-injured limb	MCP Extension Graph omitted; Grip Strength Value Graphs for Non-dominant Non-injured Limb omitted
Patient 2 *(Table A4; Figure A2)*	MCP Extension/Flexion: 4/4 PIP Extension/Flexion: 5/5 DIP Extension/Flexion: 5/5 Pain Score: 8/8	Pain Score Ranges averaged instead of reporting Full Range
Patient 3 *(Table A6; Figure A3)*	MCP Extension/Flexion: 10/20 PIP Extension/Flexion: 20/20 DIP Extension/Flexion: 20/20 Pain Score: 19/19 Grip Strength: 5/5 Dominant injured limb 5/5 Non-dominant Non-injured limb TAM: 19/19	MCP/PIP/DIP Extension Graphs Omitted
Patient 4 *(Table A8; Figure A4)*	MCP Extension/Flexion: 32/32 PIP Extension/Flexion: 32/32 DIP Extension/Flexion: 32/32 Pain Score: 12/12 Grip Strength: 0/7 Dominant injured limb 0/7 Non-dominant Non-injured limb TAM: 27/27	MCP/PIP/DIP Extension Graphs omitted; Pain Score Ranges plotted only Lower Bound; Grip Strength Value Graphs for Non-dominant Non-injured Limb Omitted
Patient 5 *(Table A10; Figure A5)*	MCP Extension/Flexion: 10/10 PIP Extension/Flexion: 10/10 DIP Extension/Flexion: 10/10 TAM: 5/5	Total Active Motion (TAM) Graph omitted