Sandra Villar,[a] Anna Mollar,[a,b] Miguel Lorenzo,[a] Gonzalo Núñez,[a] Rafael de la Espriella,[a] and Julio Núñez[a,b,*]

[a] *Servicio de Cardiología, Hospital Clínico Universitario de Valencia, Universitat de Valencia, Instituto de Investigación Sanitaria (INCLIVA), Valencia, Spain*
[b] *Centro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Spain*

* Corresponding author.
E-mail address: yulnunez@gmail.com (J. Núñez).
𝕏 @yulnunezvill (J. Núñez), @Sandra_ViCo88 (S. Villar).

Available online 22 November 2023

REFERENCES

1. van Essen BJ, Tromp J, Ter Maaten JM, et al. Characteristics and clinical outcomes of patients with acute heart failure with a supranormal left ventricular ejection fraction. *Eur J Heart Fail.* 2023;25:35–42.
2. Santas E, Llácer P, Palau P, et al. Noncardiovascular morbidity and mortality across left ventricular ejection fraction categories following hospitalization for heart failure. *Rev Esp Cardiol.* 2023. https://doi.org/10.1016/j.rec.2023.05.005.
3. Horiuchi Y, Asami M, Ide T, et al. Prevalence, characteristics and cardiovascular and non-cardiovascular outcomes in patients with heart failure with supra-normal ejection fraction: Insight from the JROADHF study. *Eur J Heart Fail.* 2023;25:989–998.
4. Paulus WJ, Zile MR. From Systemic Inflammation to Myocardial Fibrosis: The Heart Failure With Preserved Ejection Fraction Paradigm Revisited. *Circ Res.* 2021;128:1451–1467.

https://doi.org/10.1016/j.recesp.2023.11.002
1885-5857/© 2023 Sociedad Española de Cardiología. Published by Elsevier España, S.L.U. All rights reserved.

Assessing the accuracy of ChatGPT as a decision support tool in cardiology

Evaluación de la fiabilidad de ChatGPT como herramienta de soporte a la toma de decisiones en cardiología

To the Editor,

ChatGPT, an artificial intelligence dialogue-based language model, has generated strong expectations worldwide due to its surprising ability to convincingly answer complex queries formulated in plain natural language. It has been used in a wide variety of fields, including education, computer programming, and journalism, with potentially paradigm-shifting results. The medical community is no exception. ChatGPT has successfully passed the exams required to obtain a medical license,[1] draft scientific abstracts,[2] and compose complete medical reports.[3] In cardiology, the bot has provided appropriate cardiology-related assistance for common cardiovascular conditions in simulated patients[4] and has outperformed medical students in standardized cardiovascular tests.[5]

In light of the above, there is a strong temptation to try ChatGPT out as a decision-support tool in real-world clinical data. However, it is important to ask whether ChatGPT is able to process real-world medical records and suggest appropriate treatment. Most of the current literature focuses on its application in "synthetic" databases with highly preprocessed, curated texts, and/or multiple-choice answers.[1–6] Real-world accuracy cannot be directly inferred from those settings. To answer this question, we assessed the agreement between ChatGPT and a heart team consisting of cardiologists and cardiac surgeons in a specific use case: the decision-making process in patients with severe aortic stenosis.

We performed a descriptive retrospective analysis of the medical records of 50 consecutive patients with aortic stenosis presented at a heart team meeting of our institution between January 1, 2022 and February 14, 2022 (these dates were chosen to guarantee that information on the patients' eventual treatment was available). Depending on a wide variety of variables, the treatment of these patients consisted of the following options: *a)* surgical valve replacement; *b)* percutaneous valve implant; or *c)* medical treatment. The management strategies of the heart team were compared with those recommended by ChatGPT. An anonymized summary of each patient's status was produced by a cardiologist, who copy-pasted together the following sections from the electronic health record: demographics, past medical history, echocardiogram, coronary angiogram, symptoms, and diagnosis. During the second half of February 2023, this information was entered 3 times as a prompt in ChatGPT (GPT-3.5, 13 February 2023 version) as part of an enquiry about the optimal treatment. Initially, the question was "What is the best treatment for this patient?", but the responses of ChatGPT were too comprehensive and included medications and interventions for any concurrent comorbidities in the test patient. Therefore, the final prompt used for the experiments was "What is the best treatment for the aortic stenosis in the following patient?" to elicit a focused response that would facilitate data interpretation, labeling, classification, and processing. No further changes to the prompt were necessary to obtain meaningful answers. Responses were codified as *a)* surgery; *b)* transcatheter aortic valve implantation (TAVI); *c)* medical treatment; *d)* undefined intervention (ChatGPT recommended aortic valve replacement but did not specify whether the approach should be surgical or percutaneous); or *e)* inconclusive. The results were classified according to the following definitions:

- *Fully consistent*: all 3 responses recommended exactly the same treatment.
- *Partially consistent*: all 3 responses recommended a similar approach (intervention vs medical treatment).
- *Full agreement*: fully consistent response that matched the heart team's assessment.
- *Agreement on approach*: fully or partially consistent response that matched the heart team's "intervention vs medical treatment" assessment.

Figure 1 shows the results in detail. The mean age was 78 years, and 41% were men. The heart team's decision was TAVI in 56%, surgery in 40%, and medical treatment in 4% of cases. Of 150 responses generated by ChatGPT, 14 (9%) were inconclusive. A total of 70% of ChatGPT's recommendations were at least partially consistent and 38% were fully consistent. There was *agreement on approach* in 58% of the cases but *full agreement* in only 18% of cases. Fifteen recommendations were inconsistent and 6 recommendations that were consistent diverged from the heart team's decision, representing a total of 21 errors. Of these 21 cases, 10 (48%) had other concomitant valve or coronary artery disease requiring intervention, 4 (19%) were cases in which the indications

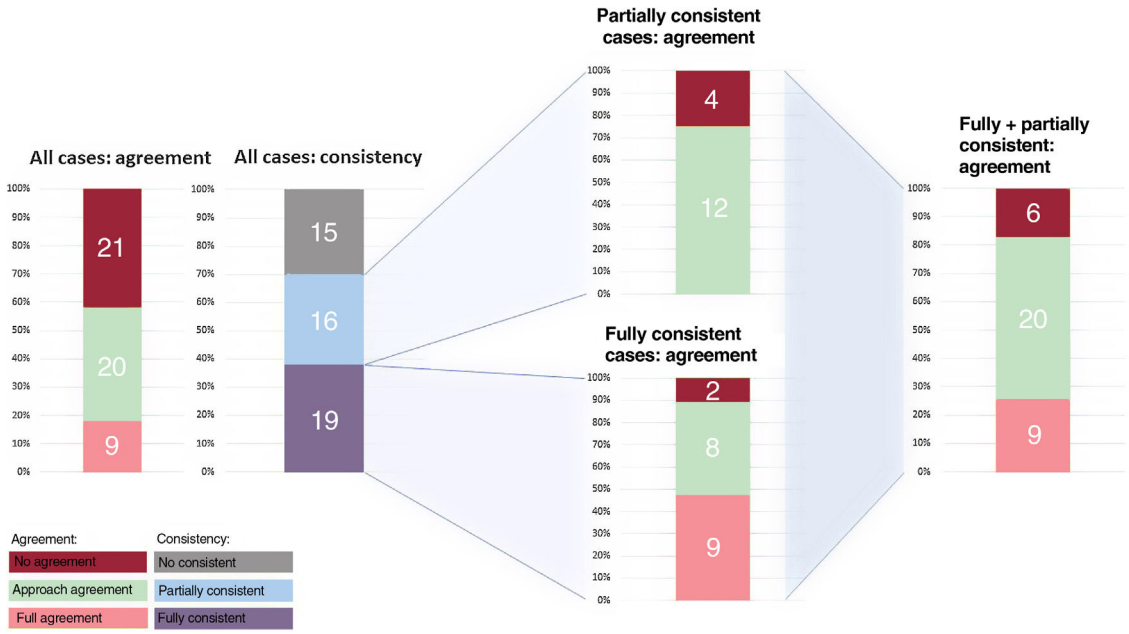

Figure 1. ChatGPT consistency and agreement.

for intervention had a lower level of evidence, and 7 (33%) were cases of isolated symptomatic aortic stenosis. However, when the recommendations were fully consistent, ChatGPT showed at least *agreement in approach* in 89% of the cases.

This study has some limitations, including its exploratory nature and small sample size. In addition, as a single-center experience, the gold standard was the decision of a particular heart team, which theoretically could have its own biases.

ChatGPT's treatment suggestions agreed with those of the medical experts in 58% of the cases. Agreement was low for specific treatments, and moderate for intervention vs medical treatment. Unsurprisingly, ChatGPT tended to agree with the heart team's decision in cases where it consistently provided similar answers to repeated instances of the same question. However, agreement and consistency were substantially impaired in clinically complex cases. These results were obtained using a system designed solely as a generic conversational bot with no specialized training in a highly challenging context (open-ended question, complex disease). There results could be markedly be improved by future versions specifically trained for medical decision support.

FUNDING

This work has not received any specific funding.

ETHICAL CONSIDERATIONS

Due to the retrospective nature of this work and the complete anonymization of data, the requirement to retrieve informed consent from patients was waived. The research protocol was reviewed and approved by *Comité de Ética de la Investigación con Medicamentos del Área de Salud Valladolid Este* with code PI 23-3194. The study design, methodology and data were analyzed for signs of any potential gender bias and none was found.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

As described in the text, ChatGPT 3.5 was used in the experiments of this work as subject of study. No IA-based tool whatsoever was used for manuscript writing or for data analysis or interpretation of the results.

AUTHORS' CONTRIBUTIONS

C. Baladrón, T. Sevilla and J.A. San Román designed the study; C. Baladrón and T. Sevilla designed and performed the experiments. C. Baladrón and J.A. San Román wrote the initial draft of the manuscript. M. Carrasco-Moraleja, I. Gómez-Salvador and J. Peral-Oliveira reviewed the data and methodology, produced the figures, and performed data analysis. All authors participated in manuscript reviewing and correction.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

Carlos Baladrón,[a,b] Teresa Sevilla,[a,b,*] Manuel Carrasco-Moraleja,[a,b] Itziar Gómez-Salvador,[a,b] Julio Peral-Oliveira,[a] and José Alberto San Román[a,b]

[a]*Servicio de Cardiología, Hospital Clínico Universitario de Valladolid, Valladolid, Spain*
[b]*Centro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Spain*

* Corresponding author.
E-mail address: tereseru@gmail.com (T. Sevilla).
@cbalzor (C. Baladrón), @TreSeru (T. Sevilla).

Available online 4 December 2023

REFERENCES

1. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ.* 2023;9:e45312.
2. Else H. Abstracts written by ChatGPT fool scientists. *Nature.* 2023;613:423.
3. Jeblick K, Schachtner B, Dexl K, et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. *Eur Radiol.* 2023 https://doi.org/10.1007/s00330-023-10213-1.
4. Fernández-Cisnal A, López-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. https://doi.org/10.1016/j.rec.2023.06.008.
5. Hariri W. Analyzing the Performance of ChatGPT in Cardiology and Vascular Pathologies. *Research square preprints [preprint].* 2023. Available at: https://doi.org/10.21203/rs.3.rs-2782768/v1. Accessed 9 Oct 2023
6. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J Am Coll Radiol.* 2023;20:990–997.

https://doi.org/10.1016/j.recesp.2023.11.011
1885-5857/© 2023 Sociedad Española de Cardiología. Published by Elsevier España, S.L.U. All rights reserved.