



Original Research

AI-Generated Graduate Medical Education Content for Total Joint Arthroplasty: Comparing ChatGPT Against Orthopaedic Fellows

Ryan DeCook, BS^{a,*}, Brian T. Muffly, MD^b, Sania Mahmood, MD^b, Christopher T. Holland, MD, MS^c, Ayomide M. Ayeni, BS^b, Michael P. Ast, MD^d, Michael P. Bolognese, MD^c, George N. Guild III, MD^b, Neil P. Sheth, MD^e, Christian A. Pean, MD, MS^c, Ajay Premkumar, MD, MPH^b

^a Philadelphia College of Osteopathic Medicine, Swannee, GA, USA

^b Department of Orthopaedic Surgery, Emory University School of Medicine, Atlanta, GA, USA

^c Department of Orthopaedic Surgery, Duke University School of Medicine, Durham, NC, USA

^d Hospital for Special Surgery, New York, NY, USA

^e Department of Orthopaedic Surgery, Perelman School of Medicine, Philadelphia, PA, USA

ARTICLE INFO

Article history:

Received 15 January 2024

Received in revised form

20 March 2024

Accepted 28 April 2024

Keywords:

Artificial intelligence

Orthopaedic surgery

Education

Total joint arthroplasty

ChatGPT

ABSTRACT

Background: Artificial intelligence (AI) in medicine has primarily focused on diagnosing and treating diseases and assisting in the development of academic scholarly work. This study aimed to evaluate a new use of AI in orthopaedics: content generation for professional medical education. Quality, accuracy, and time were compared between content created by ChatGPT and orthopaedic surgery clinical fellows. **Methods:** ChatGPT and 3 orthopaedic adult reconstruction fellows were tasked with creating educational summaries of 5 total joint arthroplasty-related topics. Responses were evaluated across 5 domains by 4 blinded reviewers from different institutions who are all current or former total joint arthroplasty fellowship directors or national arthroplasty board review course directors.

Results: ChatGPT created better orthopaedic content than fellows when mean aggregate scores for all 5 topics and domains were compared ($P \leq .001$). The only domain in which fellows outperformed ChatGPT was the integration of key points and references ($P = .006$). ChatGPT outperformed the fellows in response time, averaging 16.6 seconds vs the fellows' 94 minutes per prompt ($P = .002$).

Conclusions: With its efficient and accurate content generation, the current findings underscore ChatGPT's potential as an adjunctive tool to enhance orthopaedic arthroplasty graduate medical education. Future studies are warranted to explore AI's role further and optimize its utility in augmenting the educational development of arthroplasty trainees.

© 2024 Published by Elsevier Inc. on behalf of The American Association of Hip and Knee Surgeons. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

Recent advancements in technology, particularly in the realm of artificial intelligence (AI), have seen substantial growth and integration across numerous medical subspecialties [1]. From diagnostic imaging to surgical robotics, AI's incorporation in the field of orthopaedic surgery and arthroplasty is revolutionizing patient care and could improve treatment outcomes [2]. While the initial implementation of AI has been related to patient selection, optimization of surgical workflows, and decision-making support, the

utilization of AI extends beyond patient care alone, and its potential in the educational realm has recently gained increased attention [3,4].

High-quality surgical educational materials are essential to developing a strong foundation of clinical orthopaedic knowledge. Such educational materials can provide a structured approach for surgical trainees to consolidate their knowledge, gain exposure to various pathologies, refine their decision-making skills in familiar and relatively complex clinical scenarios, and understand and synthesize the medical literature as it applies to areas of ongoing controversy. As a result, such resources are instrumental in ultimately helping to prepare trainees to provide quality patient care in real-world clinical settings. A growing number of trainees continue to seek high-quality educational materials to support their studies

* Emory MSK Institute: 21 Ortho Ln, Atlanta, GA 30329 USA.

E-mail address: ryandecook7@gmail.com

and professional development, and the market share for these resources is expected to increase by \$11.9 billion from 2021 to 2026 [5]. Creating high-quality, subspecialized educational content can be significantly time-consuming, even for the most experienced medical educators. As such, there is a growing need to better optimize this content creation process and alleviate some of the educator burden. In this context, AI-generated content could be valuable to content creators and consumers of professional orthopaedic educational materials.

Previous research on AI and medical writing has primarily focused on the AI generation of scientific abstracts and manuscripts [6]. When comparing ChatGPT-generated to real scientific abstracts, Gao and colleagues found that AI can produce such realistic abstracts that they are difficult to differentiate from human-authored content. Similar findings have been observed in the AI creation of full academic manuscripts. ChatGPT has demonstrated the capability of generating coherent research articles that closely resemble authentic articles published by academic researchers in peer-reviewed journals [7].

Although AI has shown promise in various areas of orthopaedic surgery, its application in generating content sophisticated enough for graduate medical education remains to be discovered. To our knowledge, research has yet to examine the ability of AI to create quality educational materials in total joint arthroplasty (TJA). The

primary objective of this study was to explore the potential of ChatGPT in enhancing the educational development of orthopaedic surgery and arthroplasty trainees by providing them with efficient access to accurate and relevant educational content. To do this, the authors asked 2 questions: 1) How does the quality and accuracy of AI-generated TJA educational content compare to that of orthopaedic arthroplasty fellows at different institutions? 2) How much more efficiently can these educational materials be created? By assessing the effectiveness of AI models such as ChatGPT in generating surgical educational content, the authors hypothesize that AI language models could ultimately be optimized to improve the development of graduate medical educational materials.

Material and methods

In an effort to establish proof of concept for AI's application in generating TJA educational content, 3 current orthopaedic surgery trainees from one of 2 adult reconstruction fellowships were recruited as study participants. Each of these fellows had passed Part I of the American Board of Orthopaedic Surgery certifying examination and was in the final 3 months of their respective arthroplasty fellowship training programs. The involvement of a limited number of fellows was secondary to the study's preliminary nature, aiming to gauge the feasibility and potential future

Please check ONE box (they are check boxes) in each section below for the above passage.

Content Accuracy and Relevance

- 0: Summary contains significant inaccuracies or is not relevant to the topic
- 1: Summary contains some inaccuracies or is somewhat relevant to the topic
- 2: Summary is mostly accurate and relevant to the topic

Organization and Clarity

- 0: Summary is disorganized and difficult to understand
- 1: Summary is somewhat organized but contains unclear sections
- 2: Summary is well-organized and easy to understand

Use of Appropriate Terminology

- 0: Summary uses incorrect terminology or lacks necessary medical terms
- 1: Summary uses some incorrect terminology or is missing some necessary medical terms
- 2: Summary uses correct and appropriate medical terminology throughout

Integration of Key Points and References

- 0: Summary fails to integrate key points or concepts from the topic and/or lacks proper references
- 1: Summary partially integrates key points or concepts from the topic and/or has some references missing or incorrect
- 2: Summary effectively integrates key points or concepts from the topic and includes proper references

Writing Mechanics and Formatting

- 0: Summary contains numerous grammatical, spelling, or formatting errors
- 1: Summary contains some grammatical, spelling, or formatting errors
- 2: Summary is free of grammatical, spelling, or formatting errors

Figure 1. Scoring rubric for summary evaluation.

directions for its role in graduated medical education for TJA. Given that this research was an assessment of content generation without direct human interventions or use of patient data, institutional review board approval was not deemed necessary. Fellows and a generative AI system (ChatGPT, version 4.0) were presented with 5 separate TJA topics and tasked to produce summaries sufficient for board-review materials. The prompts can be found in Figure 2. Respondents were instructed to provide summaries of 500 words or less in length and incorporate up to 6 American Medical Association style citations (Appendix A). Fellows self-reported both response start/end times as well as the total time for each submission. Four blinded, fellowship-trained arthroplasty surgeons served as reviewers. Each reviewer has served for over 10 years in resident and fellow education and is either a current or former national arthroplasty board review course director.

The evaluation rubric was constructed to assess the educational summaries comprehensively (Fig. 1). Each category within the rubric, including “content accuracy and relevance,” “organization and clarity,” “use of appropriate terminology,” and “writing mechanics and formatting,” in addition to “integration of key points and references,” was established to address distinct and critical aspects of quality educational content. The inclusion of references within the rubric evaluated not only the presence of citations but also how effectively these references were integrated to support key points, reflecting the participants’ adeptness at crafting a scholarly narrative. The structured nature of the rubric aimed to minimize subjectivity across all domains, providing a balanced and equitable scoring process and allowing for a holistic comparison of the summaries’ quality.

The combined grades from the arthroplasty fellows for each domain of individual prompts and overall were compared to ChatGPT. Study groups were assessed using statistical means and standard deviation measures. Comparisons between the groups were conducted using the Mann-Whitney *U* tests. Nonparametric distributions were confirmed after performing Shapiro-Wilks tests. *P*-values of less than .05 were considered to indicate statistical significance. All the statistical analyses were performed using RStudio software version 4.2.2 (Foundation for Statistical Computing).

1. Write a 500 word summary appropriate for orthopaedic surgery board certification review on the considerations for total hip arthroplasty vs hemiarthroplasty for femoral neck hip fractures. Include up to 6 AMA Style citations.
2. Write a 500 word summary appropriate for orthopaedic surgery board certification review on the considerations for cemented vs uncemented femoral stem fixation in primary total hip arthroplasty. Include up to 6 AMA Style citations.
3. Write a 500 word summary appropriate for orthopaedic surgery board certification review on the role of dual mobility in revision total hip arthroplasty. Include up to 6 AMA Style citations.
4. Write a 500 word summary appropriate for orthopaedic surgery board certification review on the considerations for the diagnosis of acute v chronic knee periprosthetic joint infection after total knee arthroplasty. Include up to 6 AMA Style citations.
5. Write a 500 word summary appropriate for orthopaedic surgery board certification review on the considerations for Cementless v cemented fixation in primary total knee arthroplasty. Include up to 6 AMA Style citations.

Figure 2. Prompts provided to fellows and ChatGPT for TJA topic summaries.

Table 1
Graded scores by domain.

Domain	Mean values		<i>P</i> value
	ChatGPT	Fellows	
All	1.72 ± 0.47	1.66 ± 0.5	<.001
Content accuracy and relevance	1.7 ± 0.57	1.65 ± 0.48	.233
Organization and clarity	1.75 ± 0.44	1.57 ± 0.59	.424
Use of appropriate terminology	1.85 ± 0.37	1.85 ± 0.36	.424
Integration of key points and references	1.5 ± 0.51	1.5 ± 0.54	.006 ^a
Writing mechanics and formatting	1.8 ± 0.41	1.72 ± 0.45	.424

Values are given as mean ± standard deviation.

Bold values indicate statistical significance (*P* < .05).

^a Despite the observation of identical mean scores, the application of the Mann-Whitney *U* test reveals a statistically significant discrepancy in this domain. This phenomenon arises due to the test’s focus on the entirety of the score distribution, specifically the variations in the dispersion of individual scores (assessed through ranking), rather than solely on mean values.

Results

Twenty total responses were analyzed across 5 domains: content accuracy and relevance, organization and clarity, use of appropriate terminology, integration of key points and references, and writing mechanics and formatting.

Mean scores between the ChatGPT and fellow cohorts were found to be insignificant within the domains of content accuracy and relevance (*P* = .233), organization and clarity (*P* = .424), use of appropriate terminology (*P* = .424), and writing mechanics and formatting for all prompts (*P* = .424; Tables 1 and 2). However, the integration of key points and references domain showed a significant variation and distribution among the fellows, with ChatGPT demonstrating superior performance (*P* = .006).

When data across all 5 topics and all 5 domains were aggregated, ChatGPT demonstrated superior performance compared to the fellows. The mean scores for ChatGPT and the fellows were 1.72 ± 0.47 and 1.66 ± 0.5, respectively (*P* < .001).

In terms of efficiency, generative AI significantly outpaced the fellows. The language model generated responses with a mean time of 16.6 ± 1.14 seconds per topic. In contrast, fellows took

Table 2
Graded scores by domain and prompt.

Domain	Prompt	Mean values		P value
		ChatGPT	Fellows	
Content accuracy and relevance	1	1.75 ± 0.5	1.75 ± 0.45	1.000
	2	1.75 ± 0.5	1.83 ± 0.39	1.000
	3	1.75 ± 0.5	1.42 ± 0.51	1.000
	4	2 ± 0	1.58 ± 0.51	NA ^a
	5	1.25 ± 0.96	1.67 ± 0.49	0.346
Organization and clarity	1	1.75 ± 0.5	1.5 ± 0.67	1.000
	2	2 ± 0	1.67 ± 0.49	NA ^a
	3	1.5 ± 0.58	1.67 ± 0.49	1.000
	4	1.75 ± 0.5	1.58 ± 0.67	1.000
	5	1.75 ± 0.5	1.42 ± 0.67	NA ^a
Use of appropriate terminology	1	2 ± 0	2 ± 0	NA ^a
	2	1.75 ± 0.5	1.83 ± 0.39	1.000
	3	2 ± 0	1.83 ± 0.39	NA ^a
	4	1.75 ± 0.5	1.67 ± 0.49	1.000
	5	1.75 ± 0.5	1.92 ± 0.29	1.000
Integration of key points and references	1	1.5 ± 0.58	1.58 ± 0.51	0.346
	2	1.25 ± 0.5	1.5 ± 0.52	0.149
	3	1.5 ± 0.58	1.5 ± 0.52	1.000
	4	1.75 ± 0.5	1.42 ± 0.67	1.000
	5	1.5 ± 0.58	1.5 ± 0.52	1.000
Writing mechanics and formatting	1	1.75 ± 0.5	1.58 ± 0.51	1.000
	2	2 ± 0	1.62 ± 0.52	NA ^a
	3	1.75 ± 0.5	1.72 ± 0.45	1.000
	4	1.75 ± 0.5	1.67 ± 0.50	NA ^a
	5	1.75 ± 0.5	1.58 ± 0.51	1.000

Values are given as mean ± standard deviation.

^a Invalid scores due to ties between groups.

considerably longer to create responses, with a mean time of 5656 seconds (1:34:16 ± 1:35:50) per prompt response ($P = .002$).

Discussion

In an effort to evaluate generative AI's potential application in the field of graduate medical education, the present study aimed to assess the quality and accuracy of AI-generated TJA board-review level content compared to that provided by arthroplasty fellows. As judged by blinded arthroplasty board-review course directors, ChatGPT outscored the fellows in overall generated content and specifically in the domain of "integration of key points and references" while scoring on par with the fellows in the remaining domains. Although statistically significant, it is admittedly difficult to interpret the relevance of the small scoring difference observed between mean values when all the data were aggregated. While different methods of prompt assessment exist within the educational literature, data reporting the magnitude of relevant scoring differences is relatively lacking. In the authors' view, this metric is not just about comparison but also a validation of AI's competence in producing educationally relevant content comparable (and in some aspects superior) to that produced by fellows.

Despite the inherent potential bias associated with fellows' self-reported timing data, the substantial difference in content creation time—AI averaging 16.6 seconds vs fellows' 94 minutes per prompt—is highly unlikely to impact the conclusion regarding AI's significant time-saving advantage. Our findings indicate that ChatGPT can efficiently generate educational content that is comparable to that generated by fellows, as judged by senior authors who are arthroplasty content creators. The study findings indicate that AI may serve as a complementary tool to traditional resources in the future of orthopaedic surgery education. Nonetheless, we advocate for future research to further explore the role of AI in this educational domain. The authors believe the current findings support the future promise of AI in increasing access to accurate and efficiently generated subspecialized medical education

content, including in time-sensitive study environments. ChatGPT has the potential to be a valuable resource for arthroplasty trainees seeking accurate and up-to-date information on specialized topics in TJA.

Our research findings echo similar themes of efficiency and accuracy observed in other AI-related studies, such as the one conducted by Dergaa and colleagues. This investigation emphasized the potential benefits of natural language processing technologies, like ChatGPT, in enhancing the efficiency of academic writing and research. It mirrors our own exploration into AI-generated content, which we discovered could serve as a valuable reservoir of accurate and up-to-date knowledge for arthroplasty trainees delving into specialized topics. Just as Dergaa et al. found, such advanced technologies may alleviate the workload on content creators in the educational sphere and refine the content creation process itself in the future.

Much like current AI research, our study method used a standardized prompt format with predetermined topics to ensure a certain level of uniformity, but, as Sechopoulos and colleagues have pointed out, it may not entirely reflect the rich intricacies and diversity characteristic of real-world scenarios or the broad spectrum of inquiries medical trainees typically face during their educational journey. Their study focused on the application of AI in breast cancer detection, which aligns with our own in the sense that both are essentially stand-alone evaluations of the capabilities of new AI systems. As they astutely noted, the real test of these AI models is their performance in genuine, practical situations, which can be most accurately assessed during comprehensive screening trials. Therefore, future research should extend our research scope to incorporate real-life, patient-specific cases. This will enable us to better evaluate the AI model's adaptability and efficacy in generating educational content that caters to a broader range of clinical scenarios.

While in this study, ChatGPT was able to generate accurate and relevant content that, in the aggregate, surpassed those of surgeons completing subspecialty fellowship training, the

current body of literature suggests that generative AI should be used cautiously in sophisticated academic settings. When evaluating the accuracy and reliability of AI-generated medical responses, Johnson et al. highlighted the need for further research and model development to correct inaccuracies and validate AI-generated content [8]. In their comparison of ChatGPT-generated and human-generated scientific abstracts, Gao and colleagues observed that although ChatGPT wrote believable scientific abstracts, the system may cite unsubstantiated data to support its claims [9]. Previous work has also demonstrated that while ChatGPT can adequately summarize the topic for a general medical or lay audience, it may be inadequate for a subspecialty audience of expert authors [10]. These authors attributed this discrepancy to AI-related poor search methodology, lack of organization and flow, inaccuracies or omissions in text and references, and a lack of novelty. Our study had different results, where ChatGPT scored on aggregate better than fellows, as judged by blinded subspecialty experts. The authors believe this discrepancy may stem from a lack of understanding about how these new technologies work and the nature of the language of the prompts used in this study. Dwivedi and colleagues explain through multidisciplinary perspectives that imparting knowledge about the limitations of ChatGPT to users is vital for them to reap the full advantages of this tool [11]. Additionally, it is essential for users to comprehend how generative AI tools, like ChatGPT, stand apart from other resources, such as search engines. Just like any other technology, the onus of appropriate usage of ChatGPT squarely falls on the users, highlighting the significance of their understanding and responsible engagement. These studies highlight the ongoing discussions and challenges surrounding the ethical and acceptable use of large language models like ChatGPT in scientific writing. The current study demonstrated that content created by ChatGPT can be reliable and a viable option for generating educational content at a sophisticated level.

This study acknowledges several limitations that are intrinsic to its exploratory nature. Most notably, the inclusion of only 3 orthopaedic surgery fellows was made without conducting a formal power analysis. This methodology was primarily driven by logistical considerations and the study's preliminary nature, aiming to establish a proof-of-concept for the application of AI in generating TJA educational content. Following a power analysis, the expansion of the fellow participant could provide a more comprehensive assessment of AI-generated content compared to human responses. This was offset by a desire to limit the number of blinded reviews required by each reviewer to avoid reviewer fatigue and poor-quality reviews. Similarly, while the blinded reviewers were fellowship-trained TJA surgeons with expertise in educating fellows and teaching board review courses, response evaluation remains inherently subjective. Future studies may incorporate a larger panel of expert reviewers and a greater number of diverse content generators to enhance the robustness and objectivity of prompt assessments. As prompts only focused on TJA topics, we are unable to comment on the broader applicability and generalizability of AI-generated educational materials in other orthopaedic subspecialties. Finally, the study focused exclusively on the quality, accuracy, and efficiency of AI-generated content compared to that of orthopaedic arthroplasty fellows. Although this aspect is crucial, it leaves other relevant questions unanswered, such as medical trainees' retention and long-term comprehension of AI-generated educational materials. Future investigations may assess the retention and application of knowledge acquired from AI-generated content in clinical practice.

Conclusions

Our study provides evidence of ChatGPT's potential to enhance medical education by providing efficient access to high-quality educational content. The AI model demonstrated proficiency in generating accurate, relevant, and well-organized summaries in a fraction of the time. Building on the insights gained from this study, future research will investigate the integration of AI tools like ChatGPT in the broader context of graduate medical education. Through exploration of AI's potential applications within this realm, it may ultimately complement human expertise as well as contribute to a more effective and comprehensive learning experience for medical professionals at all stages of their careers. Although AI-generated content shows promise in medical education, it should not be viewed as a substitute for critical thinking and human intelligence. AI-generated responses should be fact-checked for accuracy. In this light, generative AI may hold promise as a powerful adjunct tool to complement and enhance graduate medical education while retaining the valuable aspects of education that rely on human creativity and expertise.

Conflicts of interest

G. Guild is a paid consultant for Smith & Nephew, receives research support from Smith & Nephew, has stock options in TJO, and is a board/committee member of AAHKS. M. Ast is a paid consultant for Surgical Care Affiliates, BD, Bioventus, Conformis, Convatec, OrthoAlign, Stryker, Smith & Nephew, a paid speaker for Stryker, Smith & Nephew, OrthoAlign, receives research support from Smith & Nephew, has stock options in Parvizi Surgical Innovations, Osso VR, Ospitek, OrthoAlign, HS2, ConveyMED, has IP royalties in OrthoAlign, is an editorial board member of *Journal of Arthroplasty* and *JOEI*, and is a board/committee member of AAOS, AAHKS, FPA, and EOA. M. Bolognese receives royalties from Smith & Nephew, Zimmer, and TJO; is a speaker for Ethicon; is a paid speaker to Zimmer Biomet, TJO; has stock options in TJO, Amedica; receives research support from Depuy Synthes, Zimmer, KCI, Exactech, Biomet; receives other financial support from Acelity, AOA Omega, Smith & Nephew; is a board/committee member of OREF, *Arthroplasty Today*, and AAHKS; and is an editorial board member of *Journal of Arthroplasty*. N. P. Sheth is a paid consultant for Medacta, Smith & Nephew, and Zimmer; receives publishing royalties and financial support from Elsevier; receives research support from Medacta; and is a board/committee member of AAOS Now, AAOS OrthoInfo, Eastern Orthopaedic Association, and AHF. C. Pean is an Arthrex speaker, has stock options in Azra Care Inc, and is a paid consultant for Azra Care Inc. and Kaizen Clinical Partners. A. Premkumar has stock options in AccuJoint, Azra Care, Osgenic, Inc.; receives publishing royalties and financial support from Elsevier; and is an editorial board member of *HSS Journal*. All other authors declare no potential conflicts of interest.

For full disclosure statements refer to <https://doi.org/10.1016/j.artd.2024.101412>.

CRedit authorship contribution statement

Ryan DeCook: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Data curation. **Brian T. Muffly:** Writing – review & editing, Supervision, Resources, Project administration. **Sania Mahmood:** Data curation. **Christopher T. Holland:** Data curation. **Ayomide M. Ayeni:** Validation, Software, Formal analysis, Data curation. **Michael P. Ast:** Data curation. **Michael P. Bolognese:** Supervision, Resources, Data curation. **George N. Guild:** Data curation. **Neil P. Sheth:** Data curation. **Christian A. Pean:**

Investigation, Data curation. **Ajay Premkumar:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization.

References

- [1] Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 2023;47:33. <https://doi.org/10.1007/s10916-023-01925-4>.
- [2] Magruder ML, Delanois RE, Nace J, Mont MA. ChatGPT and other natural language processing artificial intelligence models in adult reconstruction. *J Arthroplasty* 2023;38:2191–2. <https://doi.org/10.1016/j.arth.2023.06.030>.
- [3] Farrow L, Ashcroft GP, Zhong M, Anderson L. Using artificial intelligence to revolutionise the patient care pathway in hip and knee arthroplasty (ARCHERY): protocol for the development of a clinical prediction model. *JMIR Res Protoc* 2022;11:e37092. <https://doi.org/10.2196/37092>.
- [4] Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023;11:887. <https://doi.org/10.3390/healthcare11060887>.
- [5] T. Test Preparation Market Size, Share, Growth, Trends, Industry Analysis Forecast 2027 n.d.. <https://www.technavio.com/report/test-preparation-market-industry-in-the-us-analysis>. [Accessed 21 July 2023].
- [6] Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biol Sport* 2023;40:615–22. <https://doi.org/10.5114/biolsport.2023.125623>.
- [7] Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R. A comparison of ChatGPT-generated articles with human-written articles [published online ahead of print, 2023 Apr 14]. *Skeletal Radiol* 2023;52:1755–8. <https://doi.org/10.1007/s00256-023-04340-5>.
- [8] Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Research Square*; 2023. <https://doi.org/10.21203/rs.3.rs-2566942/v1>. PREPRINT (Version 1).
- [9] Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med* 2023;6:75. <https://doi.org/10.1038/s41746-023-00819-6>.
- [10] Wu CL, Cho B, Gabriel R, et al. Addition of dexamethasone to prolong peripheral nerve blocks: a ChatGPT-created narrative review. *Reg Anesth Pain Med* 2023. <https://doi.org/10.1136/rapm-2023-104646>.
- [11] Dwivedi YK, Kshetri N, Hughes L, Hughes L, Slade EL, Jeyaraj A, et al. Opinion paper: “So what if ChatGPT wrote it?” multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inf Manage* 2023;71:102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>.

Appendix A

Responses to Prompt 1 (Fig. 2)

Fellow 1

For the displaced femoral neck fracture (FNF) in an older individual, the management options are hemiarthroplasty (HA) vs total hip arthroplasty (THA). Currently, the American Academy of Orthopaedic Surgeons (AAOS) Clinic Practice Guidelines give a “moderate” recommendation favoring THA over HA for these injuries on the basis of improved functional benefit at the risk of increasing complications. This management decision, however, necessitates a shared decision-making discussion between the surgeon, patient, and patient’s family/caretaker(s). There are 3 primary considerations that are helpful in guiding patient selection: patient age, patient preinjury functional status, and patient burden of medical comorbidities. Octogenarians and nonagenarians, for example, have a lower remaining life expectancy and, thus, a relatively lower need for extended longevity of an arthroplasty implant. In a randomized controlled trial by Chammout and colleagues, no difference in outcomes (hip function, pain, and quality of life) between procedures at 2 years postoperatively was observed in this patient cohort. The authors concluded that HA is suitable for this group in the short term. Second, the preinjury functional status of the patient should be considered. Was the patient ambulatory in the community (community-ambulator) vs primarily ambulatory around the home (household-ambulator)? In the case of the former, registry data suggests that use of THA has better functional outcomes and is more likely to be able to restore preinjury level of functional demand. For the household-ambulator who is returning to minimal/lower levels of preinjury function, HA is an appropriate option that avoids exposing the patient to the increased risks associated with THA (higher rates of instability/dislocation, longer operative times potentially increasing blood loss or risk of infection, etc.). Finally, the burden of concurrent medical comorbidities should be incorporated into decision-making. In those with radiographic evidence of pre-existing arthritis (osteoarthritis or inflammatory), THA may be a more appropriate option. When neurologic/neurodegenerative conditions and/or cognitive impairment are present, patients are at higher risk of instability and dislocation secondary to their underlying disease process and/or inability to comply with potential hip precautions. HA may be favored in this situation, as meta-analyses have demonstrated significantly higher rates of dislocation following THA. Patients with high Charlson comorbidity index scores (surrogate for burden of medical comorbidities) have an increased risk of perioperative complications. Such patients often have lower functional demands and may benefit from a shorter operation/less time under anesthesia in the form of HA.

References

- [1] American Academy of Orthopaedic Surgeons (AAOS) Clinical Practice Guidelines: Management of Hip Fractures in Older Adults. 2022.
- [2] OKU 5: Hip and Knee Reconstruction, Chapter 25. Management of Hip Fractures: Hemiarthroplasty vs Total Hip Arthroplasty - Results and Complications. 2017.
- [3] Chammout G, Kelly-Pettersson P, Hedbeck CJ, Stark A, Mukka S, Sköldenberg O, et al. HOPE-Trial: Hemiarthroplasty Compared with Total Hip Arthroplasty for Displaced Femoral Neck Fractures in Octogenarians: A Randomized Controlled Trial. *JBJS Open Access* 2019;4:e0059.
- [4] Leonardsson O, Rolfson O, Hommel A, Garellick G, Åkesson K, Rogmark C. Patient-reported outcome after displaced femoral neck fracture: A national survey of 4467 patients. *JBJS Am* 2013;95:1693–9.

Fellow 2

For intracapsular FNFs, there are several avenues for stabilizing the fracture and restoring function to the patient. For displaced FNFs, there is debate as to whether to proceed with HA vs total hip replacement. There have been many studies to gauge the outcomes between the 2 surgeries with respect to different outcome variables. A meta-analysis by Hopley et al. looking at 15 RTCs found that overall total hip replacements had lower reoperation rates and higher functional outcome scores at all intervals after a total hip replacement in comparison to HA. The meta-analysis does, however, note that there is a tendency for higher dislocation rates with total hip replacements in comparison to HA, though the data was not statistically significant. There have also been studies showing higher postoperative complication rates with total hip replacements, though there is no significant difference in 1-year mortality rates.

Reference

- [1] Hopley C, Stengel D, Ekkernkamp A, Wich M. Primary total hip arthroplasty vs hemiarthroplasty for displaced intracapsular hip fractures in older patients: systematic review. *BMJ* 2010;340:c2332. <https://doi.org/10.1136/bmj.c2332>.

Fellow 3

Introduction

Hip fracture rates in patients over 50 years old are expected to double by 2050. These fractures can significantly impact mortality and quality of life, with up to 24% mortality in the first year and reports of 50% mortality at 5 years.[1] Two common surgical options for intracapsular FNFs are HA and total hip arthroplasty (THA). THA is typically indicated for patients with displaced intracapsular FNF who have normal cognition and are independent ambulators prior to fracture with adequate bone quality. Patients who do not meet these criteria are likely to undergo HA.

Implant Selection

In a multicenter, randomized control trial comparing cemented HA patients were found to have better quality of life, a lower risk of periprosthetic fracture at 2 months, and a trend toward improved mortality at 12 months compared to uncemented HA.[2] Uncemented HA is associated with 3 times higher revision rate for aseptic loosening compared to cemented HA in elderly patients for FNF. [3] Furthermore, the use of a cemented femoral component in THA leads to significantly lower early hip-related complications such as periprosthetic fracture.[4] Acetabular wear is a primary concern when performing HA for FNF, with rates of conversion to THA between 1% and 10%. [5] Dislocation after THA for FNF is a significant risk with higher rates of THA dislocation compared to HA and for THA not performed for FNF. When considering cost, THA is a more expensive surgical procedure. THA in the long term may be more cost-effective due to the reported lower risk of conversion surgery and improved functional outcomes.

Outcomes

The outcomes of HA and THA for FNF have been extensively studied, with improved functional scores for both groups compared to preoperative values. THA has been shown to provide improved Harris hip score, Western Ontario and McMaster Universities Arthritis Index, and Oxford hip scores at 2 and 4 years compared to

HA, with better functional outcomes and a lower risk of revision surgery in younger patients with good bone quality.[6] However, in elderly patients ≥ 80 with poor bone quality and limited life expectancy, HA has been shown to have similar outcomes to THA with hip function, health-related quality of life, hip-related complications, and reoperations.[4]

Conclusions

In conclusion, treatment of displaced intracapsular FNFs with THA or HA is a patient-specific decision that depends on patient age, bone quality, level of function, and surgical expertise available. HA has consistently been shown to improve functional outcomes compared to preoperative measures, carries a risk of revision for acetabular wear, with the benefit of lower implant costs and surgical time. THA has the potential benefit of offering improved outcome measures postoperatively, reduced risk of revision procedures, and greater longevity compared to HA; however, THA carries a higher cost, risk of dislocation, and complication rate. The use of a cemented femoral component for either HA or THA has consistently been shown to reduce periprosthetic fracture and need for revision surgery. Implant selection, complications, outcomes, and cost should be carefully considered when making the decision between THA and HA.

References

- [1] Macaulay W, Pagnotto MR, Iorio R, Mont MA, Saleh KJ. Displaced femoral neck fractures in the elderly: hemiarthroplasty vs total hip arthroplasty. *J Am Acad Orthop Surg* 2006;14:287–93. <https://doi.org/10.5435/00124635-200605000-00004>.
- [2] Fernandez MA, Achten J, Parsons N, Griffin XL, Png ME, Gould J, et al. Cemented or Uncemented Hemiarthroplasty for Intracapsular Hip Fracture. *N Engl J Med* 2022;386:521–30. <https://doi.org/10.1056/NEJMoa2108337>.
- [3] Okike K, Chan PH, Prentice HA, Paxton EV, Burri RA. Association between uncemented vs cemented hemiarthroplasty and revision surgery among patients with hip fracture. *JAMA* 2020;323:1077–84. <https://doi.org/10.1001/jama.2020.1067>.
- [4] Chammout G, Kelly-Pettersson P, Hedbeck CJ, Stark A, Mukka S, Skoldenberg O. HOPE-Trial: hemiarthroplasty compared with total hip arthroplasty for displaced femoral neck fractures in octogenarians: a randomized controlled trial. *JB JS Open Access* 2019;4:e0059. <https://doi.org/10.2106/JBJS.OA.18.00059>.
- [5] Chammout G, Muren O, Laurencikas E, Bodén H, Kelly-Pettersson P, Sjöo H, et al. More complications with uncemented than cemented femoral stems in total hip replacement for displaced femoral neck fractures in the elderly. *Acta Orthop* 2017;88:145–51. <https://doi.org/10.1080/17453674.2016.1262687>.
- [6] Hedbeck CJ, Enocson A, Lapidus G, Blomfeldt R, Törnkqvist H, Ponzer S, et al. Comparison of bipolar hemiarthroplasty with total hip arthroplasty for displaced femoral neck fractures: a concise 4-year follow-up of a randomized trial. *J Bone Joint Surg Am* 2011;93:445–50. <https://doi.org/10.2106/JBJS.J.00474>.

ChatGPT

Title: Considerations for Total Hip Arthroplasty vs Hemiarthroplasty for Femoral Neck Hip Fractures

Introduction

Femoral neck fractures represent a common and challenging clinical scenario in orthopaedic surgery. The surgical management of these injuries often involves either total hip arthroplasty (THA) or hemiarthroplasty (HA) as the primary treatment options. This summary aims to discuss the indications, advantages, and disadvantages of THA and HA in the context of femoral neck fractures, drawing from the most recent evidence.

Indications for total hip arthroplasty and hemiarthroplasty

The choice between THA and HA for femoral neck fractures depends on multiple factors, including patient age, preinjury functional status, comorbidities, and surgeon experience [1]. Generally, THA is recommended for younger, healthier, and more active patients with higher functional demands, while HA is more suitable for older patients with lower functional requirements and a higher risk of perioperative complications [2].

Advantages of total hip arthroplasty

THA provides several advantages over HA, including better functional outcomes, improved hip stability, and a lower rate of revision surgery (3). Patients who undergo THA typically experience less pain and better hip function, leading to improved quality of life and higher patient satisfaction (4). Additionally, THA offers better joint biomechanics and load distribution across the hip joint, resulting in lower risk of implant wear and loosening (5).

Disadvantages of total hip arthroplasty

Despite its advantages, THA also has some disadvantages compared to HA. It is a more complex procedure with a higher risk of perioperative complications, such as infection, bleeding, and dislocation (6). Furthermore, THA is associated with increased operative time and potentially higher cost, which may limit its applicability in certain healthcare settings (3).

Advantages of hemiarthroplasty

HA is a simpler and faster procedure compared to THA, with fewer surgical steps and a lower risk of perioperative complications [2]. This makes HA a more suitable option for elderly patients or those with multiple comorbidities who may not tolerate the longer operative time and potential complications of THA. Additionally, HA is generally less expensive than THA, making it a more cost-effective option in resource-limited settings (6).

Disadvantages of hemiarthroplasty

The main disadvantage of HA is the increased risk of long-term complications, such as acetabular erosion, implant wear, and the need for revision surgery (3). Furthermore, HA is associated with inferior functional outcomes and increased pain compared to THA, which may negatively impact the patient's quality of life (4).

Conclusions

In summary, the choice between THA and HA for femoral neck fractures depends on multiple factors, including patient age, pre-injury functional status, comorbidities, and surgeon experience. THA offers better functional outcomes and hip stability but carries a higher risk of perioperative complications. In contrast, HA is a simpler and faster procedure with a lower risk of perioperative complications, making it more suitable for elderly patients or those with multiple comorbidities. Ultimately, the decision between THA and HA should be individualized based on patient-specific factors and the surgeon's experience.

References

- [1] Bhandari M, Swiontkowski M. Management of Acute Hip Fracture. *N Engl J Med* 2017;377:2053–62. <https://doi.org/10.1056/NEJMcp1611090>.
- [2] Rogmark C, Leonardsson O. Hip arthroplasty for the treatment of displaced fractures of the femoral neck in elderly patients. *Bone Joint J* 2016;98-B:291–7.