OBSERVATIONAL STUDY

OPEN

Clinical Research With Large Language Models Generated Writing—Clinical Research with AI-assisted Writing (CRAW) Study

Ivan A. Huespe, MD[1,2]
Jorge Echeverri, MD[3]
Aisha Khalid, MD[4]
Indalecio Carboni Bisso, MD[1]
Carlos G. Musso, PhD[1,5]
Salim Surani, MD[6,7]
Vikas Bansal, MBBS, MPH[6]
Rahul Kashyap, MD[6,8]

IMPORTANCE: The scientific community debates Generative Pre-trained Transformer (GPT)-3.5's article quality, authorship merit, originality, and ethical use in scientific writing.

OBJECTIVES: Assess GPT-3.5's ability to craft the background section of critical care clinical research questions compared to medical researchers with H-indices of 22 and 13.

DESIGN: Observational cross-sectional study.

SETTING: Researchers from 20 countries from six continents evaluated the backgrounds.

PARTICIPANTS: Researchers with a Scopus index greater than 1 were included.

MAIN OUTCOMES AND MEASURES: In this study, we generated a background section of a critical care clinical research question on "acute kidney injury in sepsis" using three different methods: researcher with H-index greater than 20, researcher with H-index greater than 10, and GPT-3.5. The three background sections were presented in a blinded survey to researchers with an H-index range between 1 and 96. First, the researchers evaluated the main components of the background using a 5-point Likert scale. Second, they were asked to identify which background was written by humans only or with large language model-generated tools.

RESULTS: A total of 80 researchers completed the survey. The median H-index was 3 (interquartile range, 1–7.25) and most (36%) researchers were from the Critical Care specialty. When compared with researchers with an H-index of 22 and 13, GPT-3.5 was marked high on the Likert scale ranking on main background components (median 4.5 vs. 3.82 vs. 3.6 vs. 4.5, respectively; $p < 0.001$). The sensitivity and specificity to detect researchers writing versus GPT-3.5 writing were poor, 22.4% and 57.6%, respectively.

CONCLUSIONS AND RELEVANCE: GPT-3.5 could create background research content indistinguishable from the writing of a medical researcher. It was marked higher compared with medical researchers with an H-index of 22 and 13 in writing the background section of a critical care clinical research question.

KEY WORDS: article writing; artificial intelligence; clinical research; Generative Pre-trained Transformer-3.5; medical research

Large language models (LLMs) are rapidly increasing in number and sophistication, providing authors with tools to improve the preparation and quality of their articles and published articles. These tools include assistance with writing, grammar, language, references, statistical analysis, and reporting standards (1). One of the most exciting developments in this field is the Generative Pre-trained Transformer (GPT)-3.5, introduced by OpenAI in late 2022 (2, 3).

DOI: 10.1097/CCE.0000000000000975

KEY POINTS

Question: Can the large language models (LLMs) Generative Pre-trained Transformer (GPT)-3.5 generate a background section of a critical care clinical research question comparable to medical researchers with an H-index of 22 and 13?

Findings: Human researchers were unable to differentiate between the background section created by GPT-3.5 and those created by their peers. Furthermore, the LLMs generated backgrounds received higher quality scores.

Meaning: The study suggests that GPT-3.5 performs comparably to medical researchers with an H-index of 22 and 13 in generating the background section. However, the tool's limitation in providing references and reliable sources restricts its suitability as a standalone tool for scientific medical writing.

The GPT-3.5 has been used in multiple applications, including chatbots, customer support, language learning models (4), medical reports (5), or even to generate reflections from critical care physicians when delivering bad news (6). Furthermore, it has the potential to become a powerful tool for scientific writing, including automated draft generation, article summarization, and language translation, which could help researchers write more efficiently and effectively (7). However, only a few studies have performed qualitative and quantitative evaluations of GPT-3.5 use in medical research, including its use in scientific writing (8), generating research questions (9), and systematic review topics (10, 11).

Salvagno et al (3) reported that GPT-3.5 appears to be a useful tool in scientific writing, assisting researchers and scientists in organizing material, generating an initial draft, and/or proofreading. However, to date, there have been no publications in the field of medicine prepared using this approach.

But, the scientific community is still debating the overall quality and authorship merit of articles generated by GPT-3.5, as well as their originality and the ethical and acceptable boundaries of using them in scientific writing (1, 12, 13). Additionally, Gao et al (14) reported that experts may not be able to recognize whether abstracts were written by GPT-3.5 or not. However, due to the limited number (only four) of researchers who participated in the study by Gao et al (14), further research is needed to draw definitive conclusions about the ability of experts to distinguish between LLMs generated and human-generated text.

Our primary aim was to investigate whether a representative sample of medical researchers can accurately differentiate between a LLMs generated (GPT-3.5) background and one generated by two expert medical researchers. Additionally, we aimed to compare the quality of the LLMs generated background with that produced by the expert researchers.

METHODS

Study Design and Objectives

This study was conducted in three phases: 1) Background development: Two medical researchers and GPT-3.5 created three backgrounds of a research project based on specific instructions outlined below. 2) Questionnaire development: The research team created a background quality questionnaire to evaluate background content and assess whether it was developed by a LLMs tool or a human research expert. 3) Survey of investigators: Researchers worldwide blindly reviewed the three backgrounds and answered the electronic questionnaire. This article adheres to Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline for cross-sectional studies and the Standards for Reporting of Diagnostic Accuracy Studies reporting guidelines (**Supplementary Tables 1** and **2**, http://links.lww.com/CCX/B255) (15, 16).

The primary aim of this study was to evaluate two outcomes. First, to assess the ability of researchers to discriminate between a research background created by a LLMs tool and created solely by human researchers. Second, to evaluate the quality of a LLMs generated research background compared with backgrounds developed by human researchers with expertise in the topic. As secondary outcomes, we also evaluated the ability of researchers with experience in the field and researchers with H-index score higher than 5 to detect LLMs generated background.

Human Backgrounds Development

To create human research backgrounds, we enlisted the help of two medical researchers who are experts in acute kidney injury (AKI) and sepsis. One of them

had a PhD with an Scopus H-index of 22 on Scopus (researcher 1), and the other had Scopus H-index of 13 (researcher 2). The researchers formulated a research question related to their expertise: "Is renal Fractional sodium excretion (FeNa) a biomarker that predicts renal failure in septic patients admitted to the ICU?" To maintain consistency and adhere to the STROBE guidelines, researchers were instructed to ensure that the background section should include the following components: 1) the importance of the research topic, 2) the current state of knowledge in this field, 3) the existing research gaps in this area, and 4) the objectives of the study.

From February 5, 2023, to February 19, 2023, researchers developed their backgrounds based on medical literature without including citations. This was done since GPT-3.5 was not able to generate reliable references despite several training preparatory tests.

GPT-3.5 Background Development

The generation of the GPT-3.5 background involved an iterative process aimed at refining the output. The final prompt consists of three essential components. First, it encompasses the research question and study design. Second, it specifies the required language characteristics (medical, scientific language) and word count (ranging from 300 to 500 words). Finally, it outlines the necessary information that must be included in the background section and the preferred order: 1) the significance of the research topic, 2) the current state of knowledge in the field, 3) the research gap that exists in this area, and 4) the study's aims. The final version of the chat log and the final prompt can be found in the **Supplementary Material** (http://links.lww.com/CCX/B255).

Questionnaire Development

The questionnaire items were derived from the STROBE statement guidelines for reporting cohort studies. The initial questionnaire was created by our research team with expertise in both clinical and research fields. The questionnaire design was refined through an iterative process involving the research team and two external researchers.

To ensure the face and content validity, readability, comprehension, and user-friendliness of the questionnaire, the research team solicited feedback from both internal and external researchers, including non-native English speakers. The research team collated and reviewed the feedback in an iterative process and made amendments to the questionnaire accordingly. The team reached a consensus on items that needed to be revised, added, reduced, or reworded.

The final questionnaire consisted of two sections. The first section assessed the quality of the background in describing the concepts that should be included in a research background based on the STROBE guidelines. This section comprised five questions, which participants rated on a Likert scale ranging from 1 (poor quality) to 5 (excellent quality). The selection of the Likert scale was based on the Review Quality Instrument for Assessing Peer Reviews of Manuscripts, as outlined by van Rooyen et al (17). The questions were as follows: 1) Does the background adequately explain the significance of the research topic? 2) Does the background adequately explain the current state of knowledge in the field? 3) Does the background adequately explain the research gap that exists in this area? and 4) Does the background adequately explain the study's aims?

The second section included the question, "Do you think this Background was created with assistance from an artificial intelligence tool?" Both sections were to be answered for each of the three backgrounds. The complete questionnaire is presented in the Supplementary Material (http://links.lww.com/CCX/B255).

The questionnaire was globally launched on February 17, 2023, and data collection continued until the required sample size was obtained on March 31, 2023. The study's target population was researchers with an academic background, as quantified by their Scopus index and experience in their domain.

Data Collection

Data collection for this study was carried out using web-based surveys that were designed using Google Forms. To disseminate the study globally, the core team used a multifaceted approach to implement the questionnaire. First, the questionnaire was disseminated through scientific societies, such as the American Societies of Critical Care and the Argentine Society of Intensive Care. Additionally, the core team individually contacted recognized researchers who have expertise in the relevant areas of study, inviting them to participate in the questionnaire. The questionnaire

required participants to disclose their Scopus H-index and submit their responses. Only researchers with a Scopus index greater than 1 were included in the final analysis. To ensure complete anonymity, the collected data set was securely stored in an Excel spreadsheet that was only accessible to the study investigators.

Statistical Analysis

In this study, categorical variables were presented using absolute numbers and proportions, whereas continuous variables were reported using the either mean (SD) or median (interquartile range [IQR]), depending on their distribution.

Regarding the primary outcomes, we evaluated the sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC) to determine the researchers' ability to identify LLMs background correctly. Additionally, we compared the average scores on a Likert scale using *t* test (18, 19). Multiple comparisons were adjusted with Bonferroni's correction. Finally, we conducted a subgroup analysis focusing on researchers with an H-index higher than 5 and prior experience in AKI research.

Even though it was a pilot study, we conducted a sample size calculation based on our primary descriptive outcome to ensure adequate statistical precision. We assumed that researchers would have a 90% sensitivity to detect a LLMs generated background. To achieve the desired level of precision with a CI ranging from 81% to 95%, a sample size of 80 participants would be necessary (CI calculated with the Clopper-Pearson exact method). Finally, we performed all statistical analyses using STATA software (Version 17.0 SE; StataCorp, College Station, TX).

Ethical Consideration

This study did not involve a clinical survey or patient participation. Each researcher who participates in the study gave their consent by means of taking the study survey, and all are included as collaborative co-authors (Supplementary Material, http://links.lww.com/CCX/B255).

RESULTS

We recruited a total of 80 researchers, with a median H-index of 3 (IQR, 1–7.25). Of these, 31.2% had prior experience in AKI research, and 32% had a master's degree, PhD, or PharmD. Also, 46% of them had experience in critical care, and English was the primary language of 35%. Detailed demographic characteristics of the researchers are presented and the country distribution of the researchers that evaluated the backgrounds is presented in the (**Supplementary Tables 3** and **4** (http://links.lww.com/CCX/B255). All participants completed the questionnairy, without missing data.

For the primary outcome, we found that the sensitivity of researchers in correctly identifying backgrounds created by GPT-3.5 was 22.6% (95% CI, 15.6–31%), while the specificity was 55.2% (95% CI, 45.7–64.4%), with an AUROC of 0.38 (95% CI, 0.33–0.45).

For the second primary outcome, the GPT-3.5 background scored significantly higher in all four evaluated components compared with backgrounds written by human researchers with H-indices of 22 and 13. Notably, the GPT-3.5 background achieved scores higher than 4 in all components (**Fig. 1** and **Table 1**).

In subgroup analyses, we found that researchers with an H-index higher than 5 had a sensitivity of 19.6% (95% CI, 9.36–33.9%), a specificity of 54% (95% CI, 39.3–68.2%), and an AUROC of 0.36 (95% CI, 0.27–0.45) for identifying the GPT-3.5 background. Researchers with prior experience in AKI research had a sensitivity of 10.5% (95% CI, 2.94–24.8%), a specificity of 43.2% (95% CI, 27.1–60.5%), and an AUROC of 0.26 (95% CI, 0.17–0.36).

DISCUSSION

We conducted a pilot study to evaluate the effectiveness of a LLMs tool (GPT-3.5) in writing academic backgrounds for research questions in critical care. Our results showed that researchers were unable to distinguish between backgrounds generated by GPT-3.5 and those written by researchers. In this sense, Gao et al (14) evaluated the ability of humans to detect LLMs abstracts and found results similar to ours, with low sensitivity and specificity for researchers detecting them. Another research performed by Levin et al (8) used also GPT-3.5 to generate full abstracts by providing only the title and result sections of the abstracts from 50 real scientific publications. They found that humans could not detect abstracts generated by LLMs (8, 14). These results are consistent with

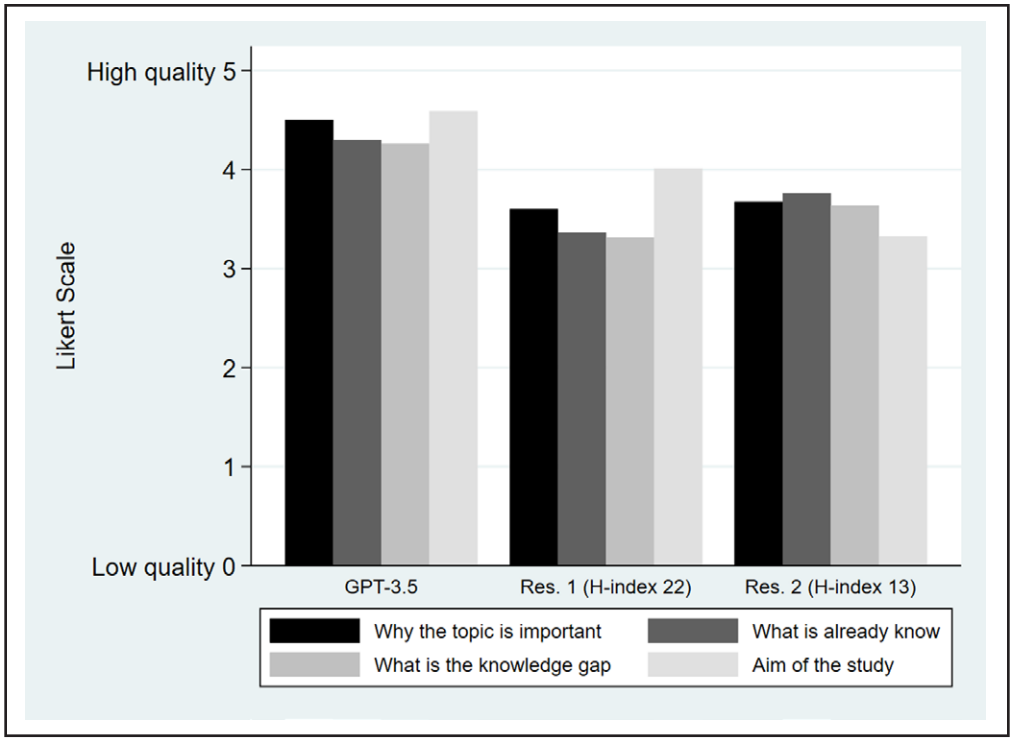

Figure 1. Evaluation of background components for Generative Pre-trained Transformer-3.5 (GPT-3.5), researcher 1 (Res. 1) (H-index 22), and researcher 2 (Res. 2) (H-index 13).

our findings that even researchers who have experience in the field and an H-index higher than 5 could not detect backgrounds generated by LLMs.

Second, we performed a structured, quantitative, and blinded comparison between the quality of a background generated with GPT-3.5 and a human expert background. In this evaluation, we found that the GPT-3.5 background consistently provided more comprehensive descriptions of the four key components recommended by the STROBE guidelines, resulting in higher scores across all components when compared with those written by human researchers. This is the first study to provide such evidence, demonstrating the potential of LLMs to improve the quality of scientific writing. GPT-3.5's ability to quickly process and analyze information is one of the key factors behind its ability to create robust scientific articles (20). In this sense, Salvagno et al (21) recently published an article in which they evaluated the summary of three studies created with GPT-3.5. They concluded that GPT-3.5 appears to be a useful tool in scientific writing, assisting researchers and scientists in organizing material, generating an initial draft, and/or proofreading (21). However, just 3 weeks after this publication, Azamfirei et al (13) observed that the GPT-3.5-generated summaries reported by Salvagno et al (21) were believable but generic, sparse in details, and contained severe mistakes and false information. Furthermore, the

TABLE 1.
Average Score of Human Backgrounds Versus Generative Pre-Trained Transformer-3.5 Background

Variable	Researcher 1 H-Index 22 ($n = 80$)	Researcher 2 H-Index 13 ($n = 80$)	GPT-3.5 Background ($n = 80$)	*p* Overall (Researcher 1 vs GPT-3.5)	*p* Overall (Researcher 2 vs GPT-3.5)
Why the topic is important	3.60 (0.99)	3.67 (1.03)	4.50 (0.73)	< 0.001	< 0.001
What is already known in the field	3.36 (1.12)	3.76 (1.08)	4.30 (0.86)	0.001	< 0.001
What is the knowledge gap	3.31 (1.13)	3.64 (1.14)	4.26 (0.87)	< 0.001	< 0.001
Aim of the study	4.01 (1.02)	3.33 (1.39)	4.59 (0.69)	< 0.001	< 0.001

GPT-3.5 = Generative Pre-trained Transformer-3.5.
After Bonferroni correction a significant *p* value should be lower than 0.006 (0.05/8 = 0.00625).

study conducted by Buholayka et al (22) examined the capability of ChatGPT in independently writing scientific case reports. Their evaluation highlighted significant limitations, as ChatGPT produced case reports with critical flaws such as incorrect diagnoses and fabricated references (22).

The research question used in this study was relatively general, allowing for the incorporation of existing knowledge and studies conducted on the topic. However, in more specific or novel research areas, the generation of false information becomes more likely as the LLMs may attempt to fill gaps with speculative or insufficiently supported data. Furthermore, it is worth noting that the citations generated by GPT-3.5 were consistently false regardless of the research question. This critical limitation raises concerns about the potential for LLMs to produce fraudulent or misleading research articles. This has significant ethical implications, particularly in scientific research where accuracy and trustworthiness are paramount (23). While GPT-4 can incorporate accurate citations into its text (if internet access is available), it still faces challenges in referencing scientific articles. Instead of citing relevant studies, it often draws on citations from public information sources which do not necessarily correspond to the content of the reference.

Regarding the prompt development, we engaged in multiple interactions with GPT-3.5, refining the generated text at each step. This iterative approach allowed us to gradually improve and shape the output to meet the desired criteria. To guide the generation of the introduction, we progressively added various components to the prompts. These components included the significance of the research topic, the current state of knowledge in the field, the research gap in the area, and the study's aims. As the prompt development progressed, we also introduced additional requirements, such as the use of scientific and medical language, as well as specifying the desired length of the introduction.

Finally, the linguistic backgrounds of the human authors in our study, including being non-native English speakers, are also important factors to consider. While these linguistics antecedents may introduce variations in writing styles and syntax patterns, it is worth noting that the authors have extensive experience in scientific writing. However, we acknowledge that linguistic backgrounds can play a role and potentially introduce biases in the evaluation of the backgrounds. In this regard, the use of GPT-3.5 as a tool for generating scientific backgrounds can be particularly beneficial for non-native English speakers, as it provides assistance in generating fluency and coherence in writing.

Strengths

The strengths of this study include the first structured evaluation of a LLMs generated background, comparing it to a background written by a human expert, and evaluations from 80 researchers across America, Europe, Asia, and Oceania, ensuring worldwide external validity.

Limitations

This study has several limitations. First, the LLMs tool used in this study was GPT-3.5 and not the latest version, GPT-4. Second, among the evaluators only 31% had experience in AKI research; however, in the subgroup analysis, we did not find a difference in this subgroup. Third, the H-index has inherent limitations as an evaluation tool, such as not considering the number of authors in an article, between others. However, in this study, we employed aiming to provide readers with an approximate idea of the experience in article writing. Fourth, the median H-index of the researchers who evaluated the backgrounds was three, which may have influenced the quality of evaluations due to their relatively lower research experience.

Additionally, it is important to note that our study compared a LLMs background with those of two individual investigators. While one of them is a PhD medical researcher with an H-index of 22, it is crucial to have a representative sample of researchers conducting scientific backgrounds to compare with a representative sample of LLMs backgrounds. Furthermore, the evaluation of these representative samples should be carried out by a separate representative sample of researchers. These considerations are important to ensure a more comprehensive assessment of the quality of researchers and the LLMs content in scientific writing.

CONCLUSIONS

Our study provides preliminary evidence suggesting the potential of GPT-3.5 in generating high-quality

scientific backgrounds. While the LLMs GPT-3.5 may help with background content development, the analyzed version was not able to generate references and reliable sources limiting its validity and readiness as an auxiliary tool for scientific medical writing. Therefore, researchers need to evaluate the quality of LLMs generated content and use these tools only to improve the grammar and coherence of article drafts, not to create them entirely. Further research with larger and more diverse samples is necessary to establish broader generalizations.

1 *Hospital Italiano de Buenos Aires, Buenos Aires, Argentina.*

2 *Universidad de Buenos Aires, Buenos Aires, Argentina.*

3 *Universidad Javeriana, Bogotá, Colombia.*

4 *Harvard Medical School, Boston, MA.*

5 *Facultad de Ciencias de la Salud, Universidad Simon Bolivar, Barranquilla, Colombia.*

6 *Mayo Clinic, Rochester, MN.*

7 *Texas A&M University, College Station, TX.*

8 *WellSpan Health, York, PA.*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccejournal).

Drs. Huespe and Echeverri both are first authors.

This study received support from the Ben Barres Spotlight Award from eLife.

The authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: ivan.huespe@hospitalitaliano.org.ar

Collaborative co-authors list is in Supplementary Material (http://links.lww.com/CCX/B255).

REFERENCES

1. Flanagin A, Bibbins-Domingo K, Berkwits M, et al: Nonhuman "Authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA* 2023; 329:637–639
2. OpenAI: Introducing ChatGPT. 2022. Available at: https://openai.com/blog/chatgpt. Accessed April 14, 2023
3. Salvagno M, Taccone FS, Gerli AG: Can artificial intelligence help for scientific writing? *Crit Care* 2023; 27:75
4. Ropek L: Everything We Know About OpenAI's ChatGPT. Gizmodo. 2023. Available at: https://gizmodo.com/chatgpt-openai-ai-finance-ai-everything-we-know-1850018307. Accessed April 10, 2023
5. Grewal H, Dhillon G, Monga V, et al: Radiology gets chatty: The ChatGPT saga unfolds. *Cureus* 2023; 15:e40135
6. Abbey A: Artificially intelligent reflection? Smoke and mirrors and a tale of two perspectives. *Intensive Care Med* 2023; 49:609–610
7. Biswas S: ChatGPT and the future of medical writing. *Radiology* 2023; 307:e223312
8. Levin G, Meyer R, Kadoch E, et al: Identifying ChatGPT-written OBGYN abstracts using a simple tool. *Am J Obstet Gynecol MFM* 2023; 5:100936
9. Lahat A, Shachar E, Avidan B, et al: Evaluating the use of large language model in identifying top research questions in gastroenterology. *Sci Rep* 2023; 13:4164
10. Gupta R, Pande P, Herzog I, et al: Application of ChatGPT in cosmetic plastic surgery: Ally or antagonist. *Aesthet Surg J* 2023; 43:NP587–NP590
11. Li J, Dada A, Kleesiek J, et al: ChatGPT in healthcare: A taxonomy and systematic review. *medRxiv* Preprint posted online March 30, 2023. doi: 10.1101/2023.03.30.23287899
12. Hosseini M, Rasmussen LM, Resnik DB: Using AI to write scholarly publications. *Account Res* 2023:1–9
13. Azamfirei R, Kudchadkar SR, Fackler J: Large language models and the perils of their hallucinations. *Crit Care* 2023; 27:120
14. Gao CA, Howard FM, Markov NS, et al: Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv* Preprint posted online December 27, 2022. doi: 10.1101/2022.12.23.521610
15. Bossuyt PM, Reitsma JB, Bruns DE, et al; STARD Group: STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015; 351:h5527
16. von Elm E, Altman DG, Egger M, et al; STROBE Initiative: The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *J Clin Epidemiol* 2008; 61:344–349
17. van Rooyen S, Black N, Godlee F: Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *J Clin Epidemiol* 1999; 52:625–629
18. Sullivan GM, Artino AR Jr: Analyzing and interpreting data from likert-type scales. *J Grad Med Educ* 2013; 5:541–542
19. Norman G: Likert scales, levels of measurement and the "laws" of statistics. *Adv Health Sci Educ Theory Pract* 2010; 15:625–632
20. King MR: The future of AI in medicine: A perspective from a Chatbot. *Ann Biomed Eng* 2022; 51:291–295
21. Salvagno M, Taccone FS, Gerli AG: Correction to: Can artificial intelligence help for scientific writing? *Crit Care* 2023; 27:99
22. Buholayka M, Zouabi R, Tadinada A: The readiness of ChatGPT to write scientific case reports independently: A comparative evaluation between human and artificial intelligence. *Cureus* 2023; 15:e39386
23. Anderson N, Belavy DL, Perle SM, et al: AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation. *BMJ Open Sport Exerc Med* 2023; 9:e001568