Journal of Orthopaedic Experience & Innovation
Clinical, Economic, and Innovative Orthopaedic Experience

Shah, Ayush, Sophia Mavrommatis, Linzie Wildenauer, Deborah Bohn, and Alexander Vasconcellos. 2024. "Performance of ChatGPT on Hand Surgery Board-Style Examination Questions." *Journal of Orthopaedic Experience & Innovation* 5 (2). https://doi.org/10.60118/001c.118938.

Research Article

Performance of ChatGPT on Hand Surgery Board-Style Examination Questions

Ayush Shah[1a], Sophia Mavrommatis, MD[1b], Linzie Wildenauer[1c], Deborah Bohn, MD[2d], Alexander Vasconcellos, MD[2e]

[1] University of Minnesota Medical School, Twin Cities, [2] Orthopedic Surgery, University of Minnesota Medical School, Twin Cities

Keywords: artificial intelligence, board examinations, ChatGPT, hand surgery, large language models

https://doi.org/10.60118/001c.118938

Journal of Orthopaedic Experience & Innovation

Vol. 5, Issue 2, 2024

Background

The performance of large-language models, such as ChatGPT, on medical and sub-specialty examinations has been preliminarily explored in fields such as radiology, obstetrics and gynecology, and orthopedic surgery. However, no literature assessing ChatGPT's ability to answer hand surgery exam questions exists. This study's purpose was to evaluate ChatGPT's performance on hand surgery board-style examination questions.

Methods

All questions from the American Society for Surgery of the Hand (ASSH) Hand 100 Exam, Beginner, and Intermediate Assessment tools were entered into ChatGPT-3.5. Responses were regenerated two times to identify inconsistencies. Duplicate questions, questions with figures and/or videos, and questions that ChatGPT refused to provide a response to were excluded. ChatGPT's correct response rate, answer modifications, and human accuracy were recorded.

Results

117 questions from the 3 assessment tools were analyzed: 49 from the ASSH Hand 100, 32 from the Beginner, and 36 from the Intermediate Assessment tools. On ChatGPT's initial attempt, 40.82% (20/49), 50.0% (16/32), 38.89% (14/36) of questions were correctly answered, respectively. Overall, ChatGPT correctly answered 50/117 (42.7%) of questions on the first try. ChatGPT excelled in topics (>60% correct) of mass/tumor, nerve, wrist, and performed poorly (<40% correct) on topics regarding anatomy/basic science/imaging, brachial plexus, congenital, elbow, tendon, and vascular disorders, trauma. On the

a Ayush Shah is a third-year medical student at the University of Minnesota Twin Cities. His research interests include studying clinical outcomes, social disparities, medical education, and the application of artificial intelligence & machine learning in orthopedic patient care and education.

Conflicts of Interest Statement for Ayush Shah

b Sophia Mavrommatis graduated from Washington University in St. Louis in 2018 and graduated from the University of Minnesota Medical School in 2024. She is starting her first year of orthopedic surgery residency training at Mayo Clinic in Rochester, MN.

Conflicts of Interest Statement for Dr. Mavrommatis

c Linzie Wildenauer is a medical student at the University of Minnesota Medical School, graduating in 2025.

Conflicts of Interest Statement for Linzie Wildenauer

d Deborah Bohn is an associate professor of orthopedic surgery at the University of Minnesota with a subspecialty interest in adult and pediatric upper extremity surgery.

Visit Dr. Bohn's Website

Conflicts of Interest Statement for Dr. Bohn

e Dr. Vasconcellos is an orthopedic surgeon with an interest in hand and upper extremity surgery.

Visit Dr. Vasconcellos's Website

Conflicts of Interest Statement for Dr. Vasconcellos

Beginner and Intermediate Exams, humans correctly answered 56.64% and 62.73% of questions, respectively.

Conclusions

ChatGPT can correctly answer simpler hand surgery questions but performed poorly when compared to humans on higher-difficulty questions.

INTRODUCTION

Artificial intelligence (AI) refers to the use of technology to mimic intelligent human behavior and thought.(Amisha, Pathania, and Rathaur 2019) AI's integration in medicine has helped facilitate innovations such as robotic assistance in the operating room, algorithm integration in electronic medical records, and the identification of genetic predispositions to diseases (Hamet and Tremblay 2017). There has been an eruption of interest in AI in medicine with the recent introduction of deep-learning models such as ChatGPT. ChatGPT is a publicly available large language model (LLM) developed by OpenAI that is trained on massive amounts of data enabling it to respond to complex questions or prompts (Bhayana, Krishna, and Bleakney 2023; Kung, Cheatham, Medenilla, et al. 2023).

ChatGPT's ability to perform clinical tasks has been assessed through attempts to answer patient questions related to total hip arthroplasty (Mika et al. 2023), answer clinical questions in obstetrics and gynecology (Grünebaum et al. 2023), and guide urologic treatments (Zhou et al. 2023). Overall, it appears ChatGPT is able to respond to relatively high-level medical questions in the aforementioned disciplines but is unable to provide recommendations using the most up-to-date information and lacks the knowledge to navigate more nuanced clinical scenarios. This may be because ChatGPT-3.5 is only programmed with information until September 2021, though newer industry search engine partners and updated versions may help widen the scope of ChatGPT's search power (Metz and Weise 2023).

ChatGPT's ability to perform on medical examinations has also been preliminarily explored. ChatGPT has taken all three of the United States Medical Licensing Exams (USMLE) and was able to exceed or nearly reach the passing threshold of 60% for each test (Kung, Cheatham, Medenilla, et al. 2023). On specialty-specific questions, ChatGPT correctly answered over two-thirds of image-less radiology-specific board questions (Bhayana, Krishna, and Bleakney 2023) and outperformed humans on a mock objective-structured clinical examination modeled off the Royal College of Obstetricians and Gynaecologists entrance exam (Li, Kemp, Logan, et al. 2023).

Orthopedic surgeons seeking subspecialty certification in hand surgery are required to take the American Board of Orthopaedic Surgery's Surgery of the Hand Examination. The American Society for Surgery of the Hand (ASSH) supplies practice questions to help train surgeons planning to sit for this certification exam: Hand 100 Assessment Tool, Hand 100 Assessment Tool - Beginner Level, and Hand 100 Assessment Tool - Intermediate Level ("HAND 100 ASSESSMENT TOOL," n.d.). **These assessments may also be used by residents to assess their knowledge of the hand and upper extremity.** ("HAND 100 ASSESSMENT TOOL," n.d.) This study aims to evaluate ChatGPT-3.5's performance on ASSH board-style examination questions and to compare it to human test takers.

MATERIALS AND METHODS

This was an exploratory prospective study performed from May 24th, 2023, to June 25th, 2023. No human participants were included in this nor was any human data obtained; and thus, no Institutional Review Board approval was sought out.

MULTIPLE-CHOICE QUESTION SELECTION

All questions from the ASSH Hand 100 Assessment Tool, Hand 100 Assessment Tool - Beginner Level, and Hand 100 Assessment Tool - Intermediate Level (publicly available on the ASSH website) were downloaded for use in this study. These assessment tools included 100, 50, and 50 questions, respectively. Questions with figures and/or videos and duplicate questions were excluded from the final analysis.

Each question was pre-categorized into one of 15 topics: "Elbow", "Wrist (distal radioulnar Joint, triangular fibrocartilage complex, Arthritis)", "Anatomy/Basic Science/Imaging", "Arthritis (including inflammatory, excluding wrist)", "Brachial Plexus (nerve trauma)", "Congenital", "Contracture/Stiffness", "Fractures/dislocations (excluding elbow)", "Bites/Infection", "Amputation", "Skin/Soft Tissue Reconstruction", "Tendon (atraumatic and traumatic)", "Masses/Tumor", "Vascular disorders, Trauma", and "Nerve (Atrau-

Table 1. Stratification of the number of questions in each category across all assessment tools

	Assessment Tool				
Category	Hand 100	Beginner Level	Intermediate Level	**Total**	**Correctly answered* (ChatGPT) (%)**
Amputation	1	0	1	2	1 (50)
Anatomy/Basic Science/ Imaging	5	2	3	10	4 (40)
Arthritis	3	3	3	9	5 (55.6)
Bites/Infection	4	2	3	9	4 (44.4)
Brachial Plexus	2	2	3	7	2 (28.6)
Congenital	3	0	2	5	1 (20)
Contracture/Stiffness	2	1	1	4	2 (50)
Elbow	4	4	1	9	0 (0)
Fractures/dislocations	3	5	3	11	5 (45.5)
Masses/Tumor	1	1	1	3	3 (100)
Nerve	8	3	4	15	9 (60)
Skin/Soft Tissue Reconstruction	3	2	0	5	3 (60)
Tendon	6	6	5	17	6 (35.3)
Vascular disorders, Trauma	3	1	4	8	3 (37.5)
Wrist	1	0	2	3	2 (66.7)
Total Questions	49	32	36	**117**	50 (42.7)
Correctly answered* (ChatGPT) (%)	20 (40.82)	16 (50.0)	14† (38.89)	50 (42.7)	

*On the first try
†Indicates P<0.05

matic/compressive)". A full breakdown of the number of questions in each category is shown in Table 1.

CHATGPT AND DATA COLLECTION

ChatGPT – 3.5 (version May 24th, 2023; OpenAI) was used to answer the questions, with no prior pre-training of the software with hand surgery content. Each multiple-choice question, along with the answer choices, was entered into ChatGPT using a standardized format (Figure 1) to ensure consistency in question presentation. The response was then regenerated two additional times to identify inconsistencies in ChatGPT's answer. To reduce memory-retention bias, a new chat session was used for each individual question. For each question and response, the following variables were recorded: 1) the question/answer choices, 2) topic of the question – as designated by the ASSH, 3) ChatGPT's selected answer/explanation, 4) if the answer changed after regeneration of the response, 5) whether ChatGPT identified the correct answer, and 6) the percentage of human respondents who selected the correct answer (when applicable).

What is the correct answer to the following question:

[COPY QUESTION HERE]

Answer choices:
[A]
[B]
[C]
[D]
[E]

Figure 1: Standardized format for questions inputted into ChatGPT

STATISTICAL ANALYSIS

Overall performance of ChatGPT on all three assessment tools was calculated, and T-tests were used to compare significant differences between human and ChatGPT response accuracy using Statistical Package for Social Sciences (Version 29.0.0.0). The significance level was set at a *P* value of <0.05.

RESULTS

Overall, 117 questions from the 3 assessment tools were included in the final analysis. Of the 117 questions, 49 were from the ASSH Hand 100 Assessment Tool, 32 were from the Beginner Assessment Tool, and 36 were from the Intermediate Assessment Tool (Figure 2). On the first try, ChatGPT correctly answered 40.82% (20/49), 50.0% (16/32),

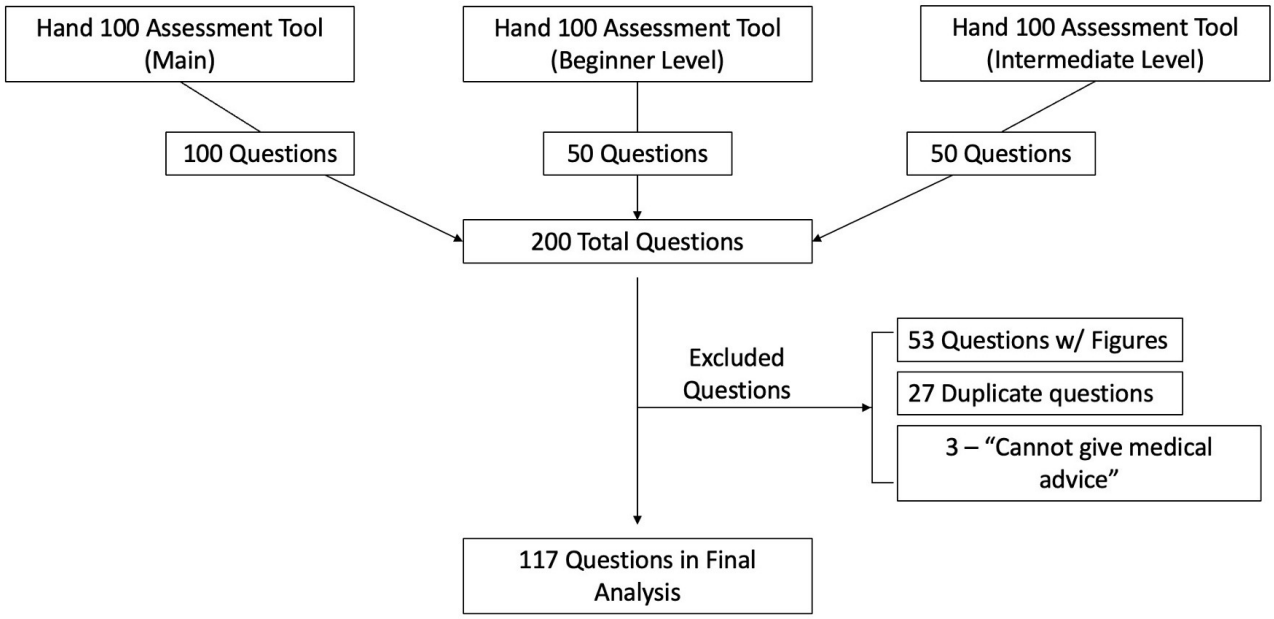

Figure 2: Flowchart of questions included in the final analysis

38.89% (14/36) of questions, respectively. Overall, ChatGPT correctly answered 50/117 (42.7%) of the questions on the first try. Upon regeneration of the response, ChatGPT changed its response to 40 (34.2%) of the 117 questions. Of these 40 questions, ChatGPT changed its response to the correct answer from an initially incorrect answer, on 10 questions.

On the Beginner and Intermediate Assessment tools, humans correctly answered 56.64% and 62.73% of questions, respectively (Figure 3), as opposed to ChatGPT's 50.0% and 38.89%, respectively. Humans performed significantly better than ChatGPT on the Intermediate assessment tool (P=0.034). There was no significant difference between ChatGPT and human performance on the Beginner Assessment tool. ChatGPT excelled in topics (>60% correct) of mass/tumor, nerve, wrist, and performed poorly (<40% correct) on topics regarding anatomy/basic science/imaging, brachial plexus, congenital, elbow, tendon, and vascular disorders, trauma.

DISCUSSION

With the recent emergence of OpenAI's LLM, ChatGPT, the use of AI in medicine has become an area of increasing interest. ChatGPT's performance on various medically-related exams has been evaluated, however, to our knowledge, no studies have examined ChatGPT-3.5's ability to answer questions from the publicly-available Hand 100, Beginner, and Intermediate assessment tools. The current study compared ChatGPT's performance on these assessment tools to human test takers. The present study reveals that while ChatGPT was able to correctly answer simpler hand surgery questions approximately 40% of the time, it struggled to answer higher-level questions, seen through its significantly worse performance on the intermediate-level question set than human test takers.

When comparing ChatGPT's performance on hand surgery questions to its performance on other orthopedic examinations, a trend emerged of ChatGPT's inability to achieve passing scores on orthopedic exams. In a study which sought to report ChatGPT's performance on the American Board of Orthopedic Surgery's (ABOS) In-Training Examination (OITE), Lum reported that ChatGPT correctly answered 47% of the 207 included examination questions. ChatGPT scored in the first percentile compared to postgraduate year (PGY) 3s, PGY4s, and PGY5s, which made it unlikely to pass the ABOS' written board examination (Lum 2023). Furthermore, when assessing ChatGPT's performance on the Orthopaedic Fellowship of the Royal College of Surgeons (FRCS) Part A exam in the United Kingdom, Saad et al. reported that ChatGPT did not achieve a passing score, correctly answering 67.5% of the 240 included questions (Saad et al. 2023). Though ChatGPT did not pass the FRCS Orthopaedic Part A exam, the authors noted that ChatGPT demonstrated better proficiency in answering "anatomy-based" questions and shorter questions. This semi-parallels the results of the Lum study, as they reported that on the ABOS examination, ChatGPT performed significantly better on "recognition and recall" and "comprehension and interpretation questions'' when compared to questions "about the application of new knowledge." In our study, ChatGPT performed better on "Beginner" level questions and performed variably across categories that we suspect have more "recognition and recall" style questions such as "Masses/Tumors" and "Anatomy/Basic Science/Imaging." However, in our study, no formal analysis stratifying the types of questions was performed, making it difficult to assess similar trends. Overall, ChatGPT performed similarly when compared to accuracy on other Or-

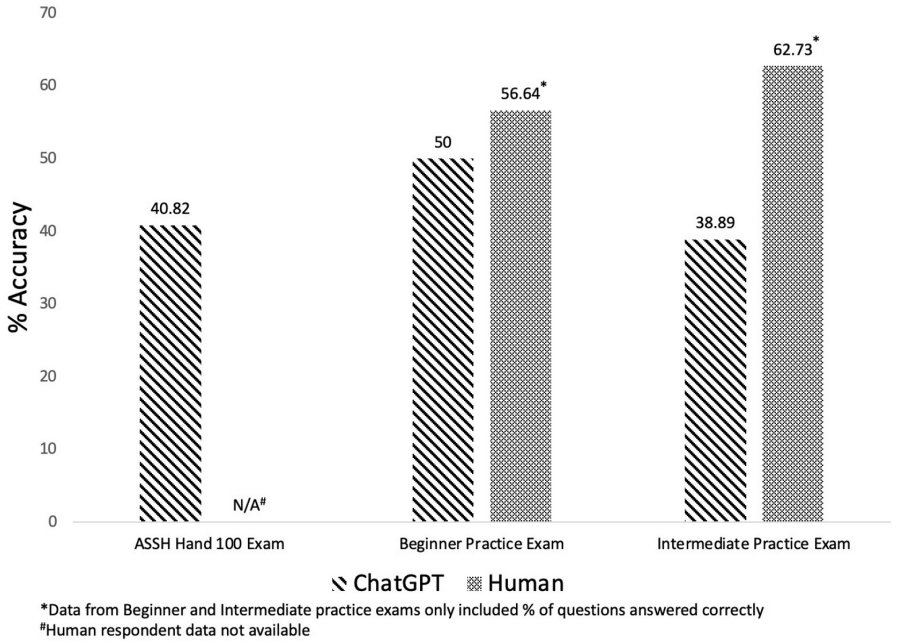

Figure 3: ChatGPT versus human response accuracy across all three question sets

thopedic examinations, both in the United States and in Europe.

To compare ChatGPT's performance on orthopedic-specific exams to other medical examinations, we sought out similar studies using the same version of ChatGPT (3.5 version) to assess performance on multiple-choice questions. Kung et al. assessed ChatGPT-3.5's performance on 350 questions from the USMLE board exams - Step 1, Step 2, and Step 3. ChatGPT answered 75%, 61.5%, and 68.8% of questions correctly, respectively. ChatGPT obtained at or above a passing score for all USMLE exams (60% threshold) (Kung, Cheatham, Medenilla, et al. 2023). Similarly, on 150 questions generated to match the Canadian Royal College and American Board of Radiology exams, ChatGPT obtained a "near pass" at 69% (Bhayana, Krishna, and Bleakney 2023). Lewandowski et al. assessed ChatGPT's performance on 358 dermatology specialty exam questions, with ChatGPT answering 74.63% correctly, well above the pass rate of 60% (Lewandowski et al. 2023). Additionally, ChatGPT achieved a 90.48% correct on 105 thoracic surgery exam questions, 7% above human students (Gencer and Aydin 2023). In conclusion, while ChatGPT performed similarly on the orthopedic exams discussed in the previous paragraph, that trend did not hold true when compared to its performance on other specialty assessments. ChatGPT-3.5 performed at least 18.8% worse and at maximum 47.8% worse on the orthopedic hand surgery questions than the non-orthopedic exams discussed above.

ChatGPT 3.5's relatively poor performance on hand surgery board-style questions may be explained by the source from which it acquires its information. ChatGPT-3.5 is not a search engine. Prior to the development of recent "Plugins", which are additional software tools to "help ChatGPT access up-to-date information, run computations, or use third-party services" ("ChatGPT Plugins" 2023) and the release of ChatGPT-4, ChatGPT did not actively search the Internet when a prompt was entered. Rather, it was trained using "Reinforcement Learning from Human Feedback" ("Introducing ChatGPT" 2023). Essentially, ChatGPT-3.5 was trained on massive amounts of information through 2021. Though we are not aware of a method to quantify this, using the number of examiners as a proxy, we broadly estimate that there is more information regarding the USMLE, and to some extent, the OITE, in comparison to a subspecialty examination, such as the hand surgery assessment questions. From 2021- 2022, over 29,000 people took Step 1 and Step 2, and over 22,000 people took Step 3 from United States and Canadian medical schools ("USMLE Administration, Minimum Passing Scores, and Performance" 2023). In comparison, in 2023, there were 188 hand surgery fellowship positions filled in the United States ("Match Results Statistics Hand Surgery - 2023: Overall Statistics" 2023). Though it is not a requirement to practice hand surgery, we assume that most hand surgery fellows will sit for the Surgery of the Hand examination, if seeking subspecialty certification. Thus, as the content of the questions becomes more subspecialized, it may become more difficult to find straightforward answers as there is also a relatively smaller pool of consensus creating the *correct* answer for these questions as compared to an exam with more examiners, such as the USMLE.

This study assessed ChatGPT-3.5's performance on hand surgery board-style examination questions. It is our hope that the information gathered in this study can guide the use of AI in the training of hand surgeons. This study has several limitations. ChatGPT-3.5 only includes data published up to 2021 and therefore, if the questions asked about recently published data, ChatGPT-3.5 may not have

been equipped to answer them as well as if its database had included data published more recently. Additionally, there was no formal analysis done to stratify the types of questions used by this study, therefore limiting the ability to assess trends across other similar studies. There was also an exclusion of image-based questions in this study, as ChatGPT-3.5 does not have the capacity for these questions. This limits the ability to fully compare ChatGPT's performance to that of a human test taker. **Lastly, it is of important note that no pre-training of ChatGPT-3.5 was conducted. Our intent was to match the actions of a standard user, testing ChatGPT's direct knowledge. Many studies investigating artificial intelligence's clinical use in orthopedics have pre-trained machine learning models with training data sets of their intended pathology/fracture of focus.** (Groot, Bongers, Ogink, et al. 2020) **Pre-training with hand surgery content may impact ChatGPT's accuracy on the Hand 100, Beginner, and Intermediate assessment tools and remains an area of future study.**

The results of this study further emphasize the potential of artificial intelligence in the field of orthopedics, along with medicine at large. Further investigation includes expanding to ChatGPT-4 or other updated versions and using free-form questions. Multiple studies have assessed ChatGPT's performance on medical exams using ChatGPT-4, a more advanced LLM, with findings to suggest superior performance when compared to the 3.5 version ("GPT-4," n.d.). Lewandowski et al. and Ali et al. found that ChatGPT- 4 significantly outperformed version 3.5 on dermatology board exam questions and neurosurgery board exam questions, respectively (Lewandowski et al. 2023; Ali, Tang, Connolly, et al. 2023). Future studies would benefit from using both version 3.5 and 4 to accurately assess the performance of this AI tool and to investigate if the trend of version 4 outperforming version 3.5 is consistent across specialties. Additionally, studies have used free-form questions, rather than multiple choice questions, to assess ChatGPT's response when it does not have answer options. Zhou et al. assessed ChatGPT-3.5's performance on 30 urologic questions generated by the research team; ChatGPT was deemed "correct" if it answered in accordance with medical guidelines (Zhou et al. 2023).

CONCLUSIONS

In conclusion, ChatGPT may be a versatile tool for medical education; however, it performed poorly on hand-surgery board-style examination questions, and its current performance did not match the level of human respondents. However, it is reasonable to assume that ChatGPT's ability to answer more complex hand surgery will improve with advanced iterations of software.

Submitted: January 28, 2024 EDT, Accepted: June 09, 2024 EDT

REFERENCES

Ali, R., O. Y. Tang, I. D. Connolly, et al. 2023. "Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations." *Neurosurgery*, August, 21. https://doi.org/10.1227/neu.0000000000002632.

Amisha, M P, M. Pathania, and V. Rathaur. 2019. "Overview of Artificial Intelligence in Medicine." *J Family Med Prim Care* 8 (7): 2328. https://doi.org/10.4103/jfmpc.jfmpc_440_19.

Bhayana, R., S. Krishna, and R.R. Bleakney. 2023. "Performance of ChatGPT on a Radiology Board-Style Examination: Insights into Current Strengths and Limitations." *Radiology* 307 (5): e230582. https://doi.org/10.1148/radiol.230582.

"ChatGPT Plugins." 2023. OpenAI. September 11, 2023. https://openai.com/blog/chatgpt-plugins.

Gencer, A., and S. Aydin. 2023. "Can ChatGPT Pass the Thoracic Surgery Exam?" *The American Journal of the Medical Sciences* 366 (4): 291–95. https://doi.org/10.1016/j.amjms.2023.08.001.

"GPT-4." n.d. OpenAI. https://openai.com/research/gpt-4.

Groot, O. Q., M. E. R. Bongers, P. T. Ogink, et al. 2020. "Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? A Systematic Review." *Clin Orthop Relat Res* 478 (12): 2751–64. https://doi.org/10.1097/CORR.0000000000001360.

Grünebaum, A., J. Chervenak, S. L. Pollet, A. Katz, and F. A. Chervenak. 2023. "The Exciting Potential for ChatGPT in Obstetrics and Gynecology." *American Journal of Obstetrics and Gynecology* 228 (6): 696–705. https://doi.org/10.1016/j.ajog.2023.03.009.

Hamet, P., and J. Tremblay. 2017. "Artificial Intelligence in Medicine." *Metabolism* 69:S36–40. https://doi.org/10.1016/j.metabol.2017.01.011.

"HAND 100 ASSESSMENT TOOL." n.d. https://www.assh.org/s/hand-100-assessment-tool.

"Introducing ChatGPT." 2023. OpenAI. September 11, 2023. https://openai.com/blog/chatgpt#OpenAI.

Kung, T. H., M. Cheatham, A. Medenilla, et al. 2023. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." Edited by A. Dagan. *PLOS Digit Health* 2 (2): e0000198. https://doi.org/10.1371/journal.pdig.0000198.

Lewandowski, M., P. Łukowicz, D. Świetlik, and W. Barańska-Rybak. 2023. "ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level Based on the Specialty Certificate Examination in Dermatology." *Clinical and Experimental Dermatology*, August, llad255. https://doi.org/10.1093/ced/llad255.

Li, S.W., M.W. Kemp, S.J.S. Logan, et al. 2023. "ChatGPT Outscored Human Candidates in a Virtual Objective Structured Clinical Examination in Obstetrics and Gynecology." *American Journal of Obstetrics and Gynecology* 229 (2): 172.e1-172.e12. https://doi.org/10.1016/j.ajog.2023.04.020.

Lum, Z. C. 2023. "Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT." *Clin Orthop Relat Res* 481 (8): 1623–30. https://doi.org/10.1097/CORR.0000000000002704.

"Match Results Statistics Hand Surgery - 2023: Overall Statistics." 2023. National Resident Matching Program. September 11, 2023. https://www.nrmp.org/wp-content/uploads/2023/05/Hand-Surgery-MRS-Report-2023.pdf.

Metz, C., and K. Weise. 2023. "Microsoft to Invest $10 Billion in OpenAI, the Creator of ChatGPT." New York Times Company. January 23, 2023. https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html.

Mika , A.P., J.R. Martin, S.M. Engstrom, G.G. Polkowski, and J.M. Wilson. 2023. "Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty." *Journal of Bone and Joint Surgery*, July, 5. https://doi.org/10.2106/JBJS.23.00209.

Saad, A., K.P. Iyengar, V. Kurisunkal, and R. Botchu. 2023. "Assessing ChatGPT's Ability to Pass the FRCS Orthopaedic Part A Exam: A Critical Analysis." *The Surgeon* 21 (5): 263–66. https://doi.org/10.1016/j.surge.2023.07.001.

"USMLE Administration, Minimum Passing Scores, and Performance." 2023. United States Medical Licensing Exam. September 11, 2023. https://www.usmle.org/performance-data.

Zhou, Z., X. Wang, X. Li, and L. Liao. 2023. "Is ChatGPT an Evidence-Based Doctor?" *European Urology* 84 (3): 355–56. https://doi.org/10.1016/j.eururo.2023.03.037.