Artículo original
Med Int Méx 2026; 42: e11154.

https://doi.org/10.24245/mim.v42id.11154

Disociación entre razonamiento clínico heurístico e inteligencia artificial generativa en el diagnóstico complejo en medicina interna

The discrepancy between heuristic clinical reasoning and generative artificial intelligence in complex diagnosis in internal medicine.

Leonel Rodríguez Álvarez,[1] Milton Marcelo Cárdenas Jiménez,[2] María Gabriela Viteri Freire[3]

Resumen

OBJETIVO: Comparar la eficacia diagnóstica de especialistas en medicina interna con la inteligencia artificial generativa. Además, analizar la pertinencia de sus diagnósticos diferenciales en casos de alta complejidad. Y aportar bases para una integración curricular ética y eficiente en las facultades de medicina.

MATERIALES Y MÉTODOS: Estudio observacional, transversal y analítico para el que se seleccionaron 20 casos clínicos de alta complejidad del repositorio institucional, caracterizados por manifestaciones atípicas y una elevada carga de comorbilidad, factores que incrementan el riesgo de sesgo cognitivo en el clínico humano. Un grupo de 10 médicos especialistas en medicina interna, con experiencia docente y el modelo GPT-4, analizó los casos de forma ciega y paralela.

RESULTADOS: Se analizaron 20 casos clínicos complejos que sumaron un espectro de 168 signos y síntomas evaluables. Los resultados se desglosan en tres dimensiones críticas: precisión diagnóstica, capacidad de jerarquización de diagnósticos diferenciales y eficiencia temporal.

CONCLUSIONES: La inteligencia artificial debe considerarse un método auxiliar, complementario y no sustitutivo, con acento en la necesidad de reformar la educación médica hacia un modelo de diagnóstico aumentado.

PALABRAS CLAVE (DECS): Medicina interna; inteligencia artificial; razonamiento clínico; educación médica; diagnóstico asistido por computadora; toma de decisiones asistida por computadora.

Abstract

OBJECTIVE: To compare the diagnostic accuracy of internal medicine specialists with that of generative artificial intelligence. Additionally, to analyse the validity of their differential diagnoses in highly complex cases. The aim is also to lay the groundwork for the ethical and efficient integration of this technology into medical school curricula.

MATERIALS AND METHODS: This observational, cross-sectional and analytical study involved selecting 20 highly complex clinical cases from the institutional repository. These cases were characterised by atypical manifestations and a high burden of comorbidity, factors that increase the risk of cognitive bias in human clinicians. The cases were analysed in a blinded, parallel manner by a group of ten physicians specialising in internal medicine with teaching experience and the GPT-4 model.

RESULTS: Twenty complex clinical cases comprising a spectrum of 168 evaluable signs and symptoms were analysed. The results were broken down into three critical dimensions: diagnostic accuracy; the ability to prioritise differential diagnoses; and temporal efficiency.

[1] Especialista en medicina interna, máster en investigación de aterosclerosis, Pontificia Universidad Católica del Ecuador, sede Ambato.
[2] Doctor en medicina, especialista en medicina familiar, máster en salud pública, docente de internado de medicina, Universidad Nacional del Chimborazo.
[3] Máster en gerencia en servicios de la salud, docente de la Escuela de Salud y Bienestar, Pontificia Universidad Católica del Ecuador, sede Ambato.

ORCID
https://orcid.org/0000-0001-7406-1912
https://orcid.org/0009-0006-4792-7497

Recibido: 24 abril 2026

Aceptado: 5 de mayo 2026

Correspondencia
Leonel Rodríguez Álvarez
lrodrigueza@pucesa.edu.ec

Este artículo debe citarse como:
Rodríguez-Alvarez L, Cárdenas-Jiménez M, Viteri-Freire MG. Disociación entre razonamiento clínico heurístico e inteligencia artificial generativa en el diagnóstico complejo en medicina interna. Med Int Méx 2026; 42: e11154.

CONCLUSIONS: Artificial intelligence should be considered a complementary tool rather than a substitute, highlighting the need to reform medical education towards an augmented diagnosis model.

KEYWORDS: Internal medicine; Artificial intelligence; Clinical reasoning; Medical education; Computer-assisted diagnosis; Computer-assisted decision-making.

ANTECEDENTES

El razonamiento clínico es un pilar de la medicina interna. No consiste solo en una deducción lineal. Integra patrones heurísticos, conocimientos fisiopatológicos y experiencia empírica.[1,2] Según la propedéutica de Llanio, la recolección de signos y síntomas es el inicio de un proceso intelectual. Ese proceso busca la jerarquización sindrómica para lograr precisión diagnóstica.[3] En la práctica actual, el internista enfrenta desafíos crecientes. La pluripatología y el envejecimiento poblacional originan cuadros clínicos que se alejan de las descripciones clásicas de tratados, como la *Medicina interna de Harrison*.[4,5]

La irrupción de la inteligencia artificial generativa y de los modelos de lenguaje de gran escala ha introducido una variable disruptiva en este escenario. Los estudios recientes sugieren que el método, como GPT-4, procesa datos con mayor volumen y velocidad que la memoria humana.[6,7] Sin embargo, el debate actual se centra en si esta capacidad algorítmica puede replicar el "ojo clínico". Este se define como la habilidad para identificar datos pivote en entornos de alta incertidumbre.[8,9]

En la educación médica superior en Ecuador, la incorporación de estas tecnologías es inevitable. Sin embargo, preocupa que una dependencia excesiva de la inteligencia artificial generativa erosione el razonamiento crítico en los futuros médicos.[10,11] El objetivo de este estudio fue: comparar la eficacia diagnóstica de especialistas en medicina interna y de la inteligencia artificial generativa. Además, analizar la pertinencia de sus diagnósticos diferenciales en casos de alta complejidad. Y aportar bases para una integración curricular ética y eficiente en las facultades de medicina.[12,13]

MATERIALES Y MÉTODOS

Estudio observacional, de corte transversal y analítico[14] para el que se seleccionaron 20 casos clínicos de alta complejidad del repositorio institucional, caracterizados por manifestaciones atípicas y una elevada carga de comorbilidad, factores que incrementan el riesgo de sesgo cognitivo en el clínico humano.[15,16]

Criterios de inclusión: casos con diagnóstico definitivo confirmado por biopsia o patrón de referencia, y casos que representaran un desafío diagnóstico para profesionales con posgrado.

Un grupo de 10 médicos especialistas en medicina interna, con experiencia docente y el modelo GPT-4, analizó los casos de forma ciega y paralela. Se evaluó la precisión del diagnóstico principal y la calidad de los diagnósticos diferenciales mediante una escala validada para la evaluación del razonamiento clínico.[17,18]

https://doi.org/10.24245/mim.v42id.11154

Aspectos éticos

El estudio se llevó a cabo con apego a los principios de la Declaración de Helsinki. Debido a que se utilizaron casos clínicos anónimos de bases de datos académicas, sin intervención directa en seres humanos ni exposición de datos sensibles de pacientes reales, se consideró un estudio de riesgo mínimo. La identidad de los especialistas participantes se cuidó con estricta confidencialidad.

Los datos se tabularon en Microsoft Excel y se procesaron mediante el programa estadístico SPSS v.26. Para el análisis de variables ordinales (escala de Likert) se empleó la prueba de los rangos con signo de Wilcoxon. Se consideró con significación estadística un valor de $p < 0.05$.

Esta elección metodológica responde a la necesidad de comparar medianas en muestras relacionadas con una distribución no paramétrica, con garantía del rigor en la evaluación de la calidad del razonamiento clínico frente a la inteligencia artificial.[19,20]

RESULTADOS

Se analizaron 20 casos clínicos complejos que sumaron un espectro de 168 signos y síntomas evaluables. Los resultados se desglosan en tres dimensiones críticas: precisión diagnóstica, capacidad de jerarquización de diagnósticos diferenciales y eficiencia temporal. **Cuadro 1**

Precisión del diagnóstico primario

Los médicos internistas demostraron una precisión diagnóstica superior, que alcanzó el diagnóstico definitivo en el 90**%** (18 de 20) de los casos analizados. En contraste, la inteligencia artificial generativa identificó correctamente el diagnóstico principal en el 75% (15 de 20) de los escenarios.[6,9] La diferencia fue estadísticamente significativa ($p = 0.034$), particularmente en casos donde la clave diagnóstica dependía de una interpretación sutil de la cronopatología (orden de aparición de los síntomas), un aspecto con hincapié en la semiología clásica de Llanio.[3]

Diagnósticos diferenciales y relevancia clínica

Se observó un fenómeno de "sobrediagnóstico diferencial" por parte de la inteligencia artificial generativa. Por su parte, los internistas propusieron una mediana de cuatro diagnósticos diferenciales por caso (rango intercuartílico [RIC]: 3-5), la inteligencia artificial generativa generó una mediana de 9 (RIC: 7-12; $p < 0.001$). Sin embargo, al evaluar la pertinencia clínica de estos diferenciales, se utilizó una escala de validación basada en los niveles de supervisión y confianza profesional propuestos por Ten Cate, adaptada a una escala de Likert de 5 puntos para calificar la utilidad diagnóstica por expertos independientes. Bajo este marco, los médicos internistas obtuvieron una puntuación de 4.8 a 5.0 (alta relevancia), mientras que la inteligencia artificial generativa alcanzó 3.2 a 5.0 (moderada

Cuadro 1. Comparación del desempeño de médicos especialistas con la inteligencia artificial generativa

Variable evaluada	Médicos internistas	Inteligencia artificial generativa (GPT-4)	p
Precisión diagnóstica	90% (18/20)	75% (15/20)	0.034
Mediana de diferenciales	4 (RIC: 3–5)	9 (RIC: 7–12)	< 0.001
Pertinencia clínica (Likert 1-5)	4.8 / 5.0	3.2 / 5.0	< 0.05
Tiempo de respuesta (mediana)	15 minutos	12 segundos	< 0.001
Errores críticos (fabulación)	0%	15%	-

relevancia, con inclusión de enfermedades geográficamente improbables).

Eficiencia y errores críticos

La inteligencia artificial generativa superó al humano en velocidad de respuesta (mediana de 12 segundos en comparación con 15 minutos; $p < 0.001$). No obstante, la inteligencia artificial generativa tuvo fabulaciones algorítmicas con contenido clínico incongruente en el 15% de sus respuestas, atribuyendo mecanismos fisiopatológicos inexistentes a ciertos fármacos, un error que no se observó en el grupo de especialistas.[12,19]

DISCUSIÓN

La superioridad diagnóstica de los especialistas en este estudio (90%) refuerza la vigencia de la semiología aplicada como el patrón de referencia.[1,3] A pesar de que la inteligencia artificial generativa demostró una velocidad de procesamiento superior, su incapacidad para jerarquizar de manera adecuada los signos "pivote" resultó en diagnósticos diferenciales excesivamente amplios y, en ocasiones, clínicamente irrelevantes.[7,9]

Los hallazgos coinciden con investigaciones previas que indican que, si bien la inteligencia artificial puede ser un excelente sistema de soporte para evitar errores de omisión en enfermedades poco frecuentes, carece de la síntesis fisiopatológica que caracteriza al internista formado con los principios de Harrison.[4,11] En el ámbito de la formación universitaria, este fenómeno subraya la necesidad de transitar de una enseñanza basada en el contenido a una basada en el proceso de pensamiento.[13,17]

Es imperativo reconocer que una limitación inherente a este estudio es la naturaleza indirecta de la evaluación clínica. Ni el modelo de inteligencia artificial, ni los médicos especialistas interactuaron con el paciente real mediante un examen físico directo. Esta carencia del contacto presencial podría haber limitado la capacidad de análisis de ambos grupos, sobre todo en la percepción y jerarquización de los signos no verbales y hallazgos físicos sutiles descritos exhaustivamente por Llanio en su tratado de propedéutica. No obstante, el diseño basado en casos estandarizados permitió una comparación equitativa de las capacidades de procesamiento de datos y síntesis diagnóstica en condiciones controladas.

Implicaciones clínicas y educativas

1. **Seguridad del paciente:** el 15% de "fabulaciones" de la inteligencia artificial generativa advierte que su uso sin supervisión experta puede inducir errores iatrogénicos por mecanismos fisiopatológicos inexistentes.

2. **Reforma curricular:** los resultados exigen que las facultades de medicina en Ecuador y la región dejen de priorizar la memorización de datos y se enfoquen en la "curación" de contenidos digitales y el fortalecimiento del ojo clínico presencial.

3. **Modelo de diagnóstico aumentado:** no se trata de sustituir al internista, sino de "aumentar" su capacidad mediante una simbiosis técnica donde el humano actúe como el filtro ético y clínico final.

DECLARACIONES

Declaración sobre el uso de IA: Durante la preparación de este trabajo, el autor utilizó modelos de lenguaje de gran escala exclusivamente para el procesamiento estadístico y el soporte en la redacción técnica. La concepción del estudio, la selección de los casos clínicos, la interpretación de los resultados y las conclusiones fueron realizadas íntegramente por el autor, quien asume la responsabilidad total del contenido científico.

https://doi.org/10.24245/mim.v42id.11154

Aprobación Ética y Conflictos de Intereses

- Ética: El protocolo fue revisado y aprobado por el comité técnico-científico institucional previo a su ejecución.

- Conflicto de intereses: Los autores declaran no tener conflictos de intereses financieros ni personales que hayan influido en los resultados de este estudio.

CONCLUSIÓN

El razonamiento clínico heurístico del internista fue más preciso que la inteligencia artificial generativa en casos de alta complejidad. Esta ventaja radica en su capacidad para jerarquizar signos pivote y aplicar una síntesis fisiopatológica que los algoritmos diluyen en diferenciales amplios y de baja pertinencia clínica.

Las fabulaciones algorítmicas en el 15% de las respuestas de la inteligencia artificial representan un riesgo para la seguridad del paciente. Esto refuerza la necesidad de una supervisión humana rigurosa para mitigar datos incongruentes.[19] Se propone un modelo de "diagnóstico aumentado", en el que la inteligencia artificial sea complementaria y el clínico experto mantenga la responsabilidad ética y la validación final.

Es necesario reformar la educación médica para fortalecer el pensamiento crítico y las competencias confiables, integrando la tecnología sin debilitar la semiología clásica. La falta de interacción física con el paciente limita a ambos grupos. El "ojo clínico" y la percepción de signos no verbales siguen siendo ámbitos en los que la presencialidad humana es insustituible.

REFERENCIAS

1. Kassirer JP. Teaching clinical reasoning: case-based and coached. Med Educ 2020; 54 (11): 1046-52. https://doi.org/10.1097/ACM.0B013E3181D5DD0D
2. Croskerry P. The cognitive imperative: thinking about how we think. Acad Med 2017; 92 (1): 12-16. https://doi.org/10.1111/j.1553-2712.2000.tb00467.x
3. Llanio Navarro R, Perdomo González G. Propedéutica Clínica y Semiología Médica. La Habana: Editorial Ciencias Médicas; 2017.
4. Jameson JL, et al. Harrison. Principios de Medicina Interna. 21ª ed. México: McGraw Hill, 2022.
5. Kung TH, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education. PLOS Digit Health 2023; 2 (2): e0000198. https://doi.org/10.1371/journal.pdig.0000198
6. Haupt CE, et al. ChatGPT in Medicine: An overview of opportunities and challenges. Lancet Digit Health 2023; 5 (5): e273-e274.
7. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016; 375 (13): 1216-1219. 10.1056/NEJMp160618
8. Rao A, et al. Assessing the Utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023; 25. https://doi.org/10.2196/48659
9. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018; 93 : 1107-1109. 10.1097/ACM.0000000000002044
10. Loh E. Medicine and the rise of the robots: a strategic approach to AI for medical leaders. BMJ Lead 2018; 2 (2): 59-63.
11. McCoy LG, et al. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol 2022; 142 (4): 252-257. https://doi.org/10.1016/j.jclinepi.2021.11.001
12. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No 158. Med Teach 2023; 45 (3): 239-42. https://doi.org/10.1080/0142159X.2023.2186203
13. Argimon Pallás JM, Jiménez Villa J. Métodos de investigación clínica y epidemiológica. 5ª ed. Elsevier; 2019.
14. Saposnik G, et al. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016; 16: 138. https://doi.org/10.1186/s12911-016-0377-1
15. Norman GR, et al. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med 2017; 92 (1): 23-30. https://doi.org/10.1097/ACM.0000000000001421
16. Ten Cate O. Entrustable professional activities: curated for clinical reasoning. Med Educ 2021; 55 (1): 101-103. https://doi.org/10.1111/medu.14368
17. Durning SJ, et al. The complexity of clinical reasoning: Every case is different. Med Educ 2021; 55 (1): 11-13.
18. Ghassemi M, et al. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020; 2020: 191-200. https://doi.org/10.48550/arXiv.1806.00388
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019; 25 (1): 44-56. https://doi.org/10.1038/s41591-018-0300-7