J. Phys. Ther. Sci. 36: 234–239, 2024

The Journal of Physical Therapy Science

Original Article

Evaluation of the accuracy of ChatGPT's responses to and references for clinical questions in physical therapy

Shogo Sawamura, RPT, PhD[1)*], Takanobu Bito, RPT, MSc[2)],
Takahiro Ando, RPT, MSc[2)], Kento Masuda, RPT, MSc[2)],
Sakiko Kameyama, RPT, MSc[1)], Hiroyasu Ishida, RPT, MSc[1)]

[1)] *Department of Rehabilitation, Heisei College of Health Sciences: 180 Kurono, Gifu City, Gifu 501-1131, Japan*
[2)] *Department of Rehabilitation, Gifu University Hospital, Japan*

Abstract. [Purpose] This study evaluated the accuracy of ChatGPT's responses to and references for five clinical questions in physical therapy based on the *Physical Therapy Guidelines* and assessed this language model's potential as a tool for supporting clinical decision-making in the rehabilitation field. [Participants and Methods] Five clinical questions from the "Stroke", "Musculoskeletal disorders", and "Internal disorders" sections of the *Physical Therapy Guidelines*, released by the Japanese Society of Physical Therapy, were presented to ChatGPT. ChatGPT was instructed to provide responses in Japanese accompanied by references such as PubMed IDs or digital object identifiers. The accuracy of the generated content and references was evaluated by two assessors with expertise in their respective sections by using a 4-point scale, and comments were provided for point deductions. The inter-rater agreement was evaluated using weighted kappa coefficients. [Results] ChatGPT demonstrated adequate accuracy in generating content for clinical questions in physical therapy. However, the accuracy of the references was poor, with a significant number of references being non-existent or misinterpreted. [Conclusion] ChatGPT has limitations in reference selection and reliability. While ChatGPT can offer accurate responses to clinical questions in physical therapy, it should be used with caution because it is not a completely reliable model.

Key words: Artificial intelligence, Language model, Physical therapy

(This article was submitted Dec. 12, 2023, and was accepted Jan. 29, 2024)

INTRODUCTION

ChatGPT, launched in November 2022, is an artificial intelligence (AI)-based large-scale language model (LLM) trained on a large set of multilingual text data. It can generate human-like responses to text inputs[1)]. The Generative Pretrained Transformer (GPT) architecture uses neural networks to process natural language and generate responses based on the context of the input text[2)].

Globally, ChatGPT has numerous users due to its simplicity of use[3)]. LLMs can perform various tasks with high levels of capability and output text ability by predicting the next word with references to past text[2, 4)]. They can be utilized in diverse areas, and their use in various additional fields is under consideration. In the medical field, ChatGPT is capable of answering questions on general medical knowledge and different diagnoses with a certain degree of accuracy, aiding clinical decision-making[5)]. For example, ChatGPT could be beneficial for collecting COVID-19 vaccination information, given its concise and neutral output content[6)]. Moreover, it can be favorably used in various medical application scenarios, such as the creation of

*Corresponding author: Shogo Sawamura (E-mail: sawarm123@yahoo.co.jp) radiology reports[7, 8], laboratory reports[9], and discharge summaries[10], as well as in medical education[11, 12]. Hence, ChatGPT can potentially be used in the medical field as an adjunct tool to improve work efficiency[13–15]. Further, ChatGPT can be used to provide patients with medical information[16] and to compose medical texts and academic articles[17–20].

However, ChatGPT can also provide inaccurate or biased outputs, such as the citation of non-existent article references or the perpetuation of sexist stereotypes[1]. Reports on the possibility of spreading misinformation[21, 22], as well as "hallucinations"[23–28], which refer to the phenomenon of responding to incorrect information as if it were correct, further indicate the dangers of readily trusting ChatGPT outputs. While both healthcare professionals and patients have reportedly been misled by its outputs[6, 27], some studies[29] have indicated that medical professionals and patients can use ChatGPT to diagnose symptoms. Therefore, an appropriate understanding of ChatGPT's performance and characteristics is required before use.

In the rehabilitation field, LLMs, such as ChatGPT, can be used in patient care and rehabilitation planning to provide useful information and advice on rehabilitation based on patients' medical information, including evaluation item enumeration, goal setting, and clinical reassignment. However, to the best of our knowledge, no research has explored the use of ChatGPT in the rehabilitation field. Therefore, this study examines the accuracy of ChatGPT's responses to clinical questions related to the rehabilitation field of physical therapy. By conducting this verification, we aimed to establish basic evidence to support the use of ChatGPT in this field.

MATERIALS AND METHODS

This study randomly extracted five clinical questions from the sections titled "Stroke", "Musculoskeletal disorders", and "Internal disorders" in the *Physical Therapy Guidelines* released by the Japanese Society of Physical Therapy[30] (Table 1). We inputted the five clinical questions for each of the three sections into ChatGPT (free version) in Japanese, as most Japanese users cannot read other languages. To standardize the quality of the responses, we set the following prompt: "I will now ask questions related to physical therapy. Please provide answers in Japanese, accompanied by references, such as PubMed IDs (PMIDs) or digital object identifiers (DOIs)". Data were collected on April 3, 2023.

Subsequently, the accuracy of the generated output and references was evaluated by six raters, two assessors for each section, who referred to the content of the *Physical Therapy Guidelines*[30]. The evaluation was conducted using a four-point scale (4=completely correct, 3=almost correct, 2=partially correct, and 1=completely correct) based on earlier research[6]. The evaluators discussed the scoring method among themselves in advance, with reference to previous studies.

The assessors provided comments whenever deductions were made. They had more than five years of clinical experience in their respective sections and held master's degrees or higher. Weighted kappa coefficients were calculated to assess their inter-rater agreement.

To evaluate the references, searches were performed using the appended PMIDs and DOIs and their assessed contents. In cases where the title of a referenced article differed from that of the search results using PMIDs or DOIs, another search was

Table 1. Clinical questions extracted from the *Physical Therapy Guidelines*

Stroke	Musculoskeletal disorders	Internal disorders
1. Is balance practice effective for stroke patients?	1. Can a combination of physical therapy and medication be recommended for patients with inflammatory phase periarthritis of the shoulder?	1. Can exercise therapy be recommended for large vessel disease (true aortic aneurysm and aortic dissection) for which surgery is not indicated?
2. Is aerobic exercise effective for patients that have had stroke?	2. Can stretching of wrist extensors be recommended for patients with lateral epicondylitis of the humerus?	2. Is home exercise therapy recommended for patients with chronic heart failure?
3. Is treadmill training (treadmill only, combined weight bearing, split belt) useful for patients that have had a stroke and have been left with gait disturbance?	3. Which is more effective in preventing the onset and progression of hip osteoarthritis: single or combined exercise therapy with physical therapy and lifestyle guidance?	3. Is abdominal and mouth-to-mouth breathing recommended for patients with stable chronic obstructive pulmonary disease?
4. Is lower extremity orthotic therapy (long and short leg orthoses) useful for patients that have had stroke and have been left with gait disturbance?	4. Is physical therapy recommended for patients with mild (Kellgren–Lawrence classification 1 or 2) knee osteoarthritis and decreased mobility?	4. Is limb muscle training recommended for patients with stable chronic obstructive pulmonary disease?
5. Is home-based physiotherapy or remote training useful for patients that have had a stroke?	5. Is strength training or a range of motion exercises recommended in physical therapy for patients prior to anterior cruciate ligament reconstruction surgery?	5. Is respiratory muscle training recommended for patients with chronic stable interstitial lung disease?

Reproduced from *Physical Therapy Guidelines* 2nd edition 2021 with permission from IGAKU-SHOIN Ltd., Copyright (2023).

conducted using the article title to evaluate the content. If an article was not found even after searching by title, PMIDs, or DOIs, it was termed a “fictional paper”.

Data on the accuracy of the content and references provided by ChatGPT were expressed as median with interquartile range (IQR). The weighted kappa values were classified as follows: <0.40 for poor strength of agreement, 0.41–0.60 for moderate agreement, 0.61–0.80 for good agreement, and 0.81–1.00 for excellent agreement[31]. A p-value of <0.05 was considered statically significant.

RESULTS

The results of our analysis are presented in Table 2. In all sections, the scores for output content and reference accuracy were 3.0 [2.25–4.00] and 1.00 [1.00–3.75] points, respectively. Among the appended PMIDs and DOIs, 18.9% (7/37) exhibited concordance with the output. Further, 40.5% (15/37) of the output references were identified as fictitious papers.

In the stroke section, 16 references were output for the five clinical questions. Accuracy of output content was scored 4.00 [3.00–4.00] by assessor 1 and 3.00 [3.00–4.00] by assessor 2. Accuracy of references was scored 1.00 [1.00–2.50] by assessor 1 and 1.00 [1.00–2.25] by assessor 2. The weighted kappa coefficient for agreement among assessors was 0.84 for accuracy and 0.98 for references. Among the appended PMIDs and DOIs, 37.5% (6/16) were consistent with the output. Further, 37.5% (6/16) of the output references were fictitious papers. The reasons for point reductions were “The article stated that an effect not mentioned in the *Guidelines* was effective”, and “The interventions and outcomes in the references differed from those in the questions”.

In the musculoskeletal disorders section, 11 references were output for the five clinical questions. Accuracy output content was scored 3.00 [3.00–4.00] and 3.50 [3.00–4.00] by assessors 3 and 4, respectively. Accuracy of references was scored 1.00 [1.00–1.00] by assessor 3 and 1.00 [1.00–1.50] by assessor 4. The weighted kappa coefficient for agreement among assessors was 0.76 for accuracy and 0.93 for references. Among the appended PMIDs and DOIs, 0.0% (0/11) were consistent with the output. Among the output references, 45.5% (5/11) were fictitious papers. The reasons for point reductions were “The article stated that an effect not mentioned in the *Guidelines* was effective”, “The recommendation differed from that in the *Guidelines*”, and “The interventions and outcomes in the reference differed from those in the questions”.

In internal disorders section, 11 references were output for the five clinical questions. Accuracy of output content was scored 3.00 [2.00–3.00] and 2.00 [2.00–3.00] by assessors 5 and 6, respectively. Accuracy of references was scored 1.50 [1.00–4.00] by assessor 5 and 1.00 [1.00–4.00] by assessor 6. The weighted kappa coefficient for agreement among assessors was 0.84 for accuracy and 0.98 for references. Among the appended PMIDs and DOIs, 9.1% (1/11) were consistent with the output. Among the output references, 36.7% (4/11) were fictitious papers. The reasons for point reductions were “The output did not include an effect described in the *Guidelines*”, “The output stated that an effect not mentioned in the *Guidelines* was effective”, and “The interventions and outcomes in the references differed from those in the questions”.

DISCUSSION

This study inputted five clinical questions based on the *Physical Therapy Guidelines*[30] to ChatGPT and evaluated the accuracy of the generated responses and references. The results revealed accurate output content but inaccurate output reference. Therefore, ChatGPT’s ability to generate accurate responses and references is unreliable.

Table 2. Summary of the scores for output content, reference accuracy and weighted kappa coefficient

	Stroke		Musculoskeletal disorders		Internal disorders	
	Assessor 1	Assessor 2	Assessor 3	Assessor 4	Assessor 5	Assessor 6
Accuracy of the generated output	4.00 [3.00–4.00]	3.00 [3.00–4.00]	3.00 [3.00–4.00]	3.5 [3.00–4.00]	3.00 [2.00–3.00]	2.00 [2.00–3.00]
Weighted kappa coefficient of the generated output	0.84* 95% CI: 0.51–1.00		0.76* 95% CI: 0.47–1.00		0.84* 95% CI: 0.51–1.00	
Accuracy of the generated references	1.00 [1.00–2.50]	1.00 [1.00–2.25]	1.00 [1.00–1.00]	1.00 [1.00–1.50]	1.50 [1.00–4.00]	1.00 [1.00–4.00]
Weighted kappa coefficient of the generated references	0.98* 95% CI: 0.94–1.00		0.93* 95% CI: 0.82–1.00		0.98* 95% CI: 0.93–1.00	
Agreement rate between PMID/DOI and the article title	37.5% (6/16)		0.0% (0/11)		9.1% (1/11)	
Rate of fictitious papers	37.5% (6/16)		45.5% (5/11)		36.7% (4/11)	

Median [first quartile–third quartile].
*p<0.05.
95% CI: 95% confidence interval; PMID: PubMed ID; DOI: digital object identifier.

All weighted kappa coefficients exceeded 0.61, indicating good inter-rater agreement. Considering each rater's educational background and clinical experience, the quality of the ratings provided herein was deemed valid.

The accuracy of the output content was relatively good, with an overall median of 3; however, there were some variations between the sections. There are reports of ChatGPT passing the U.S. National Medical Examination[32] and answering clinical reasoning questions with high accuracy[33], which agrees with our result that ChatGPT's knowledge of outputs is high. Further, Johnson et al.[16] reported that ChatGPT can provide accurate information on common cancer myths and misconceptions; similarly, the system can generate correct answers to questions requiring higher-order thinking in medical biochemistry[34] and can provide factually accurate, contextually relevant and structured answers to complex and evolving clinical questions[35]. There are many other positive reports[16, 35–38] that indicate that ChatGPT's response generation ability has a certain degree of accuracy. These findings further support our study's results.

Conversely, a review of the deduction of points by this study's assessors revealed the following issues: "The output content included an effect that was not mentioned in the *Guidelines* and stated it was effective", and "The output content did not include necessary information that was mentioned in the *Guidelines*". Previous studies have found that ChatGPT answered questions as if they were correct, despite having insufficient evidence[24, 26], and that it gave only superficial answers to highly specialized questions[8, 39]. This suggests that ChatGPT is incapable of generating completely correct output. This may be due to the lack of specialized data[40], influence of biases in the training dataset[27, 41, 42], and presence of hallucinations that cause responses of incorrect information as if they were correct, or the automatic generation of fictitious information[28, 43].

This study also found significant problems with the accuracy of the generated references. Specifically, 40.5% of all cases cited non-existent articles, and there were also cases where existing articles were referenced but interpreted incorrectly; similar findings have been reported[24, 26, 28, 41, 44, 45]. Accordingly, ChatGPT is unable to adequately generate reliable information during reference selection.

Some studies have also claimed that ChatGPT's generation ability is based on the information contained in the training dataset that is dated until September 2021, which affects the accuracy of the information generation and reference selection[1], and also leads to hallucinations[28]. However, as the *Physical Therapy Guidelines*[30], from which this study extracted its clinical questions, was based on studies published before September 2021, one would assume that its results were not affected by the problem regarding ChatGPT's data learning period. Conversely, ChatGPT's training dataset includes public information and research articles alongside general information, such as that from Wikipedia and Blogs, which may be subject to various biases. Further, the sequence of steps involved in citing references is the following: properly interpreting the question, searching for information based on the interpretation, correctly interpreting relevant articles, and citing them. In other words, citing references is considered a more challenging process than simply answering questions. This might have caused the insufficient accuracy of the reference output generated herein.

In this study, input to ChatGPT was in Japanese, and the responses were also output in Japanese. ChatGPT is trained primarily in English, and the proportion of non-English language training data is small[1]. Therefore, results similar to this study's may be obtained for other languages with a lower percentage of representation in the training data. The results of this study have important implications for non-English-speaking physical therapists and also the general public interested in rehabilitation.

As mentioned earlier, ChatGPT can output responses to physical therapy questions with a certain degree of accuracy; however, it is not a completely reliable model. Therefore, it can be used as a support tool but cannot replace professionals' specialized knowledge and experience for use in the medical field[5, 37, 46]. Accordingly, one should not blindly follow the answers provided by ChatGPT or be overdependent on them.

This study had several limitations. First, in this study, clinical questions were selected randomly from the Stroke, Musculoskeletal disorders, and Internal disorders sections; however, we have not been able to examine whether they adequately represent all areas of physical therapy. Further research needs to be conducted on a larger scale, including a comprehensive survey of the clinical questions included in the guidelines. Second, as the questions were asked in Japanese, the output's accuracy was likely limited by linguistic issues. Additionally, we used the free version of ChatGPT; therefore, it is unclear whether we would receive different results with the paid version, ChatGPT-4. The paid version, ChatGPT-4, is reported to have better capabilities to handle different languages and generate responses than the free version, 3.5[1]. Therefore, it is highly possible that ChatGPT-4 will improve the accuracy of output. In addition, the current implementation of the browsing feature is also likely to yield better results for reference extraction[1]. Therefore, further studies, such as a comparison between versions 3.5 and 4 in the latest browser, is also required. Third, the four-point scale used in this study was based on earlier research[6]. This is a rough evaluation of the accuracy of the output, and is likely to be insufficient to provide a detailed evaluation of the output content. In the future, more detailed evaluation scales need to be established and studied. In addition, as the scoring is based on the subjective judgment of the evaluator, the possibility of bias cannot be ruled out. In the future, it will be necessary to use methods that ensure objectivity, such as the formulation of clear scoring rules. Moreover, the present study only performed a statistical analysis of the agreement among the ratings. Therefore, an evaluation regarding the quality of responses, such as the complexity and depth of answers, is lacking. Future studies regarding the complexity and depth of the output content using statistical methods are required. Finally, ChatGPT has a parameter (temperature) that causes "fluctuations" in its answers; hence, the same trial may not necessarily yield the same results. Therefore, further research should be conducted on various models, including ChatGPT-4. Moreover, questions based on more clinical patient information, rather than on guiding clinical questions, should be asked, and their accuracy should be studied.

This study evaluated the accuracy of ChatGPT's responses to and references for five clinical questions based on the Japanese *Physical Therapy Guidelines*[30]. ChatGPT demonstrated a high degree of accuracy for the response content but insufficient reference accuracy. Therefore, ChatGPT is not a completely reliable model and should be used with caution. However, as it is easy to use and can generate clinical questions that have a certain level of accuracy, it can be a useful tool to support clinical decision-making. ChatGPT requires further improvements to enhance the accuracy of generated references and mitigate the biases and limitations inherent in training datasets.

Funding

This study was funded by the Reiwa 5th Year President's Recommended Research Activity Grant of Heisei College of Health Sciences. The Heisei College of Health Sciences had no role in the design or conduct of the study, interpretation of the data, or preparation and approval of the manuscript.

Conflicts of interest

None.

ACKNOWLEDGMENT

We thank Mr. Kengo Kohiyama for his useful discussions.

REFERENCES

1) Open AI: API. Open AI. https://platform.openai.com (Accessed Jun. 1, 2023)
2) Brown TB, Mann B, Ryder N, et al.: Language models are few-shot learners. Adv Neural Inf Process, 2020, 33: 1877–1901.
3) Thorp HH: ChatGPT is fun, but not an author. Science, 2023, 379: 313. [Medline] [CrossRef]
4) Wei J, Tay Y, Bommasani R, et al.: Emergent abilities of large language models. Transactions on machine learning research. 2022. https://openreview.net/forum?id=yzkSU5zdwD (Accessed Jun. 4, 2023)
5) Rao A, Pang M, Kim J, et al.: Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv, 2023. [CrossRef]
6) Sallam M, Salim NA, Al-Tammemi AB, et al.: ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online searches for information. Cureus, 2023, 15: e35029. [Medline]
7) Jeblick K, Schachtner B, Dexl J, et al.: ChatGPT makes medicine easy to walk: an exploratory case study of simplified radiological reports. Eur Radiol, 2022. [Medline]
8) Lyu Q, Tan J, Zapadka ME, et al.: Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art, 2023, 6: 9. [Medline] [CrossRef]
9) Cadamuro J, Cabitza F, Debeljak Z, et al.: Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Med, 2023, 61: 1158–1166. [Medline] [CrossRef]
10) Patel SB, Lam K: ChatGPT: the future of discharge summaries? Lancet Digit Health, 2023, 5: e107–e108. [Medline] [CrossRef]
11) Eysenbach G: The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ, 2023, 9: e46885. [Medline] [CrossRef]
12) Tsang R: Practical applications of ChatGPT in undergraduate medical education. J Med Educ Curric Dev, 2023, 10: 23821205231178449. [Medline] [CrossRef]
13) Verhoeven F, Wendling D, Prati C: ChatGPT: when artificial intelligence replaces the rheumatologist in medical writing. Ann Rheum Dis, 2023, 82: 1015–1017. [Medline] [CrossRef]
14) Zhou Z: Evaluation of ChatGPT's capabilities in medical report generation. Cureus, 2023, 15: e37589. [Medline]
15) Dave T, Athaluri SA, Singh S: ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell, 2023, 6: 1169595. [Medline] [CrossRef]
16) Johnson SB, King AJ, Warner EL, et al.: Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr, 2023, 7: pkad015. [Medline] [CrossRef]
17) Macdonald C, Adeloye D, Sheikh A, et al.: Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health, 2023, 13: 01003. [Medline] [CrossRef]
18) Fatani B: ChatGPT for future medical and dental research. Cureus, 2023, 15: e37285. [Medline]
19) Kitamura FC: ChatGPT is shaping the future of medical writing but still requires human judgment. Radiology, 2023, 307: e230171. [Medline] [CrossRef]
20) Le DP, Hall SC: Medical literature writing with ChatGPT: a rare case of choriocarcinoma syndrome with hemorrhagic brain metastases due to burned out metastatic mixed testicular cancer. Cureus, 2023, 15: e36655. [Medline]
21) Ahn C: Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation, 2023, 185: 109729. [Medline] [CrossRef]
22) De Angelis L, Baglivo F, Arzilli G, et al.: ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health, 2023, 11: 1166120. [Medline] [CrossRef]

23) Wagner MW, Ertl-Wagner BB: Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. Can Assoc Radiol J, 2024, 75: 69–73. [Medline]

24) Fijačko N, Gosak L, Štiglic G, et al.: Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation, 2023, 185: 109732. [Medline] [CrossRef]

25) Khan RA, Jawaid M, Khan AR, et al.: ChatGPT—reshaping medical education and clinical management. Pak J Med Sci, 2023, 39: 605–607. [Medline] [CrossRef]

26) Ali MJ: ChatGPT and lacrimal drainage disorders: performance and scope of improvement. Ophthalmic Plast Reconstr Surg, 2023, 39: 221–225. [Medline] [CrossRef]

27) : The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health, 2023, 5: e102. [CrossRef]

28) Athaluri SA, Manthena SV, Kesapragada VS, et al.: Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus, 2023, 15: e37432. [Medline]

29) Shahsavar Y, Choudhury A: User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors, 2023, 10: e47564. [Medline] [CrossRef]

30) Japanese Society of Physical Therapy: Physical therapy guidelines, 2nd ed. Tokyo: Igaku-Shoin, 2021. https://www.jspt.or.jp/guideline/2nd/ (Accessed Jun. 1, 2023)

31) Kundel HL, Polansky M: Measurement of observer agreement. Radiology, 2003, 228: 303–308. [Medline] [CrossRef]

32) Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health, 2023, 2: e0000198. [Medline] [CrossRef]

33) Strong E, DiGiammarino A, Weng Y, et al.: Performance of ChatGPT on free-response, clinical reasoning exams. medRxiv, 2023. [CrossRef]

34) Sabry Abdel-Messih M, Kamel Boulos MN: ChatGPT in clinical toxicology. JMIR Med Educ, 2023, 9: e46876. [Medline] [CrossRef]

35) Li SW, Kemp MW, Logan SJ, et al. National University of Singapore Obstetrics and Gynecology Artificial Intelligence (NUS OBGYN-AI) Collaborative Group: ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol, 2023, 229: 172.e1–172.e12. [Medline] [CrossRef]

36) Schulte B: Capacity of ChatGPT to identify guideline-based treatments for advanced solid tumors. Cureus, 2023, 15: e37938. [Medline]

37) Yeo YH, Samaan JS, Ng WH, et al.: Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol, 2023.

38) Sinha RK, Deb Roy A, Kumar N, et al.: Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus, 2023, 15: e35237. [Medline]

39) Hegde A, Srinivasan S, Menon G: Extraventricular neurocytoma of the posterior fossa: a case report written by ChatGPT. Cureus, 2023, 15: e35850. [Medline]

40) Giannos P, Delardas O: Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ, 2023, 9: e47737. [Medline] [CrossRef]

41) D'Amico RS, White TG, Shah HA, et al.: I asked a ChatGPT to write an editorial about how we can incorporate chatbots into neurosurgical research and patient care. Neurosurgery, 2023, 92: 663–664. [Medline] [CrossRef]

42) Sallam M: ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel), 2023, 11: 887. [Medline] [CrossRef]

43) Alkaissi H, McFarlane SI: Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus, 2023, 15: e35179. [Medline]

44) Cahan P, Treutlein B: A conversation with ChatGPT on the role of computational systems biology in stem cell research. Stem Cell Reports, 2023, 18: 1–2. [Medline] [CrossRef]

45) Huang J, Tan M: The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res, 2023, 13: 1148–1154. [Medline]

46) Liu S, Wright AP, Patterson BL, et al.: Assessing the value of ChatGPT for clinical decision support optimization. medRxiv, 2023. [CrossRef]