COMUNICACIÓN BREVE

Precisión de ChatGPT en el diagnóstico de entidades clínicas en el ámbito de la medicina interna

Accuracy of ChatGPT for the diagnosis of clinical entities in the field of internal medicine

Carlos A. Andrade-Castellanos,[1,2] Ma. Teresa Tapia-de la Paz[2] y Pedro E. Farfán-Flores[3]*

[1]Programa de Maestría en Educación en Ciencias de la Salud, Centro Universitario en Ciencias de la Salud, Universidad de Guadalajara; [2]Servicio de Medicina Interna. Hospital Civil de Guadalajara "Dr. Juan I. Menchaca"; [3]Coordinador de Posgrado, Centro Universitario de Ciencias de la Salud, Universidad de Guadalajara. Jalisco, México

ChatGPT es un modelo de inteligencia artificial (IA) diseñado para conversaciones. Su implementación en la resolución de dilemas clínicos abre nuevas posibilidades y permite a los médicos plantear casos clínicos y obtener respuestas en tiempo real.[1,2] También es útil en el proceso del diagnóstico diferencial, aunque es importante tener en cuenta los sesgos, como las alucinaciones.[3]

La IA debe utilizarse de manera responsable y ética en la educación médica. Los programas de formación deben ser diseñados y supervisados por educadores humanos, mientras que la IA debe ser utilizada como una herramienta complementaria, en lugar de ser considerada como un reemplazo de la interacción humana en el proceso educativo.[4]

El objetivo de este estudio fue evaluar la capacidad de ChatGPT en el diagnóstico de entidades clínicas en el ámbito de la medicina interna, para lo cual se utilizaron descripciones de casos. Se emplearon los casos del Medical Knowledge Self-Assessment Program (MKSAP), tal como se presentan en el sitio web de *ACP Internist Weekly*, del American College of Physicians, dentro de la sección *Test Yourself* (https://acpinternist.org/).[5] Estos casos están diseñados específicamente con fines educativos y han sido utilizados en el aprendizaje continuo desde 1968. Se copiaron los casos publicados desde el 19 de octubre de 2021 hasta el 11 de julio de 2023 directamente en ChatGPT versión 3.5 (https://chat.openai.com/), seguidos de dos preguntas: *What is the most likely diagnosis?* y *What is the differential diagnosis?* Excluimos aquellos que no implicaban dilemas diagnósticos, como los centrados en determinar el manejo más apropiado, y los que requerían imágenes para establecer un diagnóstico, según lo determinado por consenso.

El desenlace primario consistió en la coincidencia del diagnóstico principal de ChatGPT con el diagnóstico final del caso. Los desenlaces secundarios incluyeron la presencia del diagnóstico final en la lista diferencial de ChatGPT y la puntuación de calidad del diferencial obtenida con un sistema de clasificación ordinal de cinco puntos (previamente publicado),[6] el cual califica precisión y utilidad (se otorgan cinco puntos a una lista diferencial que incluye el diagnóstico exacto y cero puntos cuando no identifica diagnósticos cercanos). Todos los casos fueron evaluados de forma independiente por dos de los autores de este artículo y las discrepancias fueron resueltas por el tercero. Se realizó estadística descriptiva y se calculó el coeficiente kappa de Cohen para determinar la confiabilidad entre los evaluadores mediante el programa estadístico SPSS versión 21 (IBM).

De 110 casos evaluados, se excluyeron 100 (81 no constituyeron dilemas diagnósticos y 19 requerían imágenes para establecer el diagnóstico). El valor de kappa de Cohen estimado para la confiabilidad entre los evaluadores fue de 0.70 (acuerdo sustancial). Los detalles de cada caso y las respuestas completas generadas por ChatGPT están disponibles en el material

***Correspondencia:**
Carlos A. Andrade-Castellanos
E-mail: caandrade@hcg.gob.mx

Fecha de recepción: 18-07-2023
Fecha de aceptación: 06-09-2023
DOI: 10.24875/GMM.23000297

Gac Med Mex. 2023;159:452-455
Disponible en PubMed
www.gacetamedicademexico.com

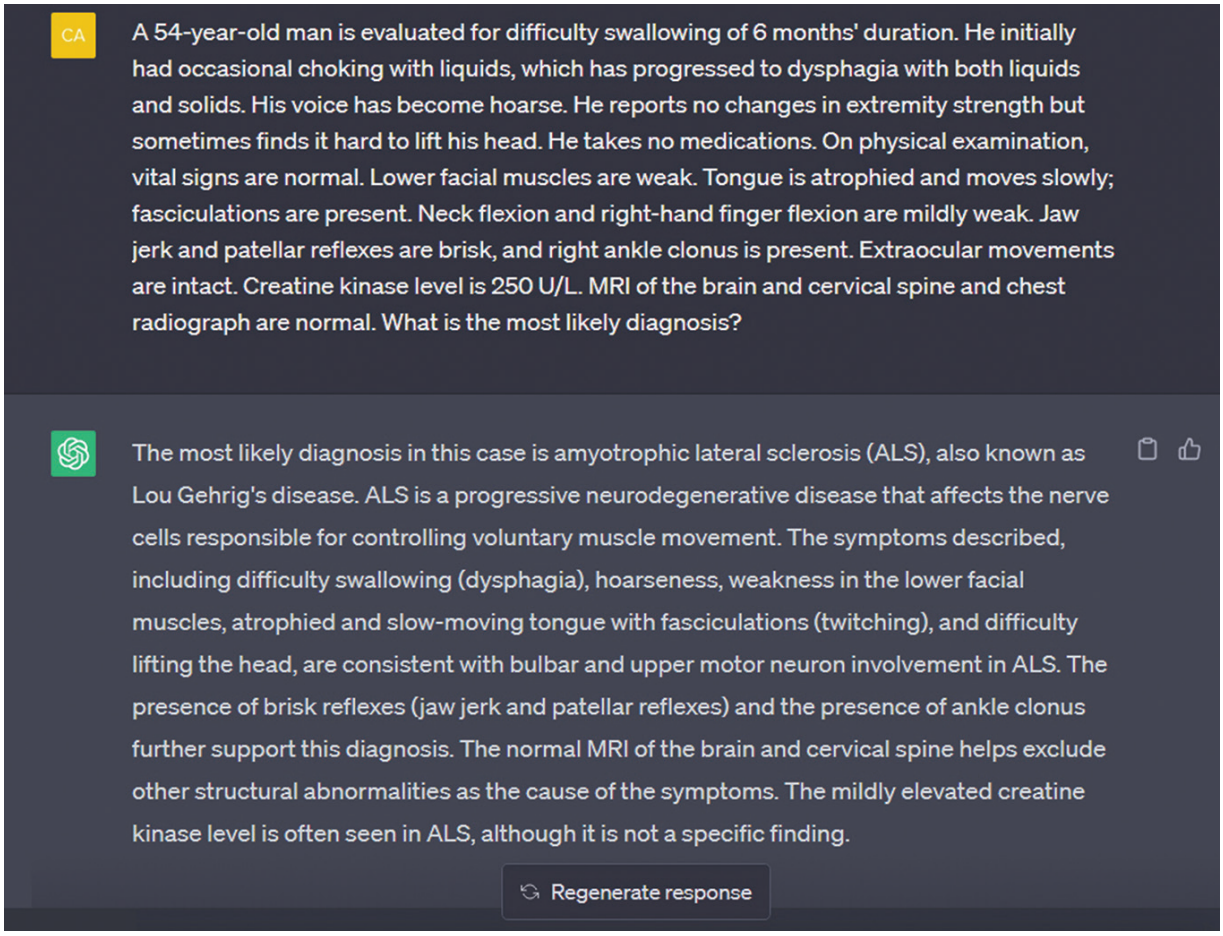

Figura 1. *Ejemplo del texto de entrada proporcionado a ChatGPT con su correspondiente resultado. El modelo de inteligencia artificial realizó el diagnóstico correcto y proporcionó información adicional relacionada con la condición en el contexto de este paciente.*

suplementario de este artículo. En la Figura 1 se muestra un ejemplo de un caso junto con la respuesta generada por ChatGPT. El diagnóstico principal del modelo de IA coincidió con el diagnóstico final en 70 % de los casos y el modelo incluyó el diagnóstico final en su lista diferencial en 80 % de los casos (Tabla 1). Cuando ChatGPT proporcionó el diagnóstico correcto en su lista diferencial, la posición media del diagnóstico fue de 1.5 ± 1.06 y la media de calidad del diagnóstico diferencial fue de 4.4 ± 1.07.

El modelo generativo de IA ChatGPT demostró un alto porcentaje de aciertos en el diagnóstico. Además, logró incluir el diagnóstico final en su lista diferencial en un elevado número de casos. Estos resultados son comparables con estudios previos que evaluaron el desempeño de ChatGPT.[7-9] Recientemente se evaluó su rendimiento utilizando casos clinicopatológicos del *New England Journal of Medicine* (NEJM). El modelo identificó el diagnóstico correcto en 39 % de los casos y el diagnóstico final en su lista diferencial en 64 %. La calidad media de los diagnósticos diferenciales obtenidos fue de 4.2, ligeramente inferior en comparación con la obtenida en nuestro estudio.[10] Es importante resaltar que los casos del NEJM suelen ser detallados, exhaustivos y a menudo requieren apoyo radiológico.

La tecnología de IA conversacional presenta limitaciones significativas. Una de ellas es su capacidad para generar respuestas que suenan plausibles, pero que son incorrectas desde el punto de vista factual. Además, estos modelos pueden ser sensibles a la formulación de la entrada o *prompt* utilizada para generar una respuesta; es probable que una descripción insatisfactoria, carente de precisión o redactada en un idioma distinto derive en resultados insatisfactorios.[11]

Es factible emplear ChatGPT con fines clínicos. No obstante, en el actual estadio de desarrollo de la IA, estas asistencias son todavía concebidas

Tabla 1. Diagnósticos provisionales y diferenciales de 10 casos formulados por ChatGPT

Categoría del caso	Diagnóstico principal de ChatGPT	Diagnóstico final (MKSAP 19)	Diagnóstico diferencial de ChatGPT
Medicina cardiovascular (pregunta 75 MKSAP 19)	Corazón de atleta	Corazón de atleta	1. Corazón de atleta 2. Miocardiopatía hipertrófica 3. Bradicardia sinusal fisiológica 4. Repolarización precoz 5. Otras miocardiopatías
Reumatología (pregunta 72 MKSAP 19)	Síndrome nefrótico	Amiloidosis renal	1. Síndrome nefrótico 2. Amiloidosis 3. Enfermedad de cambios mínimos 4. Glomeruloesclerosis focal y segmentaria 5. Trombosis de la vena renal
Neumología y cuidados intensivos (pegunta 59 MKSAP 19)	Fibrosis quística	Fibrosis quística	1. Fibrosis quística 2. Discinesia ciliar primaria 3. Aspergilosis broncopulmonar alérgica 4. Bronquiectasias 5. Enfermedad pulmonar obstructiva crónica
Enfermedades infecciosas (pregunta 35 MKSAP 19)	Síndrome de Guillain-Barré	Síndrome de Guillain-Barré secundario a infección por *Campylobacter*	1. Síndrome Guillain-Barré 2. Polineuropatía desmielinizante inflamatoria crónica 3. Compresión del cordón espinal 4. Esclerosis múltiple 5. Mielitis trasversa
Neurología (pregunta 48 MKSAP 19)	Esclerosis lateral amiotrófica	Esclerosis lateral amiotrófica	1. Esclerosis lateral amiotrófica 2. Parálisis bulbar 3. Miastenia gravis 4. Enfermedad de la motoneurona asociada a demencia frontotemporal 5. Infarto del tronco encefálico
Nefrología (pregunta 73 MKSAP 19)	Poliquistosis renal autosómica dominante	Enfermedad de membrana basal delgada	1. Enfermedad de membrana basal delgada 2. Síndrome de Alport 3. Nefropatía IgA 4. Hematuria inducida por ejercicio 5. Otras causas raras
Gastroenterología y hepatología (pregunta 56 MKSAP 19)	Disfagia orofaríngea relacionada con enfermedad de Parkinson	Disfagia orofaríngea	1. Disfagia orofaríngea relacionada con enfermedad de Parkinson 2. Efecto adverso a fármacos 3. Síndrome Parkinson plus 4. Enfermedad por reflujo gastroesofágico 5. Causas estructurales
Enfermedades infecciosas (pregunta 3 MKSAP 19)	Psitacosis	Infección por *Chlamydia psittaci*	1. Neumonía 2. Bronquitis aguda 3. Tuberculosis pulmonar 4. Embolismo pulmonar 5. Lesión por inhalación 6. Enfermedad pulmonar intersticial
Neurología (pregunta 16 MKSAP 19)	Encefalopatía traumática crónica	Encefalopatía traumática crónica	1. Enfermedad de Alzheimer 2. Demencia frontotemporal 3. Enfermedad de Parkinson 4. Encefalopatía traumática crónica 5. Demencia vascular 6. Hidrocefalia normotensiva
Nefrología (pregunta 24 MKSAP 19)	Nefritis intersticial aguda	Nefritis intersticial crónica	1. Lesión renal aguda 2. Enfermedad renal crónica 3. Glomerulonefritis 4. Nefrolitiasis 5. Enfermedades sistémicas

MKSAP 19: Medical Knowledge Self-Assessment Program, en su versión 19 (lanzado el 31 de enero de 2022).

como un copiloto en el proceso de diagnóstico. Por otro lado, su aplicación como herramienta de apoyo educativo es viable y podría ser considerada como una posible "zona de desarrollo próximo", conforme la concepción de Vygostky. Al proporcionar información clínica razonada, ChatGPT puede ayudar a los estudiantes a desarrollar esquemas que faciliten la asimilación y la acomodación de aprendizajes significativos (enfoque basado en problemas). Esta tecnología posee relevancia para las generaciones actuales y las venideras, lo que conlleva la necesidad de redefinir los enfoques educativos con el propósito de abordar sus requerimientos y expectativas de manera adecuada.

Financiamiento

Los autores declaran no haber recibido financiación para este estudio.

Conflicto de intereses

Los autores declaran no tener conflicto de intereses.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que para esta investigación no se realizaron experimentos en seres humanos ni en animales.

Confidencialidad de los datos. Los autores declaran que en este artículo no aparecen datos de pacientes.

Derecho a la privacidad y consentimiento informado. Los autores declaran que en este artículo no aparecen datos de pacientes.

Uso de inteligencia artificial para generar textos. Los autores declaran que no han utilizado ningún tipo de inteligencia artificial generativa en la redacción de este manuscrito ni para la creación de figuras, gráficos, tablas o sus correspondientes pies o leyendas.

Material suplementario

El material suplementario se encuentra disponible en DOI: 10.24875/GMM.23000297. Este material es provisto por el autor de correspondencia y publicado *online* para el beneficio del lector. El contenido del material suplementario es responsabilidad única de los autores.

Bibliografía

1. Lanzagorta-Ortega D, Carrillo-Pérez DL, Carrillo-Esper R. Inteligencia artificial en medicina: presente y futuro. Gac Med Mex. 2022;158(Supl.1):55-9. DOI: 10.24875/GMM.M22000688.
2. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023;25:e48568. DOI: 10.2196/48568.
3. Vidal-Ledo M, Diego-Olite F, Armenteros-Vera I, Morales-Suárez I, Acosta-Domínguez A, Pérez-Pedro J. Chat en la educación médica. Educación Médica Superior [Internet]. 2023 [Citado 2023 Jul 14];37(2):e3879. Disponible en: https://ems.sld.cu/index.php/ems/article/view/3879
4. Palencia-Díaz R, Palencia-Vizcarra RJ. El potencial de la inteligencia artificial para disminuir errores médicos y mejorar la educación médica continua. Med Int Mex. 2023;39(3):419-21. DOI: 10.24245/mim.v39i3.8934
5. ACP Internist Weekly [Internet]. Estados Unidos: Test Yourself. American College of Physicians. Disponible en: https://acpinternist.org
6. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med. 2012;27(2):213-9. DOI: 10.1007/s11606-011-1804-8
7. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv. 2023:2023.02.21.23285886. DOI: 10.1101/2023.02.21.23285886
8. Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. JFO Open Opthalmology. 2023;1:100005. DOI: 10.1016/j.jfop.2023.100005
9. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by Generative Pretrained Transformer 3 Chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health. 2023;20(4):3378. DOI: 10.3390/ijerph20043378
10. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78-80. DOI: 10.1001/jama.2023.8288
11. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. arXiv [Internet]. 2021 [Citado 2023 Jul 16];arXiv:2107.13586. Disponible en: http://arxiv.org/abs/2107.13586.