CONSIDERACIONES ÉTICAS

El estudio se atuvo a los principios indicados en la Declaración de Helsinki y fue aprobado por el comité ético local (Hospital Clínico Universitario de Valencia). Todos los pacientes otorgaron previamente el consentimiento informado. Se han tenido en cuenta las directrices SAGER.

DECLARACIÓN SOBRE EL USO DE INTELIGENCIA ARTIFICIAL

No se ha utilizado ninguna herramienta de inteligencia artificial en la preparación de este trabajo.

CONTRIBUCIÓN DE LOS AUTORES

Los autores no tienen ninguna otra financiación, relaciones económicas ni conflictos de intereses que declarar en relación con este trabajo.

CONFLICTO DE INTERESES

Ninguno.

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

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

* Autor para correspondencia.
Correo electrónico: yulnunez@gmail.com (J. Núñez).
𝕏 @Sandra_ViCo88 (S. Villar), @yulnunezvill (J. Núñez).

On-line el 22 de diciembre de 2023

BIBLIOGRAFÍA

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.2023.11.001

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

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

Sr. Editor:

ChatGPT, un modelo de lenguaje conversacional de inteligencia artificial, ha despertado grandes expectativas en todo el mundo debido a su sorprendente capacidad de responder de manera convincente a preguntas complejas formuladas utilizando lenguaje natural. Se ha utilizado en campos muy diversos, como la enseñanza, la programación informática y el periodismo, con resultados potencialmente revolucionarios. La comunidad médica no es ninguna excepción. ChatGPT ha sido capaz de superar con éxito exámenes necesarios para obtener la autorización para ejercer la medicina[1], redactar resúmenes científicos[2] y elaborar historias clínicas completas[3]. En cardiología, el bot ha podido ofrecer consejo clínico adecuado a pacientes imaginarios con cuadros clínicos cardiovasculares comunes[4] y ha superado a los estudiantes de medicina en exámenes estandarizados del área cardiovascular[5].

En vista de estos éxitos, existe una gran tentación de probar ChatGPT en la vida real como herramienta de soporte a la toma de decisiones basada en datos clínicos. Sin embargo, es importante preguntarse si ChatGPT puede procesar correctamente estas historias clínicas de la vida real y sugerir un tratamiento adecuado. La mayor parte de la bibliografía actual se centra en su aplicación en bases de datos «sintéticas» con textos muy pretratados y depurados, y/o respuestas de opción múltiple[1–6]. La fiabilidad de la vida real no se puede deducir directamente de esos entornos. Para responder a esta pregunta, se evaluó la concordancia entre ChatGPT y un equipo médico-quirúrgico de expertos formado por cardiólogos y cirujanos cardiacos en un caso práctico específico: el proceso de toma de decisiones en pacientes con estenosis aórtica grave.

Se realizó un análisis retrospectivo descriptivo de las historias clínicas de 50 pacientes consecutivos con estenosis aórtica que se presentaron en la sesión médico-quirúrgica cardiaca de este centro entre el 1 de enero de 2022 y el 14 de febrero de 2022 (se eligieron estas fechas para garantizar que la información sobre el tratamiento final de los pacientes estuviera disponible). Los profesionales asignaron a cada paciente a una de las siguientes opciones de tratamiento: *a)* reemplazo valvular quirúrgico; *b)* implante percutáneo de prótesis valvular, o *c)* tratamiento médico. Las estrategias de tratamiento decididas en la sesión médico-quirúrgica se compararon con las recomendadas por ChatGPT. Un cardiólogo elaboró un resumen completamente anonimizado del estado de cada paciente copiando las siguientes secciones de la historia clínica electrónica: características socio-demográficas, antecedentes médicos, ecocardiograma, coronariografía, síntomas y diagnóstico. Durante la segunda quincena de febrero de 2023, se introdujo toda esta información como un diálogo en ChatGPT (GPT-3.5, versión de 13 de febrero de 2023) asociada a una pregunta sobre el tratamiento óptimo. La pregunta se repetía 3 veces por paciente. Inicialmente, la pregunta era: «¿Cuál es el mejor tratamiento para este paciente?», pero las respuestas de ChatGPT fueron demasiado exhaustivas e incluían medicamentos e intervenciones para cualquier comorbilidad concurrente en el paciente en cuestión. Por tanto, la pregunta que se utilizó finalmente como entrada a ChatGPT en los experimentos fue: «¿Cuál es el mejor tratamiento para la estenosis aórtica del paciente que se presenta a continuación?» para obtener una respuesta más concreta que facilitara la interpretación, el

etiquetado, la clasificación y el tratamiento de los datos. No fueron necesarios más cambios en la pregunta planteada a ChatGPT para obtener respuestas útiles. Las recomendaciones devueltas por ChatGPT se codificaron como: *a)* cirugía; *b)* implante percutáneo de válvula aórtica (TAVI); *c)* tratamiento médico; *d)* intervención sin determinar (ChatGPT recomendó un reemplazo valvular aórtico, pero no especificó si el abordaje debía ser quirúrgico o percutáneo), o *e)* no concluyente. Los resultados se clasificaron según las siguientes definiciones:

- *Totalmente coherente:* las 3 respuestas recomendaron exactamente el mismo tratamiento.
- *Parcialmente coherente:* las 3 respuestas recomendaron un enfoque similar (intervención frente a tratamiento médico).
- *Acuerdo total:* respuesta totalmente coherente que coincidió con la recomendación de la sesión médico-quirúrgica.
- *Acuerdo en el enfoque:* respuesta total o parcialmente coherente que coincidió con la evaluación «intervención frente a tratamiento médico» de la sesión médico-quirúrgica.

La figura 1 muestra los resultados en detalle. La media de edad fue 78 años y el 41% eran hombres. La decisión de la sesión médico-quirúrgica fue TAVI en el 56%, cirugía en el 40% y tratamiento médico en el 4% de los casos. De las 150 respuestas generadas por ChatGPT, 14 (9%) no fueron concluyentes. Un total del 70% de las recomendaciones de ChatGPT fueron, al menos, parcialmente coherentes y el 38% fueron totalmente coherentes. Hubo *acuerdo en el enfoque* en el 58% de los casos, pero *acuerdo total* en solo el 18% de los casos. Fueron incoherentes 15 recomendaciones y 6 recomendaciones que eran coherentes fueron distintas de la decisión de la sesión médico-quirúrgica, lo que representa un total de 21 errores. De estos 21 casos, 10 (48%) presentaban otra valvulopatía o arteriopatía coronaria concomitante que requería intervención, 4 (19%) eran casos en los cuales las indicaciones de intervención tenían un menor nivel de evidencia y 7 (33%) eran casos de estenosis aórtica sintomática aislada. Sin embargo, cuando las recomendaciones eran totalmente coherentes, ChatGPT mostró, al menos, *acuerdo en el enfoque* en el 89% de los casos.

Este estudio tiene algunas limitaciones, como su carácter experimental y el pequeño tamaño de la muestra. Además, como experiencia de un solo centro, el criterio de referencia era la decisión de un equipo cardiológico en particular, que teóricamente podría tener sus propios sesgos.

Las sugerencias de tratamiento de ChatGPT coincidieron con las de los expertos médicos en el 58% de los casos. El acuerdo fue bajo en tratamientos específicos y moderado en la decisión de intervención frente al tratamiento médico. Como parece lógico, ChatGPT manifestó una tendencia a estar de acuerdo con la decisión del equipo cardiológico en los casos en que siempre ofreció respuestas similares a repeticiones de la misma pregunta. Sin embargo, el acuerdo y la coherencia se vieron sustancialmente reducidos en casos clínicamente complejos. Cabe señalar que estos resultados se obtuvieron utilizando un sistema diseñado únicamente como bot conversacional genérico, sin entrenamiento especializado, en un contexto muy exigente (pregunta abierta, enfermedad compleja). Los resultados podrían mejorarse notablemente con futuras versiones específicamente entrenadas para el apoyo a las decisiones médicas.

FINANCIACIÓN

Este trabajo no ha recibido ninguna financiación específica.

CONSIDERACIONES ÉTICAS

Debido al carácter retrospectivo de este trabajo y la anonimización completa de los datos, el trabajo se consideró exento de la obligación de obtener el consentimiento informado de los pacientes. El protocolo de investigación fue revisado y aprobado por el Comité de Ética de la Investigación con Medicamentos del Área de Salud Valladolid Este con el código PI 23-3194. El diseño del estudio, la metodología y los datos se analizaron en busca de signos de cualquier posible sesgo de género y no se encontró ninguno.

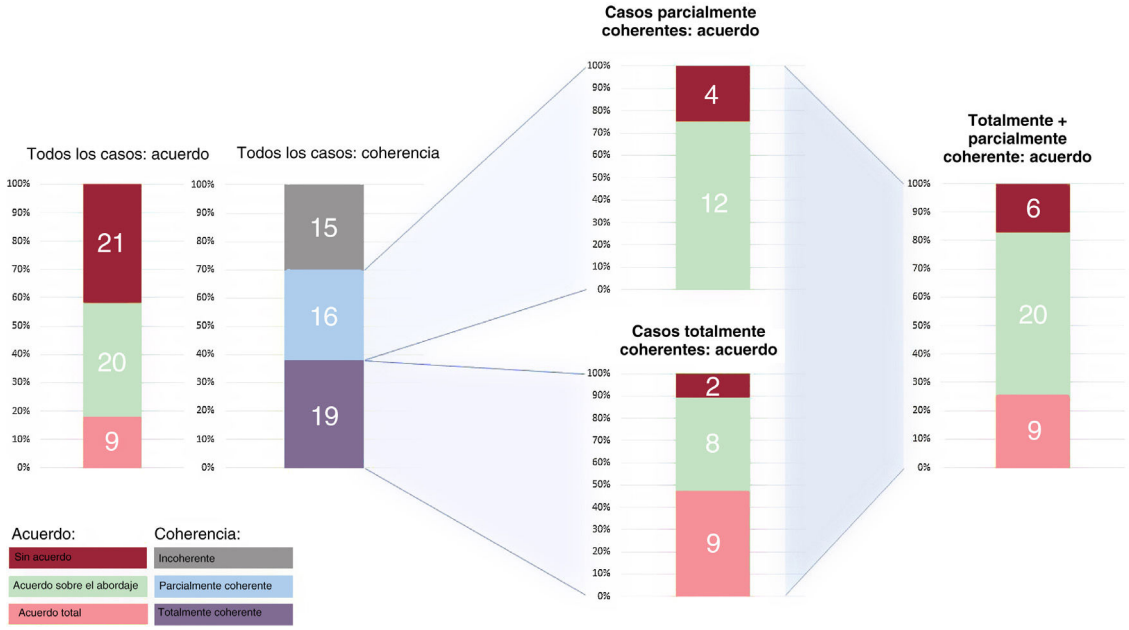

Figura 1. Coherencia y acuerdo de ChatGPT.

DECLARACIÓN SOBRE EL USO DE LA INTELIGENCIA ARTIFICIAL

Como se describe en el texto, ChatGPT 3.5 se utilizó en los experimentos de este trabajo como objeto de estudio. No se utilizó ninguna herramienta de IA para la redacción del original ni para el análisis de datos o la interpretación de los resultados.

CONTRIBUCIÓN DE LOS AUTORES

C. Baladrón, T. Sevilla y J.A. San Román diseñaron el estudio; C. Baladrón y T. Sevilla diseñaron y realizaron los experimentos. C. Baladrón y J.A. San Román redactaron el borrador inicial del original. M. Carrasco-Moraleja, I. Gómez-Salvador y J. Peral-Oliveira revisaron los datos y la metodología, crearon las figuras y realizaron el análisis de los datos. Todos los autores participaron en la revisión y corrección del artículo.

CONFLICTO DE INTERESES

Los autores no tienen conflictos de intereses relevantes que declarar.

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

[a]*Servicio de Cardiología, Hospital Clínico Universitario de Valladolid, Valladolid, España*

[b]*Centro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), España*

[*] Autor para correspondencia.
Correo electrónico: tereseru@gmail.com (T. Sevilla).
𝕏 @cbalzor (C. Baladrón) @TreSeru (T. Sevilla).

On-line el 12 de enero de 2024

BIBLIOGRAFÍA

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.
5. Hariri W. Analyzing the Performance of ChatGPT in Cardiology and Vascular Pathologies. *Research square preprints [preprint].* 2023. Disponible en: https://doi.org/10.21203/rs.3.rs-2782768/v1. Consultado 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.2023.11.014