Rev Clin Esp. 2024;224(Supl 1):S959

Revista Clínica Española

https://www.revclinesp.es

1812 - FIABILIDAD DE CHATGPT-4O PARA LA INTERPRETACIÓN DE LAS GUÍAS CLÍNICAS EN EMBOLIA PULMONAR

David Sergio Sánchez García[1], *Ramón Puchades Rincón de Arellano*[2], *Luis Ramos Ruperto*[3], *Yale Tung Chen*[2], *Giorgina Salgueiro Origlia*[2], *Teresa Sancho Bueso*[2], *Alicia Lorenzo Hernández*[2] *y Carmen Fernández Capitán*[2]

[1]*Hospital Universitario La Paz, Madrid, España.* [2]*Unidad de Enfermedad Tromboembólica, Hospital Universitario La Paz, Madrid, España.* [3]*Unidad VIH, Hospital Universitario La Paz, Madrid, España.*

Resumen

Objetivos: Analizar la precisión y concordancia de ChatGPT-4o para determinar la clase de recomendación y el nivel de evidencia en relación a una guía de práctica clínica de sobre diagnóstico y tratamiento de la embolia pulmonar.

Métodos: Se seleccionaron las recomendaciones de la Guía ESC 2019 para el diagnóstico, pronóstico y tratamiento de la embolia pulmonar aguda. A partir de las recomendaciones, se preguntó a ChatGPT-4o por el nivel de evidencia y la clase de recomendación que establecía la guía, registrando los aciertos y errores. Se realizó un analizó cualitativo de los errores (sobreestimación vs. infraestimación) de la recomendación, así como la clase y el nivel de evidencia donde se presentaron más frecuentemente.

Resultados: Se evaluaron un total de 56 recomendaciones. El número de errores globales en la determinación de la clase de recomendación y nivel de evidencia por ChatGPT-4o fueron 30 (53%). Respecto a la clase de recomendación, la determinó de forma errónea en 11 (20%) apartados, y en 26 (46%) el nivel de evidencia. En 7 (12%) apartados, el error se objetivó ambas. En cuanto a la dirección del error, en la clase de recomendación 8 (88%) fueron por sobreestimación y 1 (12%) por infraestimación; mientras que en el nivel de evidencia 11 (48%) fueron por sobreestimación y 12 (52%) por infraestimación. La distribución de errores por niveles de recomendación es: I (2/25, 8%), IIa (6/20, 30%), IIb (3/5, 60%), no encontrándose ninguno en la clase de recomendación III (0/6). Para el nivel de evidencia es: A (6/15, 40%), B (11/21, 52%) y C (9/20, 45%).

Errores ChatGPT-4o →				
Total ítems ↓	Error total	Clase de recomendación	Nivel de evidencia	Ambos
Diagnóstico (21)	15 (71%)	6 (28%)*	12 (57%)*	3 (14%)
Evaluación de la gravedad (5)	3 (60%)	2 (40%)	2 (40%)	1 (20%)
Tratamiento agudo (19)	8 (42%)	2 (10%)*	8 (42%)*	2 (10%)
Tratamiento crónico y prevención (10)	4 (40%)	1 (10%)	4 (40%)	1 (10%)
*p < 0,05.				

Discusión: La tasa de errores observada en este trabajo y próxima al 50%, es similar a la descrita en otros trabajos utilizando ChatGPT-4 para responder a preguntas de práctica clínica basadas en guías de Oncología[1]. En el "test de estrés" realizado para este trabajo, ChatGPT-4o no parece ser fiable actualmente como fuente de evidencia. En concreto, las recomendaciones dentro del apartado de diagnóstico presentaron fallos en 3/4 de las recomendaciones, si bien en general los errores se observaron en los niveles más bajos de la clase de recomendaciones y el nivel de evidencia. Por otra parte, se evidenció una tendencia a la sobreestimación recurrente la clase de recomendación, mientras en el nivel de evidencia, el porcentaje de sobreestimación e infraestimación fue similar de casos. Para que los sistemas de procesamiento de lenguaje puedan ser usados en la práctica clínica, tienen que ser capaces de dar información fiable, evitando "alucinaciones". La correcta identificación del nivel de evidencia y la clase recomendación por estos sistemas, resulta un aspecto crucial para ofrecer recomendaciones basadas en la evidencia.

Conclusiones: La utilidad de ChatGPT-4o para establecer recomendaciones basadas en la evidencia en el manejo de la embolia pulmonar aguda fue limitada. Es preciso mejorar el procesamiento e interpretación de las guías por ChatGPT-4o para obtener información válida y aplicable en la práctica clínica.

Bibliografía

1. Ferber D, Wiest IC, Wölflein G, N et al. GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines. NEJM AI, 2024;AIcs2300235.