ETHICAL CONSIDERATIONS

The study protocol was reviewed and approved by the reference clinical research ethics committee (Hospital Universitario of Bellvitge, IRB00005523). Informed consent was appropriately obtained from all patients prior to study enrollment and kept on file. A specific sex-based approach was not taken according to the SAGER guidelines.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

No artificial intelligence tool was used.

AUTHOR CONTRIBUTIONS

A. Ariza-Solé and F. Formiga contributed to the study conception, data analysis, and the writing of this article. E. Calvo contributed to data collection and article revision. J. Comín-Colet, D. Monterde, and E. Vela contributed to critical review of the manuscript.

CONFLICTS OF INTEREST

D. Monterde and E. Vela are developers of the AMG tool. There are no other potential conflicts of interest.

Albert Ariza-Solé,[a,b,c,d,*] Francesc Formiga,[c,e] David Monterde,[f,g] Emili Vela,[g,h] Elena Calvo,[a,d,i,j] and Josep Comín-Colet[a,b,c,d]

[a]*Servicio de Cardiología, Àrea de Malalties del Cor, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat, Barcelona, Spain*
[b]*Bioheart, Grup de Malalties Cardiovasculars, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain*
[c]*Facultat de Medicina, Universitat de Barcelona, Barcelona, Spain*
[d]*Centro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Spain*
[e]*Unidad de Geriatría, Servicio de Medicina Interna, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat, Barcelona, Spain*
[f]*Institut Català de la Salut, Departament de Salut, Generalitat de Catalunya, Barcelona, Spain*
[g]*Digitalización para la Sostenibilidad del Sistema Sanitario (DSc), Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL), Barcelona, Spain*
[h]*Servei Català de la Salut, Departament de Salut, Generalitat de Catalunya, Barcelona, Spain*
[i]*Grup de Recerca Infermera (GRIN), Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain*
[j]*Facultat d'Infermeria, Universitat de Barcelona, Barcelona, Spain*

[*] Corresponding author.
E-mail address: aariza@bellvitgehospital.cat (A. Ariza-Solé).
𝕏 @AlbertAriza3

Available online 12 March 2024

REFERENCES

1. Sanchis J, García Acuña JM, Raposeiras S, et al. Comorbidity burden and revascularization benefit in elderly patients with acute coronary syndrome. *Rev Esp Cardiol.* 2021;74:765–772.
2. Monterde D, Vela E, Clèries M; grupo colaborativo GMA. Los grupos de morbilidad ajustados: nuevo agrupador de morbilidad poblacional de utilidad en el ámbito de la atención primaria [Adjusted morbidity groups: A new multiple morbidity measurement of use in Primary Care]. *Aten Primaria.* 2016;48:674–682.
3. Clèries M, Monterde D, Vela E, Guarga Àgae, García Eroles L, Pérez Sust P; Grupo de validación. Validación clínica de 2 agrupadores de morbilidad en el ámbito de atención primaria. *Aten Primaria.* 2020;52:96–103.

https://doi.org/10.1016/j.rec.2024.03.001

ChatGPT-4 versus human assessment in cardiology peer review

ChatGPT-4 frente a evaluación humana para la revisión por pares en cardiología

To the Editor,

Generative language models, especially ChatGPT, have impacted science and society.[1,2] While artificial intelligence (AI) has made significant inroads in plagiarism detection and curating studies for systematic reviews,[3] its application in scientific peer review is unexplored. Peer review, a resource-intensive process both economically and in terms of human effort, may benefit from the efficiency of AI in speed of data processing, accuracy, and the ability to synthesize vast amounts of information. This study evaluated the ability of ChatGPT to generate valid scientific reviews in cardiology compared with human experts.

The study included consecutive scientific letters from May 2022 to May 2023 that underwent peer review in *Revista Española de Cardiología* (*Rev Esp Cardiol*), the official scientific journal of the Spanish Society of Cardiology, founded in 1947, and ranked within the first quartile of cardiovascular journals in Journal Citation Reports 2022.[4,5] Original articles and reviews were excluded because they exceeded the maximum text length of ChatGPT. For each scientific letter, a review (GPTr) was generated using the ChatGPT model. A custom prompt was developed through iterative testing with published scientific letters to guide ChatGPT's responses when reviewing scientific letters. This prompt was refined for *Rev Esp Cardiol* standards and was used to generate all GPTr. The Application Programming Interface was used with the "gpt-4-0613" model.

The quality of GPTr and human review (Hr) were evaluated by the associate editors of *Rev Esp Cardiol* (P. Avanzas, D. Filgueiras-Rama, P. García-Pavía) and its editor-in-chief (L. Sanchis). The standard review process for scientific letters in *Rev Esp Cardiol* includes 2 reviewers, and the associate editor in charge of the letter assigns a score of 0 to 100 points to each review for overall quality. The reviewer selected as reviewer number 1 during the standard review process was considered the Hr. The same editor who initially managed the manuscript during the standard review process also evaluated the overall quality of GPTr, scoring it from

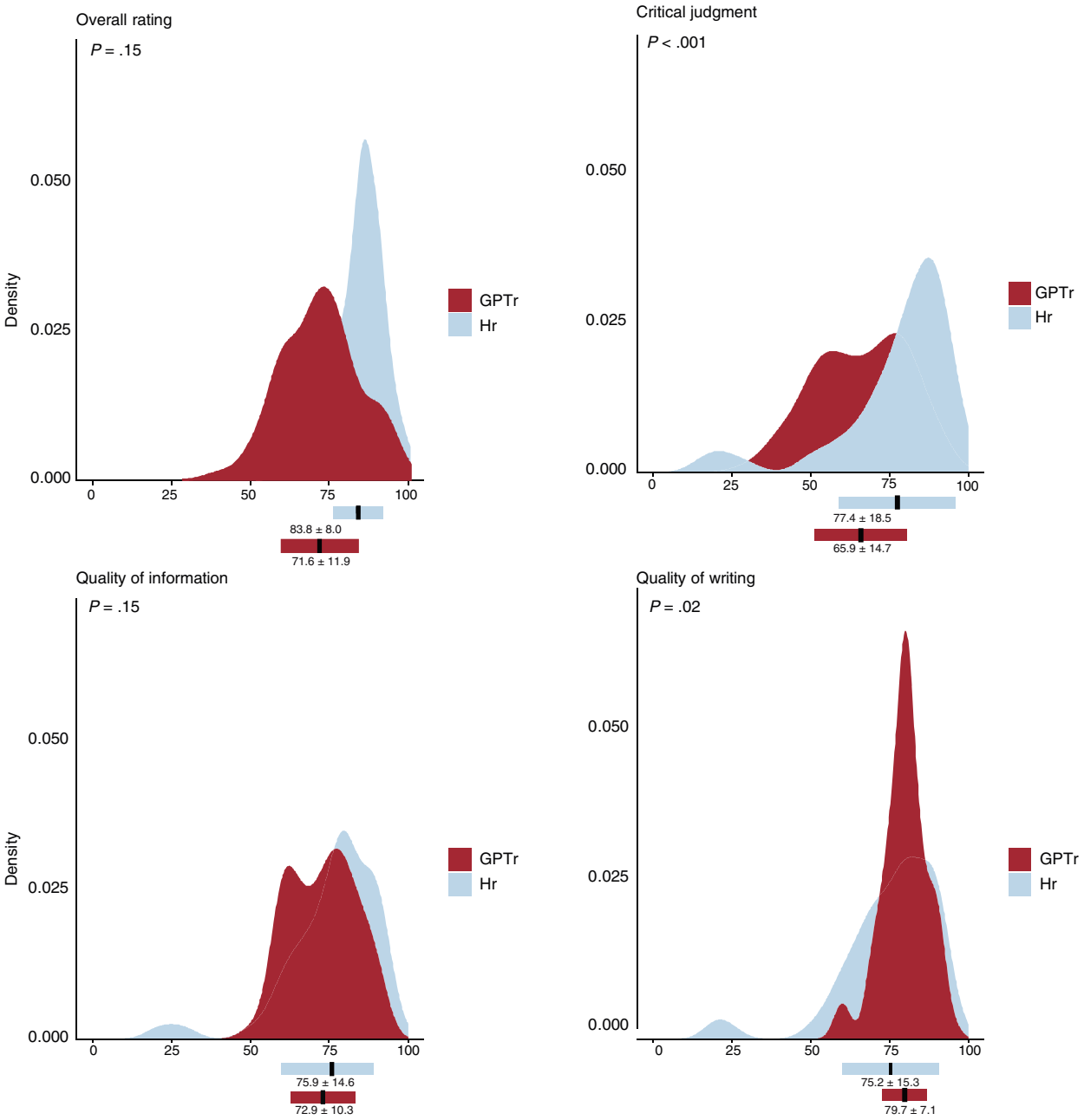

Figure 1. Density plots showing the distribution of GTPr and Hr assessments for quality of information, quality of writing, and critical judgment. GTPr, review generated using the ChatGPT model; Hr, human review.

0 to 100. A second randomly selected editor evaluated the Hr and GPTr in a blinded fashion. For this purpose, Hr and GPTr were randomly presented, anonymized, and labelled as 'Response1' or 'Response2' with the scientific letter. The second editor analyzed the following domains: information quality, writing quality, and critical judgment, providing a score of 0 to 100 points for each domain. The editors were also asked to guess which review was the Hr/GPTr and determine which review was better.

The Student t-test for independent samples was used to compare average Hr and GPTr quality scores and the chi-square test was used for categorical variables. The endpoint selected to estimate the sample size was the editor's review preference (GPTr or Hr). Assuming an alpha risk of 0.05 and a beta risk of 0.2 in a bilateral contrast, a minimum of 48 response pairs (each comprising a GPTr and Hr) were needed to detect a 20% difference in response preference, assuming 65% vs 45% preference for Hr vs GPTr. This study was carried out in accordance with the latest edition of the International Committee of Medical Journal Editors' recommendations.[6]

All 85 scientific letters received by *Rev Esp Cardiol* during the study period and subjected to peer review were initially selected. Ten letters (11.7%) were excluded because they were originally submitted and reviewed as original articles. In these, the authors were offered the chance of converting their articles into scientific

letters after the review process. Therefore, 75 scientific letters were included in the analysis. A total of 911 907 tokens were sent to ChatGPT, with 483 681 being completion tokens, generating 75 GPTr for $56.38 ($0.66/review). Hr received a better average overall rating than GPTr when evaluated by the unblinded original editor (83.8 ± 8.8 vs 71.6 ± 11.9 points; $P < .001$) (figure 1). The correlation between the 2 evaluations was poor (R = 0.209; $P = .079$).

The blinded editor's assessment showed that the information quality was similar for GPTr and Hr (72.9 ± 10.3 vs 75.9 ± 14.6 points; $P = .15$; GPTr better in 32 [43%] letters). GPTr obtained a higher score in writing quality (79.6 ± 7.1 vs 75.2 ± 15.3 points; $P = .02$; GPTr better in 51 [68%] letters), while Hr exhibited greater critical judgment (65.87 ± 14.69 vs 77.4 ± 18.5 points; $P < .001$; GPTr better in 21 [28%] letters) (figure 1). Hr assessments had more outliers, while GPTr assessments were more homogeneous (figure 1). The editor correctly assessed whether the review was GPTr or Hr in most instances (n = 74, 99%). Interestingly, GPTr was considered better than Hr in 27 cases (36%).

In this study, we evaluated the quality of a generative natural language model for generating scientific editorial reviews in cardiology and compared them with human reviews. We found that Hr provided a better review overall, particularly in the critical judgment domain. However, GPTr was considered better in around one-third of letters and had more homogeneous quality scores. In contrast, Hr quality exhibited greater dispersion as a result of the poor quality of some reviews. Indeed, finding good reviewers is currently a challenge. Our results could be of interest in an era when AI is increasingly applied in different fields, scientific publications are growing exponentially, and scientific evaluation is becoming expensive and problematic. The quality of information was similar, but GPTr had better writing quality, which can be attributed to the ability of the model to generate well-structured responses based on large amounts of prior data.[6] HR outperformed GPTr in critical judgment, likely due to human experience, intuition, and specialized expertise. Despite being adept at data pattern analysis, GPTr lacks nuanced discernment. This underscores the irreplaceability of human analysis in contexts requiring critical judgment. Nevertheless, ChatGPT-4 could be used as an initial screening tool in the peer review process, helping reviewers to organize and write their evaluations better.

The limitations of this study include: *a)* its retrospective nature; *b)* its exclusive focus on one journal, Rev Esp Cardiol, which could restrict the generalizability of our results to other publications and fields; and *c)* its evaluation of scientific letters only, not original articles, which could limit our findings due to differences in format and content depth between these article types.

In summary, the concerns raised by funding agencies about confidentiality and originality in AI-generated peer reviews underscore the need for ethical and methodological safeguards. In our opinion, AI might help the review process by summarizing article contents and helping reviewers not to overlook relevant information. However, reviewers' critical judgment and original thoughts are unique attributes essential for a good review.

FUNDING

None.

ETHICAL CONSIDERATIONS

The work did not require approval from the ethics committee. No patient data were used. Sex and gender bias was not considered as sex/gender was not analyzed.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

ChatGPT-4 was used to generate scientific reviews as part of the methodology of this work.

AUTHORS' CONTRIBUTIONS

All authors contributed to the design of the study. A. Fernández-Cisnal and J. Sanchis wrote the first draft of the article. P. Avanzas, D. Filgueiras-Rama, P. Garcia-Pavia and L. Sanchis reviewed the article.

CONFLICTS OF INTEREST

J. Sanchis is editor-in-chief of *Rev Esp Cardiol*, and P. Avanzas, D. Filgueiras-Rama, P. Garcia-Pavia and L. Sanchis are associate editors of *Rev Esp Cardiol*. The journal's editorial procedure to ensure impartial processing of the manuscript has been followed. The authors have no other conflicts of interest to declare.

ACKNOWLEDGEMENTS

The authors wish to thank the editorial office of *Rev Esp Cardiol* for their work preparing the scientific letters.

Agustín Fernández-Cisnal,[a] Pablo Avanzas,[b,c,d,e] David Filgueiras-Rama,[e,f,g] Pablo Garcia-Pavia,[e,g,h] Laura Sanchis,[i,j] and Juan Sanchis[a,e,k,*]

[a]*Department of Cardiology, Hospital Clínico Universitario de València, INCLIVA, Valencia, Spain*
[b]*Department of Cardiology, Hospital Universitario Central de Asturias, Oviedo, Asturias, Spain*
[c]*Instituto de Investigación del Principado de Asturias, Oviedo, Asturias, Spain*
[d]*Facultad de Medicina, Universidad de Oviedo, Oviedo, Asturias, Spain*
[e]*Centro de investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Spain*
[f]*Instituto de Investigación Sanitaria del Hospital Clínico San Carlos (IdISSC), Instituto Cardiovascular, Madrid, Spain*
[g]*Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain*
[h]*Department of Cardiology, Hospital Universitario Puerta de Hierro-Majadahonda, Instituto de Investigación Sanitaria Puerta de Hierro-Segovia de Arana (IDIPHISA), Madrid, Spain*
[i]*Department of Cardiology, Hospital Clínic, Barcelona, Spain*
[j]*Institut d'Investigacions Biomèdiques Agustí Pi i Sunyer (IDIBAPS), Barcelona, Spain*
[k]*Facultad de Medicina, Universidad de València, València, Spain*

[*] Corresponding author.
E-mail address: sanchis_juafor@gva.es (J. Sanchis).
𝕏 @JuanSanchisFor (J. Sanchis)

Available online 27 February 2024

REFERENCES

1. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare.* 2023;11:887.

2. Fernández-Cisnal A, Lopez-Ayala P, Miñana G, Boeddinghaus J, Mueller C, Sanchis J. Performance of an artificial intelligence chatbot with web search capability in cardiology-related assistance: a simulation study. *Rev Esp Cardiol.* 2023;76:1065–1067.
3. Elali FR, Rachid LN. AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns.* 2023;4:100706.
4. Shah NB. Challenges, experiments, and computational solutions in peer review. *Commun ACM.* 2022;65:76–87.
5. Sanchis J, Avanzas P, Filgueiras-Rama D, García-Pavía P, Sanchis L. Revista Española de Cardiología 2022. *Rev Esp Cardiol.* 2023;76:370–378.
6. International Committee of Medical Journal Editors (ICMJE). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Working Medical Journals. 2024. Available at https://www.icmje.org/recommendations/. Accessed 10 Jan 2024.

https://doi.org/10.1016/j.rec.2024.02.004