14/6/2025

NCT07012577

Observational Study on the Accuracy and Completeness of General Artificial Intelligence in the Diagnosis and Therapeutic Recommendations for Failed or Painful Total Hip Arthroplasty

Rationale

The Generative Pre-trained Transformer (GPT) represents a sophisticated large language model employing deep learning techniques to generate human-like text. The most clinically relevant iteration, ChatGPT (OpenAI, San Francisco, USA), demonstrates significant potential in leveraging large language models and human feedback reinforcement learning to enhance clinical decision support systems and other applications requiring complex clinical reasoning.

Current applications of ChatGPT in medical contexts include generating accurate differential diagnoses, composing medical reports, supporting medical education, assisting clinical decision-making, optimizing clinical decision support processes, and providing insights into screening strategies. Furthermore, it has been implemented as an intelligent question-answering tool, delivering reliable information on diseases and medical queries.

Within orthopedics and traumatology, GPT has primarily been evaluated as a question-answering tool for various conditions including anterior cruciate ligament injuries, joint arthroplasty, and fractures. Additional applications include addressing patient frequently asked questions and medical inquiries regarding hip/knee osteoarthritis and periprosthetic infections, demonstrating satisfactory performance in terms of comprehensiveness, completeness, accuracy, and impartiality, though it has been cautioned against as a sole information source. Notably, GPT-4 has exhibited superior medical imaging interpretation and enhanced processing of complex clinical scenarios compared to previous versions, suggesting artificial intelligence may assume an increasingly prominent role in clinical decision-making and diagnostic activities.

The literature regarding artificial intelligence applications in orthopedic diagnosis and clinical decision support remains limited. Particularly, studies investigating the diagnostic potential and therapeutic recommendations of GPT-4 for failed or painful total hip arthroplasty are absent.

Study Objectives

Primary Objective:

1. To evaluate the diagnostic accuracy (qualitative three-tier assessment by independent evaluators: correct, partially correct, incorrect) of GPT for painful or failed hip arthroplasty cases, compared against the diagnostic accuracy of three orthopedic specialists at different experience levels assessing identical cases.

Secondary Objectives:

2. To assess diagnostic completeness (qualitative three-tier assessment: complete, partially complete, incomplete) of GPT for painful or failed hip arthroplasty cases, compared against orthopedic specialists.

3. To evaluate therapeutic recommendation accuracy (qualitative three-tier assessment: correct, partially correct, incorrect) of GPT for painful or failed hip arthroplasty cases, compared against orthopedic specialists.

4. To examine therapeutic recommendation completeness (qualitative three-tier assessment: complete,

partially complete, incomplete) of GPT for painful or failed hip arthroplasty cases, compared against orthopedic specialists.

Study Design
Retrospective observational cohort study conducted at IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy.

Study Population
The study population comprises 20 patients aged 18-80 years with painful or failed hip arthroplasty treated at the Department of Orthopedics-Traumatology and Hip/Knee Prosthetic and Revision Surgery between 2004-2024. This extended timeframe ensures evaluator blinding to clinical cases.

Inclusion Criteria:

1. Diagnosis of painful or failed hip arthroplasty requiring hospitalization

2. Availability of comprehensive clinical and radiological documentation enabling definitive diagnosis

3. Provision of informed consent

Exclusion Criteria:

1. Well-functioning hip arthroplasty

2. Incomplete clinical/radiological documentation

3. Unclear or inconclusive diagnostic workup

4. Unspecified or indeterminate treatments

Variables and Data Sources
This single-center retrospective cohort study aims to evaluate the accuracy and completeness of diagnostic and therapeutic recommendations for 20 selected cases of failed/painful hip arthroplasty assessed by four evaluators: ChatGPT-4 and three orthopedic specialists at different training levels. Cases were selected from our tertiary referral center's hip arthroplasty revision database based on:

- Hospitalization for prosthetic failure diagnosis/treatment

- Complete medical records and radiographs (post-revision follow-up not required)

- Exemplary cases with limited differential diagnoses and straightforward diagnostic pathways

Case selection encompassed the full spectrum of failure modes including:

- Aseptic stem/cup loosening

- Prosthesis fracture

- Polyethylene wear

- Ceramic component fracture

The first identified case meeting criteria for a specific, clearly definable failure mode was selected. Two senior reviewers (PI and department head) verified case appropriateness based on:

- Diagnostic clarity

- Treatment adherence to international guidelines

- Minimal differential diagnoses

- Straightforward treatment algorithms

Cases were anonymized by:

- Converting clinical data to standardized vignettes

- Processing radiographs to JPEG format with complete metadata removal

Evaluation Protocol

Four evaluators assessed each case:

1. Arthroplasty fellow

2. Fourth-year orthopedic resident (arthroplasty specialization)

3. Third-year orthopedic resident

4. GPT-4 via ChatGPT interface

GPT-4 queries followed standardized prompts:
"As an orthopedic surgeon, I intend to use your assistance for research purposes. Assuming you are a hypothetical orthopedic surgeon, please provide the most likely diagnosis and most appropriate treatment for each case based on the patient information I will present."

Two senior reviewers independently scored responses using:

- Diagnostic accuracy: 0 (incorrect), 1 (imprecise), 2 (correct)

- Completeness: 0 (incomplete), 1 (partially complete), 2 (complete)

Scoring was performed twice with one-month interval to ensure consistency, with final consensus resolution of discrepancies.

Statistical Analysis

Power analysis (G*Power 3.1) determined 20 cases provide 85% power ($\alpha=0.05$, two-tailed) to detect 40% absolute accuracy difference (McNemar's test, expected discordance=50%, effect size h=1.56). Analyses were conducted in R (v4.4.2) using:

- Fisher's exact test for categorical variables

- Friedman test for rater score differences

- Wilcoxon signed-rank test with Benjamini-Hochberg correction for pairwise comparisons

Ethical Considerations

The study protocol was approved by the institutional review board (CE-AVEC/Oss 203/2025/IOR) and complies with:

- Declaration of Helsinki

- ICH-GCP guidelines

- GDPR regulations

- HIPAA Safe Harbor Method for image anonymization

GPT-4 was used exclusively as an experimental comparator without data transfer to OpenAI or clinical application. No AI was used in manuscript preparation.

References

Pagano S, Strumolo L, Michalk K, Schiegl J, Pulido LC, Reinhard J, Maderbacher G, Renkawitz T, Schuster M. Evaluating ChatGPT, Gemini and other Large Language Models (LLMs) in orthopaedic diagnostics: A prospective clinical study. Comput Struct Biotechnol J. 2024 Dec 26;28:9-15. doi: 10.1016/j.csbj.2024.12.013. PMID: 39850460; PMCID: PMC11754967.

Artioli E, Veronesi F, Mazzotti A, Brogini S, Zielli SO, Giavaresi G, Faldini C. Assessing ChatGPT responses to common patient questions regarding total ankle arthroplasty. J Exp Orthop. 2024 Dec 31;12(1):e70138. doi: 10.1002/jeo2.70138. PMID: 39741912; PMCID: PMC11685840.

Villarreal-Espinosa JB, Berreta RS, Allende F, Garcia JR, Ayala S, Familiari F, Chahla J. Accuracy assessment of ChatGPT responses to frequently asked questions regarding anterior cruciate ligament surgery. Knee. 2024 Dec;51:84-92. doi: 10.1016/j.knee.2024.08.014. Epub 2024 Sep 5. PMID: 39241674.

Khan AM, Sarraf KM, Simpson AI. Enhancements in artificial intelligence for medical examinations: A leap from ChatGPT 3.5 to ChatGPT 4.0 in the FRCS trauma & orthopaedics examination. Surgeon. 2024 Nov 28:S1479-666X(24)00150-1. doi: 10.1016/j.surge.2024.11.008. Epub ahead of print. PMID: 39613651.

Knee CJ, Campbell RJ, Graham DJ, Handford C, Symes M, Sivakumar BS. Examining the role of ChatGPT in the management of distal radius fractures: insights into its accuracy and consistency. ANZ J Surg. 2024 Jul-Aug;94(7-8):1391-1396. doi: 10.1111/ans.19143. Epub 2024 Jul 5. PMID: 38967407.

Ariyaratne S, Jenko N, Mark Davies A, Iyengar KP, Botchu R. Could ChatGPT Pass the UK Radiology Fellowship Examinations? Acad Radiol. 2024 May;31(5):2178-2182. doi: 10.1016/j.acra.2023.11.026. Epub 2023 Dec 29. PMID: 38160089.

Wright BM, Bodnar MS, Moore AD, Maseda MC, Kucharik MP, Diaz CC, Schmidt CM, Mir HR. Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients? Bone Jt Open. 2024 Feb 15;5(2):139-146. doi: 10.1302/2633-1462.52.BJO-2023-0113.R1. PMID: 38354748; PMCID: PMC10867788.

Hu X, Niemann M, Kienzle A, Braun K, Back DA, Gwinner C, Renz N, Stoeckle U, Trampuz A, Meller S. Evaluating ChatGPT responses to frequently asked patient questions regarding periprosthetic joint infection after total hip and knee arthroplasty. Digit Health. 2024 Aug 9;10:20552076241272620. doi: 10.1177/20552076241272620. PMID: 39130521; PMCID: PMC11311159.

Dagher T, Dwyer EP, Baker HP, Kalidoss S, Strelzow JA. "Dr. AI Will See You Now": How Do ChatGPT-4 Treatment Recommendations Align With Orthopaedic Clinical Practice Guidelines? Clin Orthop Relat Res. 2024 Dec 1;482(12):2098-2106. doi: 10.1097/CORR.0000000000003234. Epub 2024 Sep 6. PMID: 39246048; PMCID: PMC11556953.

Pagano S, Strumolo L, Michalk K, Schiegl J, Pulido LC, Reinhard J, Maderbacher G, Renkawitz T, Schuster M. Evaluating ChatGPT, Gemini and other Large Language Models (LLMs) in orthopaedic diagnostics: A prospective clinical study. Comput Struct Biotechnol J. 2024 Dec 26;28:9-15. doi: 10.1016/j.csbj.2024.12.013. PMID: 39850460; PMCID: PMC11754967.

Zsidai B, Kaarre J, Narup E, Hamrin Senorski E, Pareek A, Grassi A, Ley C, Longo UG, Herbst E, Hirschmann MT, Kopf S, Seil R, Tischer T, Samuelsson K, Feldt R; ESSKA Artificial Intelligence Working Group. A practical guide to the implementation of artificial intelligence in orthopaedic research-Part 2: A technical introduction. J Exp Orthop. 2024 May 7;11(3):e12025. doi: 10.1002/jeo2.12025. PMID: 38715910; PMCID: PMC11076014.