Endocrinología, Diabetes y Nutrición 71 (2024) S472

Endocrinología, Diabetes y Nutrición

417 - EVALUACIÓN DEL RENDIMIENTO DE CHATGPT-4 EN LA RESOLUCIÓN Y RAZONAMIENTO CLÍNICO DE PREGUNTAS DE ENDOCRINOLOGÍA Y NUTRICIÓN DEL EXAMEN MIR Y EIR 2024

L. Arsís García[1], I. Modrego Pardo[2] y C. Marco Alacid[1]

[1]Endocrinología y Nutrición, Hospital Verge dels Lliris, Alcoy. [2]Hospital Marina Baixa, Villajoyosa.

Resumen

Introducción: El desarrollo de la inteligencia artificial (IA) y el uso de ChatGPT se postula como una herramienta útil en la educación sanitaria e investigación científica. El objetivo del estudio fue evaluar la eficacia de ChatGPT-4 para responder preguntas de endocrinología y nutrición (EyN) de los exámenes MIR (médico interno residente) y EIR (enfermero interno residente) de 2024 y su capacidad para justificar su respuesta.

Métodos: Se incluyeron 14 preguntas correspondientes a EyN del examen MIR y 23 del EIR. Se evaluó el nivel de acierto de ChatGPT-4 y 3 evaluadores determinaron su razonamiento clínico. Se estudió la presencia de diferencias en cuanto a la dificultad propuesta por la IA frente a los evaluadores y se analizó si el formato de pregunta podía influir en el grado de acierto de la IA.

Resultados: ChatGPT-4 alcanzó un nivel de acierto en el MIR del 100% y en el EIR del 60,9%. El 42,9% de las respuestas del MIR se clasificaron por parte de los evaluadores como 'Razonamiento aceptable aunque incompleto' y un 57,1% como 'Razonamiento completamente correcto' y en el EIR un 60,9% como 'Razonamiento completamente correcto' y un 39,1% como 'Incorrecto o pobre'. No se encontró correlación entre el grado de dificultad de las preguntas MIR o EIR según ChatGPT y el evaluador MIR (correlación 0,357; $p > 0,05$) o EIR (correlación -0,028; $p > 0,05$) y tampoco la presencia de concordancia (Kappa 0,31; $p > 0,05$). No se objetivó que el tipo de pregunta, ni la habilidad evaluada, la disciplina, el abordaje o si la pregunta era formulada en negativo podía predecir el acierto o fallo de la pregunta ($p > 0,05$).

Conclusiones: ChatGPT-4 demostró ser una herramienta eficaz para responder preguntas de EyN correspondientes al examen MIR y, en menor medida, del EIR, con un adecuado razonamiento clínico. El rendimiento de la IA fue homogéneo independientemente del formato de pregunta o nivel de dificultad, aunque aún pueden existir errores en la interpretación por parte de la IA.