

FINANCIACIÓN

Este proyecto ha sido financiado por el Departament de Salut, Generalitat de Catalunya (Pla Estratègic de Recerca i Innovació en Salut [PERIS] 2016–2020). SLT006/17/00080 (18DPS008).

CONSIDERACIONES ÉTICAS

El protocolo del estudio fue revisado y aprobado por el comité ético de investigación clínica de referencia (Hospital Universitario de Bellvitge, IRB00005523). Se obtuvo y se archivó de forma pertinente el consentimiento informado por parte de todos los pacientes antes de ser incluidos en el estudio. No se realizó un abordaje específico de las variables de sexo según las directrices SAGER.

DECLARACIÓN SOBRE EL USO DE INTELIGENCIA ARTIFICIAL

No utilizada.

CONTRIBUCIÓN DE LOS AUTORES

A. Ariza-Solé y F. Formiga contribuyeron a la concepción del estudio, el análisis de los datos y la escritura del manuscrito. E. Calvo contribuyó a la recolección de los datos y la revisión de manuscrito. J. Comín-Colet, D. Monterde y E. Vela contribuyeron a la revisión crítica del manuscrito.

CONFLICTO DE INTERESES

D. Monterde y E. Vela son desarrolladores de la herramienta GMA. No existe otro potencial conflicto de intereses.

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} y Josep Comín-Colet^{a,b,c,d}

^aServicio de Cardiología, Àrea de Malalties del Cor, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat, Barcelona, España

^bBioheart, Grup de Malalties Cardiovasculars, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, España

^cFacultat de Medicina, Universitat de Barcelona, Barcelona, España

^dCentro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), España

^eUnidad de Geriátrica, Servicio de Medicina Interna, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat, Barcelona, España

^fInstitut Català de la Salut, Departament de Salut, Generalitat de Catalunya, Barcelona, España

^gDigitalización para la Sostenibilidad del Sistema Sanitario (DSc), Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL), Barcelona, España

^hServei Català de la Salut, Departament de Salut, Generalitat de Catalunya, Barcelona, España

ⁱGrup de Recerca Infermera (GRIN), Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, España

^jFacultat d'Infermeria, Universitat de Barcelona, Barcelona, España

* Autor para correspondencia.

Correo electrónico: aariza@bellvitgehospital.cat (A. Ariza-Solé).

✉ @AlbertAriza3

On-line el 1 de febrero de 2024

BIBLIOGRAFÍA

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.recesp.2024.01.008>

0300-8932/© 2024 Sociedad Española de Cardiología. Publicado por Elsevier España, S.L.U. Todos los derechos reservados.

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



ChatGPT-4 versus human assessment in cardiology peer review

Sr. Editor:

Los modelos de lenguaje generativos, y en especial ChatGPT, han tenido un gran impacto en la ciencia y la sociedad^{1,2}. Aunque la inteligencia artificial (IA) ha logrado avances significativos en la detección de plagios y la selección de estudios para revisiones sistemáticas³, no se ha explorado su posible aplicación a la revisión científica por pares. La revisión por pares es un proceso que requiere muchos recursos tanto económicos como de esfuerzo humano, y es posible que pueda beneficiarse de la eficiencia de la IA en cuanto a la rapidez del procesamiento de datos, la exactitud y la capacidad de síntesis de cantidades enormes de información. En este estudio se evaluó la capacidad de ChatGPT de generar revisiones científicas válidas en cardiología en comparación con expertos humanos.

Se incluyeron en el estudio cartas científicas consecutivas entre mayo de 2022 y mayo de 2023 que fueron objeto de una revisión

por pares en la Revista Española de Cardiología (*Rev Esp Cardiol*), que es el órgano científico oficial de la Sociedad Española de Cardiología, fundada en 1947, y se encuentra en el primer cuartil de las revistas cardiovasculares en el *Journal Citation Reports* de 2022^{4,5}. Se excluyeron los artículos originales y las revisiones, ya que superaban la máxima extensión de texto para ChatGPT. De cada carta científica, se generó una revisión con el modelo de ChatGPT (RGPT). Mediante pruebas iterativas con cartas científicas publicadas, se elaboró un *prompt* (indicación) específico para orientar las respuestas de ChatGPT en la revisión de cartas científicas. Este *prompt* se afinó según las normas de *Rev Esp Cardiol* y se utilizó para generar todas las RGPT. Se utilizó la interfaz de programación de aplicaciones *Application Programming Interface* con el modelo «gpt-4-0613».

Los editores asociados de *Rev Esp Cardiol* (P. Avanzas, D. Filgueiras-Rama, P. García-Pavía y L. Sanchis) y su editor jefe (J. Sanchis) revisaron la calidad de la RGPT y la revisión humana (RH). El proceso de revisión estándar para las cartas científicas en *Rev Esp Cardiol* incluye la participación de 2 revisores, y el editor asociado encargado de la carta asigna una puntuación de 0 a 100 a la calidad general. Se tomó como RH la del elegido como revisor

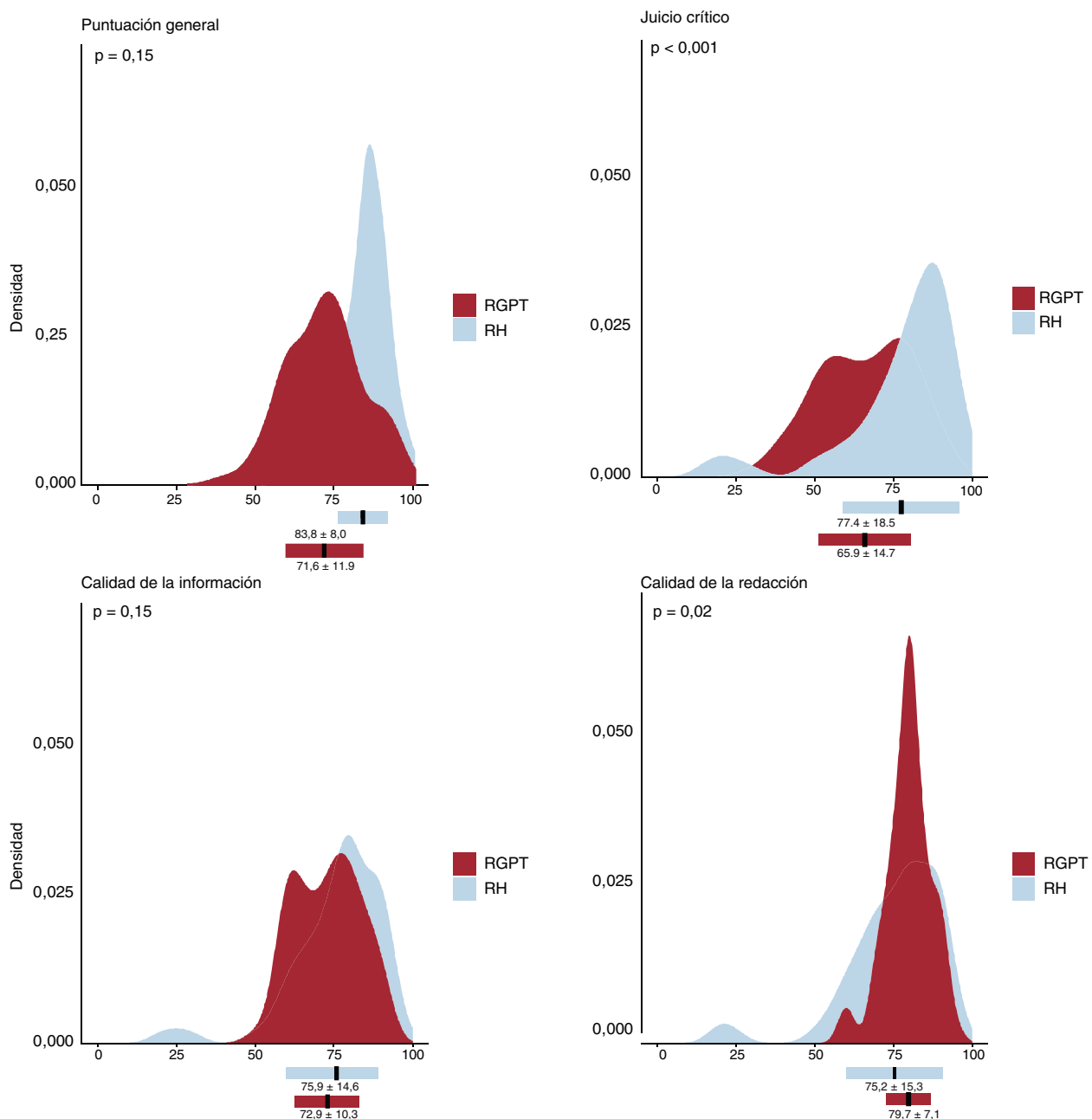


Figura 1. Gráficos de densidad que muestran la distribución de las evaluaciones de las RGPT y las RH en cuanto a la calidad de la información, la calidad de la redacción y el juicio crítico. RGPT: revisión generada con el modelo de ChatGPT; RH: revisión humana.

número 1 durante el proceso estándar de revisión. El mismo editor que había gestionado inicialmente el manuscrito durante el proceso de revisión estándar evaluó también la calidad general de la RGPT, y le asignó una puntuación de 0 a 100. Un segundo editor elegido de manera aleatoria evaluó la RH y la RGPT utilizando un diseño con enmascaramiento. Para este fin, la RH y la RGPT se presentaron en un orden aleatorio, anonimizadas y designadas como «Respuesta1» o «Respuesta2» para la carta científica. El segundo editor analizó los siguientes dominios: calidad de la información, calidad de la redacción y juicio crítico, y asignó una puntuación de 0 a 100 a cada dominio. Se pidió también a los editores que adivinaran qué revisión era la RH y cuál la RGPT y determinarían cuál era mejor.

Se utilizó una prueba de la t de Student para muestras independientes para comparar las puntuaciones medias de calidad

de la RH y la RGPT y la prueba de la χ^2 para las variables cualitativas. El objetivo elegido para estimar el tamaño de la muestra fue la preferencia de revisión (RGPT o RH) elegida por el editor. Partiendo del supuesto de un riesgo alfa de 0,05 y un riesgo beta de 0,2 en un contraste bilateral, se necesita un mínimo de 48 pares de respuestas (cada uno formado por una RGPT y una RH) para detectar una diferencia del 20% en la preferencia de respuesta, suponiendo unas preferencias del 65 frente al 45% por la RH frente a la RGPT. El estudio se llevó a cabo según lo indicado en la edición más reciente de las recomendaciones del Comité Internacional de Editores de Revistas Médicas⁶.

Se seleccionaron inicialmente las 85 cartas científicas recibidas en *Rev Esp Cardiol* durante el periodo de estudio y se les aplicó la revisión por pares. Se excluyeron 10 cartas (11,7%) porque se habían enviado y revisado inicialmente en forma de artículos

originales. En estos casos se ofreció a los autores la posibilidad de convertir sus artículos en cartas científicas tras el proceso de revisión. En consecuencia, se incluyeron en el análisis 75 cartas científicas. Se enviaron a ChatGPT un total de 911.907 *tokens* (unidades o elementos de texto), de los cuales 483.681 fueron *completion tokens* (*tokens* de finalización o de completado), que generaron 75 RGPT con un coste de 56,38 dólares (0,66 dólares por revisión). La RH obtuvo una puntuación general media mejor que la RGPT en la evaluación realizada por el editor original sin enmascaramiento ($83,8 \pm 8,8$ frente a $71,6 \pm 11,9$ puntos; $p < 0,001$) (figura 1). Hubo una mala correlación entre las 2 evaluaciones ($r = 0,209$; $p = 0,079$).

La evaluación del editor con enmascaramiento mostró que las calidades de la información de la RGPT y la RH eran similares ($72,9 \pm 10,3$ frente a $75,9 \pm 14,6$ puntos; $p = 0,15$; la RGPT fue mejor en 32 [43%] de las cartas). La RGPT obtuvo una puntuación más alta en cuanto a la calidad de la redacción ($79,6 \pm 7,1$ frente a $75,2 \pm 15,3$ puntos; $p = 0,02$; la RGPT fue mejor en 51 [68%] de las cartas), mientras que la RH mostró un mejor juicio crítico ($65,87 \pm 14,69$ frente a $77,4 \pm 18,5$ puntos; $p < 0,001$; la RGPT fue mejor en 21 [28%] de las cartas) (figura 1). En las evaluaciones de la RH hubo más valores atípicos, mientras que las evaluaciones de la RGPT fueron más homogéneas (figura 1). El editor determinó correctamente si la revisión era la RGPT o la RH en la mayor parte de los casos ($n = 74$, 99%). Es interesante señalar que la RGPT se consideró mejor que la RH en 27 casos (36%).

En este estudio se evalúa la calidad de un modelo de lenguaje natural generativo para revisiones editoriales científicas en cardiología y se compara con la de las RH. Se observa que la RH proporciona una mejor revisión general, en especial en el dominio del juicio crítico. Sin embargo, la RGPT se considera mejor en alrededor de una tercera parte de las cartas y obtiene unas puntuaciones de calidad más homogéneas. En cambio, la calidad de la RH muestra una mayor dispersión como resultado de la mala calidad de algunas revisiones. De hecho, encontrar buenos revisores hoy es un verdadero reto. Nuestros resultados podrían ser de interés en una época en la que, mientras la IA se aplica cada vez más en diferentes campos, las publicaciones científicas están aumentando exponencialmente, y la evaluación científica está pasando a ser cara y problemática. La calidad de la información fue similar, pero la RGPT mostró una mejor calidad de redacción, lo cual puede atribuirse a la capacidad del modelo de generar respuestas bien estructuradas utilizando grandes cantidades de datos previos⁶. La RH superó a la RGPT en juicio crítico, probablemente por la experiencia, la intuición y el conocimiento experto especializado de los humanos. A pesar de su carácter experto en el análisis de patrones de datos, la RGPT carece de un discernimiento matizado. Esto subraya que el análisis humano es irremplazable en contextos que requieren juicio crítico. No obstante, ChatGPT-4 podría usarse como instrumento de cribado inicial en el proceso de la revisión por pares y podría ayudar a los revisores a organizar y redactar mejor sus evaluaciones.

Las limitaciones de este estudio son: a) su carácter retrospectivo; b) se centra únicamente en una revista, *Rev Esp Cardiol*, lo cual podría limitar la posibilidad de generalizar los resultados a otras publicaciones y otros campos, y c) la evaluación solo de cartas científicas, no de artículos originales, que podría limitar los resultados debido a las diferencias de formato y profundidad del contenido entre estos tipos de artículos.

En resumen, las preocupaciones planteadas por los organismos de financiación acerca de la confidencialidad y la originalidad de las revisiones por pares generadas mediante IA subrayan la necesidad de salvaguardas éticas y metodológicas. En nuestra opinión, la IA podía ser útil en el proceso de revisión al resumir el contenido del artículo y ayudar a los revisores a no pasar por alto información pertinente. Sin embargo, el juicio crítico y los

pensamientos originales de los revisores son atributos únicos que resultan esenciales en una buena revisión.

FINANCIACIÓN

Ninguno.

CONSIDERACIONES ÉTICAS

El trabajo no requirió aprobación del comité de ética. No se utilizaron datos de pacientes. No se consideró el sesgo de sexo y género, ya que no se analizaron el sexo ni el género.

DECLARACIÓN SOBRE EL USO DE INTELIGENCIA ARTIFICIAL

Se utilizó ChatGPT-4 para generar revisiones científicas como parte de la metodología de este trabajo.

CONTRIBUCIÓN DE LOS AUTORES

Todos los autores participaron en el diseño del estudio. A. Fernández-Cisnal y J. Sanchis redactaron la versión preliminar del artículo. P. Avanzas, D. Filgueiras-Rama, P. García-Pavía y L. Sanchis revisaron el artículo.

CONFLICTO DE INTERESES

J. Sanchis es editor jefe de *Rev Esp Cardiol* y P. Avanzas, D. Filgueiras-Rama, P. García-Pavía y L. Sanchis son editores asociados de *Rev Esp Cardiol*. Se ha seguido el procedimiento editorial establecido por la Revista para garantizar un procesamiento imparcial del manuscrito. Los autores no tienen otros conflictos de intereses que declarar.

AGRADECIMIENTOS

Los autores desean agradecer a la oficina editorial de *Rev Esp Cardiol* su trabajo de preparación de las cartas científicas.

Agustín Fernández-Cisnal^a, Pablo Avanzas^{b,c,d,e}, David Filgueiras-Rama^{e,f,g}, Pablo García-Pavía^{e,g,h}, Laura Sanchis^{i,j} y Juan Sanchis^{a,e,k}.

^a Department of Cardiology, Hospital Clínico Universitario de València, INCLIVA, Valencia, España

^b Department of Cardiology, Hospital Universitario Central de Asturias, Oviedo, Asturias, España

^c Instituto de Investigación del Principado de Asturias, Oviedo, Asturias, España

^d Facultad de Medicina, Universidad de Oviedo, Oviedo, Asturias, España

^e Centro de investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), España

^f Instituto de Investigación Sanitaria del Hospital Clínico San Carlos (IdISSC), Instituto Cardiovascular, Madrid, España

^g Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, España

^h Department of Cardiology, Hospital Universitario Puerta de Hierro-Majadahonda, Instituto de Investigación Sanitaria Puerta de Hierro-Segovia de Arana (IDIPHISA), Madrid, España

ⁱ Department of Cardiology, Hospital Clínic, Barcelona, España

^j Institut d'Investigacions Biomèdiques Agustí Pi i Sunyer (IDIBAPS),
Barcelona, España

^k Facultad de Medicina, Universidad de València, València, España

* Autor para correspondencia.

Correo electrónico: sanchis_juafor@gva.es (J. Sanchis).

✉ [@JuanSanchisFor](#) (J. Sanchis)

On-line el 9 de abril de 2024

BIBLIOGRAFÍA

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. Disponible en: <https://www.icmje.org/recommendations/>. Consultado 10 Ene 2024.

<https://doi.org/10.1016/j.recesp.2024.02.018>

0300-8932/© 2024 Sociedad Española de Cardiología. Publicado por Elsevier España, S.L.U. Todos los derechos reservados.