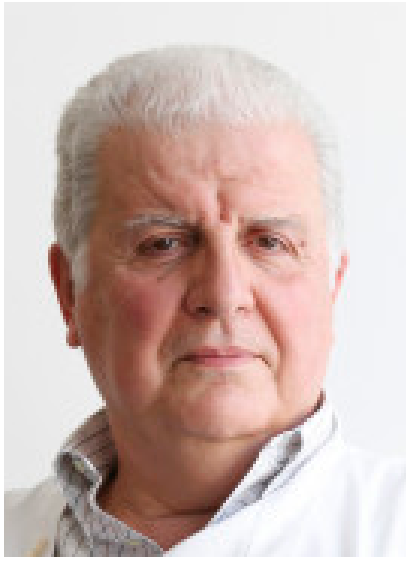

Inteligência Artificial em Medicina e a Replicação do Raciocínio Clínico: O Início de Uma Nova Era

Artificial Intelligence in Medicine and the Replication of Clinical Reasoning: The Dawn of a New Era

António Vaz Carneiro

DOI: https://doi.org/10.29315/gm.1259

INTRODUÇÃO AO RACIOCÍNIO CLÍNICO

O raciocínio clínico (RC) é o processo cognitivo através do qual os médicos integram informação de múltiplas fontes para diagnosticar, tratar, monitorizar e acompanhar os seus doentes (ou pacientes, no caso da prevenção de doença).

Para se poder compreender a emulação do RC por inteligência artificial (IA), é preciso compreender que existem vários tipos de RC (por exemplo, reconhecimento de padrões, abordagens fisiopatológicas, terapêuticas, de monitorização, etc.). Neste artigo iremos concentrar-nos apenas no raciocínio diagnóstico indutivo-dedutivo (afinal, o mais importante).

A anamnese é a base deste processo. Uma anamnese bem estruturada — abrangendo a queixa principal, o seu início e evolução, sintomas associados, antecedentes médicos passados, uso de medicação e contexto familiar e social — gera o conjunto inicial de hipóteses de diagnóstico, frequentemente designado por diagnóstico diferencial. Estudos sugerem que a anamnese, por si só, contribui para o diagnóstico correto na grande maioria dos casos, salientando a sua importância mesmo antes da recolha de quaisquer dados físicos ou laboratoriais.[1]

O exame físico serve, então, para confirmar, refinar ou refutar estas hipóteses. Achados como os sinais vitais, a auscultação, a palpação e manobras neurológicas ou musculoesqueléticas específicas, fornecem corroboração objetiva das queixas subjetivas. O exame físico é mais valioso quando orientado pelas hipóteses geradas durante a anamnese, em vez de ser realizado como uma rotina sem foco.

Os exames laboratoriais acrescentam informações quantitativas e, muitas vezes, mais específicas. A sua interpretação requer a compreensão de medidas de associação como por ex. sensibilidade, especificidade, valores preditivos positivos ou negativos, *likelihood ratios*, etc.

Coordenador, Conselho para a Inteligência Artificial em Medicina da Ordem dos Médicos, Lisboa, portugal. Presidente, Instituto de Saúde Baseada na Evidência, (FMUL/FFUL), Lisboa, Portugal

Recebido/Received: 2026-08-28. Aceite/Accepted: 2026-08-28. Publicado/Published: 2026-09-14.

Pelo seu outro lado, os exames de imagem ampliam a capacidade diagnóstica, ao visualizar anormalidades anatómicas ou funcionais não acessíveis apenas pela anamnese ou exame físico. No entanto, os achados imagiológicos devem ser interpretados juntamente com o contexto clínico: achados incidentais não relacionados com a queixa principal são comuns e podem induzir em erro, se não forem adequadamente contextualizados.[2]

À medida que as ferramentas de apoio à decisão clínica e a inteligência artificial auxiliam cada vez mais neste processo, a compreensão da estrutura subjacente do raciocínio clínico continua a ser essencial para que os médicos avaliem criticamente, em vez de aceitarem passivamente, as sugestões algorítmicas na procura de precisão diagnóstica.

COMPONENTES DO RACIOCÍNIO CLÍNICO

A discussão anterior destina-se a definir e compreender a complexidade do RC, como base de entendimento comparativo entre o que o médico faz *versus* o que um algoritmo de IA faz.

Os componentes do RC estão descritos na tabela.

COMPONENTES DO RACIOCÍNIO CLÍNICO	
Geração de hipóteses	Formulação em segundos de uma lista concisa de diagnósticos prováveis após a audição da queixa principal, com base no reconhecimento de padrões (*scripts* de doença)
Recolha de dados	Anamnese e dados do exame físico, escolhidos para diferenciar entre as hipóteses, sem serem feitos de forma exaustiva
Refinamento de hipóteses	Atualização das probabilidades à medida que novos dados estão disponíveis
Verificação diagnóstica	Comparar a hipótese principal com a evidência, considerando alternativas que não se podem falhar
Metacognição	Saber quando se está razoavelmente certo, quando passar do Sistema 1 (noviço) para o Sistema 2 (perito) e vice-versa

EMULAÇÃO DO RACIOCÍNIO CLÍNICO POR INTELIGÊNCIA ARTIFICIAL

Os sistemas de inteligência artificial, particularmente os grandes modelos de linguagem (LLMs), podem presentemente emular o raciocínio clínico, integrando informações heterogéneas, gerando diagnósticos diferenciais e produzindo respostas racionais de forma comparável à cognição médica (ainda que persistam limitações significativas no raciocínio diagnóstico inicial e na gestão da incerteza).[3]

É, portanto, fundamental investigar comparativamente os desempenhos dos médicos directamente com os algoritmos nos mesmos casos clínicos, já que existe evidência de boa qualidade que demonstra a existência de um conjunto de problemas que vão da ignorância de dados clínicos relevantes à existência de vieses e invenções.[4]

Naquele que é provavelmente o melhor estudo à data sobre o tema, PG Brodeur e colaboradores publicaram um artigo comparando médicos com o modelo o1 da OpenAI em tarefas de raciocínio clínico, abordando uma lacuna de longa data na investigação em IA aplicada à medicina: a falta de bases robustas de comparação com o desempenho humano.[5]

Em seis experiências, o modelo o1-*preview*/o1 foi comparado com centenas de médicos e gerações anteriores de IA de apoio ao diagnóstico. Quando aplicado aos casos das conferências clínico-patológicas (CPCs) do NEJM (casos clínicos complexos), o o1-*preview* atingiu o diagnóstico correcto em 78,3% dos 136 casos, e 97,9% quando foram incluídos diagnósticos quase correctos, tendo superado tanto uma base histórica de médicos, como o GPT-4 (88,6% *vs* 72,9% de diagnósticos exactos/muito próximos, P =0,015). O algoritmo seleccionou ainda correctamente o teste de diagnóstico seguinte em 87,5% dos casos. Nos casos de doentes virtuais do *NEJM Healer*, o o1-*preview* alcançou pontuações perfeitas no teste de raciocínio *Revised-IDEA* em 78 de 80 casos, superando significativamente o GPT-4 e os médicos assistentes em diversos graus da carreira. Nos casos de gestão "Grey Matters" (casos reais para testar raciocínio de gestão clínica) e nos casos de diagnóstico "landmark" (estudo clássico de sistemas de diagnóstico computorizados),[6] o o1-preview obteve pontuações substancialmente mais elevadas do que o GPT-4 e os médicos (com ou sem ajuda de IA), embora as diferenças nos casos "landmark" tenham sido apenas marginalmente significativas. No raciocínio probabilístico diagnóstico (estimativa de probabilidade pré-teste/pós-teste), o o1-preview teve um desempenho semelhante ao do GPT-4 e outros), apresentando ambos menor variabilidade do que os médicos humanos (figura adaptada do artigo, com permissão).

Foi ainda incluído um estudo ocultado com 76 casos práticos de urgência demonstrando que o o1 superou

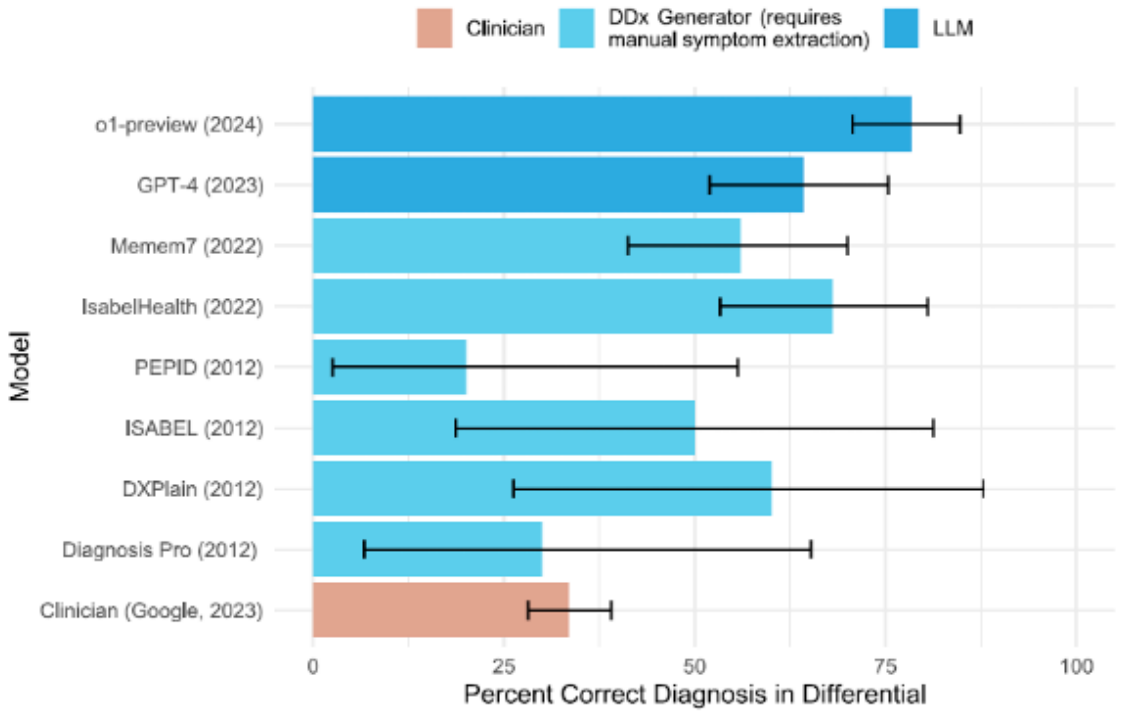

FIGURA 1. **Performance of differential diagnosis generators and LLMs on NEJM clinicopathologic conferences (CPCs), 2012 to 2024.** Bar plot showing the accuracy of including the correct diagnosis in the differential for differential diagnosis (DDx) generators and LLMs on the NEJM CPCs, sorted by year. Data for other LLMs or DDx generators were obtained from the literature (materials and methods). The 95% CIs were computed with a one-sample binomial test.

dois médicos assistentes e o GPT-4o em três pontos de contacto diagnóstico (rastreio, consulta médica e internamento), sendo a vantagem mais acentuada no rastreio inicial, quando a informação disponível era mais escassa. De realçar que os médicos raramente conseguiram distinguir os diagnósticos diferenciais gerados pela IA, dos gerados por humanos.

Os autores concluem que os grandes modelos de linguagem (LLMs) "ultrapassaram agora a maioria dos parâmetros de referência do raciocínio clínico", mas alertam que o seu estudo se limita a entradas textuais, se centra em contextos de medicina interna/urgência e reflecte uma versão de modelo por definição já obsoleta. Defendem a realização de ensaios clínicos prospetivos e o desenvolvimento de estruturas mais robustas para a integração segura do apoio ao diagnóstico por IA nos fluxos de trabalho clínicos reais, sublinhando que a precisão diagnóstica, por si só, não garante uma implementação segura e eficaz.[7]

CONCLUSÕES

- A introdução de uma nova tecnologia na saúde (como é a IA) levanta desafios importantes, que importa analisar e resolver.[8]
- As potenciais utilizações na saúde são imensas, desde áreas clínicas (apoio para diagnósticos diferenciais por ex.) até suporte à decisão terapêutica, ensino médico, apoio administrativo, investigação básica/clínica, etc.[9]
- As consequências de uma implementação pouco cuidadosa de uma tecnologia destas abre a porta a potenciais complicações a todos os níveis dos sistemas de saúde, até porque existem dúvidas sobre o seu impacto global.[10]
- A utilização da IA em medicina abre caminho a três hipóteses com impacto na prática clínica e que devem estar conscientemente presentes em todos nós:
  - ***Deskilling:*** Perda de uma competência - raciocínio ou técnica - previamente adquirida e dominada, por desuso decorrente da delegação sistemática da tarefa à IA
  - ***Mis-skilling:*** Desenvolvimento de competências ou raciocínios distorcidos ou inadequados
  - ***Never-skilling:*** Não aquisição *ab initio* da competência de base, porque a IA esteve presente desde o início da formação, não dando aso à auto-aprendizagem).

- A maioria das avaliações testa a capacidade de reconhecer padrões em informações completas e pré-definidas do caso clínico em análise. Pelo seu lado, o raciocínio clínico real é iterativo, incerto e envolve decidir que informação procurar a seguir - uma tarefa cognitiva fundamentalmente diferente de responder a um caso totalmente especificado.[11]
- O RC da IA deverá estar aberto a análises específicas, com eliminação das "caixas-negras" dos programas de IA disponíveis.[12]
- Constrangimentos éticos e morais devem ser tidos em conta com rigor.[13]
- Para maximizar o impacto e responder se a IA está a melhorar a assistência médica ou não, devemos submetê-la aos mesmos padrões de análise científica que outras intervenções clínicas bem estabelecidas: não basta demonstrar precisão ou viabilidade e mostrar potencial para que seja melhor. Os resultados é que justificam (ou não) a adoção da tecnologia.[14]
- Todos os aspectos éticos/legais, regulatórios, de protecção de dados deverão ser levados em linha de conta na utilização desta tecnologia.[15]
- Uma das abordagens mais sólidas – designada por Medical Holistic Evaluation of Language Models (Med-HELM) – define cinco domínios em estudo: apoio à decisão clínica, geração de notas médicas, comunicação com o doente, assistência à investigação e fluxos de trabalho administrativos.[16]
- A promessa da IA na assistência médica será cumprida quando tivermos evidência de que a IA altera significativamente as decisões clínicas e que essas decisões melhoram os resultados dos doentes/pacientes.

RESPONSABILIDADES ÉTICAS

CONFLITOS DE INTERESSE: Os autores declaram a inexistência de conflitos de interesse.

APOIO FINANCEIRO: Este trabalho não recebeu qualquer subsídio, bolsa ou financiamento.

PROVENIÊNCIA E REVISÃO POR PARES: Solicitado; sem revisão externa por pares.

ETHICAL DISCLOSURES

CONFLICTS OF INTEREST: The authors have no conflicts of interest to declare.

FINANCIAL SUPPORT: This work has not received any contribution grant or scholarship.

PROVENANCE AND PEER REVIEW: Commissioned; without external peer review.

REFERÊNCIAS

1. Sox HC, Higgins MC, Owens DK, Sanders Schmidler G. Medical decision making. 3rd ed. Hoboken: Wiley Blackwell; 2024.
2. Kherad O, Carneiro AV; Choosing Wisely Working Group of the European Federation of Internal Medicine. Ten year anniversary of choosing wisely campaigns. Eur J InternMed. 2022;103:118-9. doi: 10.1016/j.ejim.2022.05.028.
3. Levi M. Artificial intelligence in diagnostics and medical decision making. Eur J Intern Med. 2026;4:107056. doi: 10.1016/j.ejim.2026.107056.
4. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med. 2024;177:210-20. doi: 10.7326/M23-2772.
5. Brodeur PG, Buckley TA, Kanjee Z, Goh E, Ling EB, Jain P, et al. Performance of a large language model on the reasoning tasks of a physician. Science. 2026;392:524-7. doi: 10.1126/science.adz4433.
6. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330:1792-6. doi: 10.1056/NEJM199406233302506.
7. Hopkins AM, Cornelisse E. AI can reason like a physician--what comes next? Science. 2026;392:466-7. doi: 10.1126/science.aeg8766.
8. Kleinman RA, Torous J, Danilewitz M. Use of Large-Language Models for Therapy: Promise and Perils. Ann Intern Med. 2026;179:600-1. doi: 10.7326/ANNALS-25-04914.
9. Carneiro VA. Large Language Models in Medicine. JSCMed 2025;169:8-11 - doi: 10.57849/ulisboa.fm.jscml.0000024.2025
10. Goldenberg A, Wiens J. Is AI actually improving healthcare? Nat Med. 2026;32:1182-1183. doi: 10.1038/s41591-026-04329-2.
11. Chen SF, Alyakin A, Seas A, Yang E, Choi JJ, Lee JV, et al. LLM-assisted systematic review of large language models in clinical medicine. Nat Med. 2026;32:1152-9. doi: 10.1038/s41591-026-04229-5.
12. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, et al. AI, Health, and Health Care Today and Tomorrow. The JAMA Summit Report on Artificial Intelligence. JAMA. 2025;334:1650-64. doi: 10.1001/jama.2025.18490
13. Conselho Nacional de Ética para as Ciências da Vida. Inteligência artificial (IA): inquietações sociais, propostas éticas e orientações políticas: livro branco. Lisboa: CNECV; 2024.
14. Palmieri S, Robertson CT, Cohen IG. New Guidance on Responsible Use of AI. JAMA. 2026;335:207-8. doi: 10.1001/jama.2025.23059.
15. Kaplan A, Haenlein M. Sri, Siri in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons.2019;62:15-25. doi: 10.1016/j.bushor.2018.08.04
16. Bedi S, Cui H, Fuentes M, Unell A, Wornow M, Banda JM, et al. Holistic evaluation of large language models for medical tasks with MedHELM. Nat Med. 2026;32:943-51. doi: 10.1038/s41591-025-04151-2.