Rev Esp Cardiol. 2024;77 (Supl 1): 1161

Revista Española de Cardiología

6075-469. RENDIMIENTO DE SISTEMAS DE CHAT ALIMENTADOS CON ARTÍCULOS DE INVESTIGACIÓN EN UN ENTORNO CLÍNICO ESPECÍFICO: LA ENFERMEDAD VALVULAR CARDIACA

Alain García Olea[1], Ane García Domingo-Aldama[2], Marcos Merino Prado[2], Ignacio Díez González[1], Aitziber Atutxa Salazar[2], Josu Goikoetxea Salutregi[2], Koldo Gojenola Galletebeitia[2], Mikel Maeztu Rada[1], Iván Cano González[1], Adrián Costa Santos[1], Iván García Díaz[1], Fernando Díaz González[1], Irene Hernández Pérez[1], Uxue Millet Oyarzabal[1] y José Miguel Ormaetxe Merodio[1]

[1]Hospital Universitario de Basurto, Bilbao (Vizcaya), España y [2]Escuela de Ingeniería, Departamento de Lenguajes y Sistemas Informáticos. Universidad del País Vasco, Bilbao (Vizcaya), España.

Resumen

Introducción y objetivos: La inteligencia artificial (IA) generativa ha experimentado un rápido crecimiento en el ámbito médico, destacando los *chatbots* basados en procesamiento del lenguaje natural (NLP) como herramientas prometedoras para la asistencia en decisiones clínicas. Sin embargo, la fiabilidad de estos modelos podría variar según su diseño y los datos con los que han sido entrenados. Este estudio explora cómo los chatbots responden a preguntas clínicas específicas sobre enfermedades valvulares cardiacas, considerando el efecto del entrenamiento de estos con información actualizada.

Métodos: El estudio comparó la fiabilidad de cuatro *chatbots* para responder preguntas clínicas sobre enfermedades valvulares cardiacas, utilizando el simulacro del Examen Europeo en Cardiología Básica (EECC) de 2023. Se seleccionaron dos *chatbots* de acceso libre no entrenados (Bing Chat y ChatGPT-3,5), junto a dos *chatbots* personalizables (Dante y Scispace Copilot), alimentados con las Guías ESC/EACTS 2021 sobre valvulopatías. Se enfrentaron a las nueve preguntas sobre valvulopatías del EECC 2023, se compararon sus respuestas con las oficiales, con las guías médicas y se solicitó justificación en las opciones discordantes. Finalmente, se categorizaron en error menor (opción menos correcta con argumentación lógica) o mayor (opción que condicionase un cambio en la atención clínica) las respuestas discordantes mediante evaluación independiente de las respuestas por dos cardiólogos.

Resultados: Una pregunta sobre valvulopatías excedía el contenido de las guías y todas presentaban casos clínicos con cinco opciones de respuesta. Bing Chat y ChatGPT-3,5 respondieron correctamente a 4 de 8 preguntas, mientras que los *chatbots* entrenados acertaron 3 de 8 (figura A). La inclusión en el análisis de la pregunta no tratada en las guías mejoró la puntuación de Bing Chat. En general, los modelos no entrenados tuvieron un mejor rendimiento que los personalizados. Tras analizar las justificaciones de las respuestas, los errores de Scispace Copilot y Bing Chat eran mayoritariamente menores.

Hoja de respuesta de chatbots *(A). Conversación con Scispace Copilot mostrando seguridad en la respuesta errónea (B).*

Conclusiones: La fiabilidad de los modelos de lenguaje analizados es limitada en este entorno clínico. La ambigüedad de las respuestas entre los *chatbots* y la seguridad en el discurso equivocado (figura B) subrayan la necesidad de una investigación concisa y responsable para implementar el uso de la IA en decisiones clínicas.