

PESQUISA ORIGINAL IDENTIFICADO

DESEMPENHO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL NO RACIOCÍNIO DIAGNÓSTICO DA SÍNDROME DEMENCIAL: ESTUDO COMPARATIVO COM VINHETAS CLÍNICAS PADRONIZADAS TÍPICAS E ATÍPICAS

PERFORMANCE OF ARTIFICIAL INTELLIGENCE MODELS IN THE DIAGNOSTIC REASONING OF DEMENTIA SYNDROME: A COMPARATIVE STUDY USING STANDARDIZED TYPICAL AND ATYPICAL CLINICAL VIGNETTES

Magno Conceição das Mercês¹, Alana Pimentel Costa²

¹ Professor do curso de Medicina do Centro Universitário Dom Pedro II – UNIDOMPEDRO Afya, Salvador, Bahia, Brasil.

² Acadêmica do curso de Medicina do Centro Universitário Dom Pedro II – UNIDOMPEDRO Afya, Salvador, Bahia, Brasil. E-mail do autor principal: alanapimentel25@gmail.com

Eixo Temático: Inovação, Tecnologia, Sustentabilidade e Transformação Global

RESUMO

Introdução: A utilização de modelos de inteligência artificial generativa como ferramentas de apoio ao raciocínio clínico tem crescido na Medicina, porém ainda existem lacunas quanto ao seu desempenho diante de apresentações clínicas atípicas. **Objetivo:** Avaliar e comparar o desempenho de cinco modelos de inteligência artificial no raciocínio diagnóstico sintomático e etiológico da síndrome demencial, considerando apresentações clássicas e atípicas da Doença de Alzheimer. **Método:** Foi realizado estudo metodológico, observacional, transversal e comparativo, utilizando dez vinhetas clínicas padronizadas, sendo cinco clássicas e cinco atípicas. As vinhetas foram apresentadas a cinco modelos de inteligência artificial por meio de prompt padronizado. Foram analisadas 50 respostas quanto à identificação da síndrome, etiologia, diagnósticos diferenciais, raciocínio clínico, exames complementares e clareza, com escore máximo de 30 pontos. **Resultados:** A acurácia geral para identificação da síndrome demencial foi de 96,0%, enquanto a identificação etiológica alcançou 78,0%. Nos casos clássicos, a acurácia etiológica foi de 96%, em comparação a 60% nos casos atípicos, com diferença estatisticamente significativa. O escore médio foi de $26,88 \pm 2,98$ nos casos clássicos e $25,16 \pm 2,81$ nos atípicos. Claude e Gemini apresentaram os maiores percentuais de classificação correta e maiores escores médios. Todos os modelos erraram a etiologia de uma apresentação atípica que mimetizava demência frontotemporal. **Conclusão:** Os modelos avaliados apresentaram desempenho superior diante de apresentações clássicas, especialmente na identificação etiológica, demonstrando vulnerabilidade diante de fenótipos atípicos. Os resultados reforçam a necessidade de supervisão humana especializada para utilização dessas ferramentas no raciocínio diagnóstico neurológico.

Palavras-chave: Inteligência Artificial. Neurologia. Doença de Alzheimer. Raciocínio Diagnóstico. Demência.

ABSTRACT

Introduction: The use of generative artificial intelligence models as tools to support clinical reasoning has increased in Medicine; however, gaps remain regarding their performance in atypical clinical presentations.

Objective: To evaluate and compare the performance of five artificial intelligence models in syndromic and etiological diagnostic reasoning of dementia syndrome, considering typical and atypical presentations of Alzheimer's disease. **Method:** A methodological, observational, cross-sectional, comparative study was conducted using ten standardized clinical vignettes, five typical and five atypical. The vignettes were presented to five artificial intelligence models using a standardized prompt. Fifty responses were analyzed regarding syndrome identification, etiology, differential diagnoses, clinical reasoning, complementary examinations, and clarity, with a maximum score of 30 points. **Results:** Overall accuracy for identifying dementia syndrome was 96.0%, whereas etiological identification reached 78.0%. In typical cases, etiological accuracy was 96%, compared with 60% in atypical cases, with a statistically significant difference. The mean score was 26.88 ± 2.98 in typical cases and 25.16 ± 2.81 in atypical cases. Claude and Gemini showed the highest percentages of correct classifications and the highest mean scores. All models incorrectly identified the etiology of an atypical presentation mimicking frontotemporal dementia. **Conclusion:** The evaluated models showed superior performance in typical presentations, particularly in etiological identification, demonstrating vulnerability to atypical phenotypes. These findings reinforce the need for specialized human supervision when using these tools in neurological diagnostic reasoning.

Keywords: Artificial Intelligence. Neurology. Alzheimer Disease. Diagnostic Reasoning. Dementia.

INTRODUÇÃO

A utilização de modelos de inteligência artificial (IA) generativa como apoio ao raciocínio clínico tem se expandido na prática médica, sendo a Neurologia um campo especialmente relevante pela necessidade de integrar anamnese, exame clínico e conhecimento sindrômico-etiológico. A Doença de Alzheimer, principal causa de síndrome demencial, apresenta importante heterogeneidade fenotípica: além da apresentação amnésica clássica, fenótipos atípicos podem envolver linguagem, comportamento e função visuoespacial.¹⁻⁶ A maioria das avaliações de desempenho de IA em diagnóstico médico utiliza casos típicos, existindo lacuna quanto à confiabilidade desses modelos diante de apresentações atípicas, nas quais o diagnóstico diferencial é mais desafiador.⁷⁻⁸ Assim, o presente estudo teve como objetivo avaliar e comparar o desempenho de cinco modelos de IA no raciocínio diagnóstico sindrômico e etiológico da síndrome demencial, a partir de vinhetas clínicas padronizadas representando apresentações clássicas e atípicas da Doença de Alzheimer.

MÉTODO

Foi realizado estudo metodológico, observacional, transversal e comparativo. Foram elaboradas dez vinhetas clínicas padronizadas: cinco clássicas, construídas a partir de diretrizes

internacionais para diagnóstico da Doença de Alzheimer,^{1-3,9-10} e cinco atípicas, construídas a partir de relatos e séries de casos publicados na literatura, selecionados por apresentarem manifestações incomuns.⁴⁻⁸ Todas as vinhetas foram elaboradas especificamente para o estudo, sem correspondência integral a pacientes reais, contemplando identificação, história clínica, exame físico e neurológico, avaliação cognitiva, exames laboratoriais, neuroimagem e biomarcadores, quando aplicável.

Cada vinheta foi apresentada integralmente a cinco modelos de IA (ChatGPT GPT-5.5, Gemini, Claude, Microsoft Copilot e Perplexity AI) por meio de prompt padronizado, solicitando síndrome, etiologia mais provável, até três diagnósticos diferenciais, raciocínio clínico, exames complementares e grau de confiança. Cada resposta foi classificada em correta, parcialmente correta ou incorreta, e recebeu escore de 0 a 30 pontos, distribuído entre seis componentes (síndrome, etiologia, diagnósticos diferenciais, raciocínio clínico, exames complementares e clareza). Variáveis categóricas foram comparadas pelo teste do qui-quadrado com correção de Yates e teste exato de Fisher; o escore total foi comparado pelo teste de Mann-Whitney, adotando-se $\alpha = 0,05$, com auxílio das bibliotecas Python NumPy e SciPy.

RESULTADOS E DISCUSSÃO

Foram avaliadas 50 respostas (25 de casos clássicos e 25 de casos atípicos). A acurácia geral para identificação da síndrome demencial foi de 96,0% (48/50; IC95%: 86,5–98,9%) e a acurácia etiológica geral foi de 78,0% (39/50; IC95%: 64,8–87,2%), com 74,0% de respostas corretas, 22,0% parcialmente corretas e 4,0% incorretas. O escore médio geral foi de $26,02 \pm 2,99$ pontos (Tabela 1).

Tabela 1 – Desempenho dos modelos de inteligência artificial nas dez vinhetas clínicas

Modelo	N	Síndrome correta	Etiologia correta	Classificação correta	Escore médio (DP)
ChatGPT GPT-5.5	10	100%	70%	70%	$25,5 \pm 3,3$
Gemini	10	100%	80%	80%	$27,4 \pm 2,4$
Claude	10	100%	80%	80%	$27,8 \pm 2,1$
Microsoft Copilot	10	100%	80%	80%	$24,5 \pm 2,9$
Perplexity AI	10	80%	80%	60%	$24,9 \pm 3,1$

Fonte: Elaborado pelos autores, 2026.

Claude e Gemini apresentaram os maiores percentuais de classificação correta (80%) e os maiores escores médios (27,8 e 27,4, respectivamente); a Perplexity AI foi o único modelo a errar a síndrome clínica (dois casos clássicos).

A análise por tipo de apresentação revelou diferença relevante: a acurácia etiológica foi de 96% nos casos clássicos e de 60% nos atípicos ($\chi^2 = 7,459$; $p = 0,0063$; odds ratio = 16,0; $p = 0,0046$), e o escore médio de qualidade foi maior nos casos clássicos ($26,88 \pm 2,98$) do que nos atípicos ($25,16 \pm 2,81$; Mann-Whitney U = 439,0; $p = 0,0137$), embora a identificação sindrômica tenha sido numericamente superior nos casos atípicos (100% vs. 92%) (Tabela 2).

Tabela 2 – Desempenho dos modelos de inteligência artificial segundo o tipo de caso

Tipo de caso	N	Síndrome correta	Etiologia correta	Classificação correta	Escore médio (DP)
Clássico	25	92%	96%	88%	$26,88 \pm 2,98$
Atípico	25	100%	60%	60%	$25,16 \pm 2,81$

Fonte: Elaborado pelos autores, 2026.

O achado mais expressivo ocorreu em um caso atípico com desinibição, apatia, irritabilidade e redução da fluência verbal, associado à preservação inicial da memória: todos os cinco modelos identificaram a síndrome demencial, porém atribuíram a etiologia a degeneração lobar frontotemporal, e quatro declararam confiança alta apesar do erro. Esse resultado corrobora a literatura sobre heterogeneidade fenotípica da Doença de Alzheimer e sobreposição clínica com outras síndromes neurodegenerativas,⁴⁻⁸ sugerindo que os modelos tendem a associar o fenótipo predominante à etiologia mais compatível com aquele padrão, mesmo quando atípica. A confiança elevada acompanhando respostas incorretas reforça que esse parâmetro não deve ser interpretado isoladamente como medida de acurácia diagnóstica, podendo favorecer ancoragem diagnóstica no uso clínico dessas ferramentas.^{1-3,9-10} Entre as limitações do estudo, destacam-se o tamanho amostral reduzido (dez vinhetas) e a impossibilidade de generalização para outras síndromes neurológicas.

CONCLUSÃO

Os modelos de IA avaliados apresentaram elevado desempenho na identificação da síndrome demencial (96,0%), porém desempenho etiológico significativamente inferior diante de apresentações atípicas da Doença de Alzheimer (60% vs. 96% nos casos clássicos). Os resultados indicam que essas ferramentas podem auxiliar tarefas estruturadas de raciocínio diagnóstico, mas não substituem a avaliação clínica especializada, sobretudo diante de fenótipos atípicos, nos quais o risco de ancoragem diagnóstica é maior, reforçando-se a necessidade de supervisão humana especializada.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Jack CR Jr, Andrews JS, Beach TG, Buracchio T, Carrillo MC, Dunn B, et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimers Dement*. 2024;20(8):5143-5169. doi:10.1002/alz.13859.
2. Atri A, Dickerson BC, Clevenger CK, et al. Alzheimer's Association clinical practice guideline for the Diagnostic Evaluation, Testing, Counseling, and Disclosure of Suspected Alzheimer's Disease and Related Disorders (DETeCD-ADRD): executive summary of recommendations for primary care. *Alzheimers Dement*. 2025;21(6):e14333. doi:10.1002/alz.14333.
3. Dickerson BC, et al. The Alzheimer's Association clinical practice guideline for the Diagnostic Evaluation, Testing, Counseling, and Disclosure of Suspected Alzheimer's Disease and Related Disorders (DETeCD-ADRD): executive summary of recommendations for specialty care. *Alzheimers Dement*. 2025;21(1):e14337. doi:10.1002/alz.14337.
4. Galton CJ, Patterson K, Xuereb JH, Hodges JR. Atypical and typical presentations of Alzheimer's disease: a clinical, neuropsychological, neuroimaging and pathological study of 13 cases. *Brain*. 2000;123(3):484-498. doi:10.1093/brain/123.3.484.
5. Gallucci M, et al. More atypical than atypical Alzheimer's disease phenotypes: a Treviso Dementia (TREDEM) Registry case report. *J Alzheimers Dis Rep*. 2021;5(1):365-374. doi:10.3233/ADR-210009.
6. Zhu L, et al. Case of early-onset Alzheimer's disease with atypical manifestation. *Gen Psychiatr*. 2021;34:e100283. doi:10.1136/gpsych-2020-100283.
7. Zajicek K, et al. An atypical presentation of autopsy confirmed rapidly progressive early onset Alzheimer's disease with borderline CSF biomarkers: a case study. *Alzheimers Dement*. 2026;21(Suppl 2):e107697. doi:10.1002/alz70856_107697.
8. Namrouti A, et al. Atypical early onset Alzheimer's disease in a young female: a case report. *Cureus*. 2025;17(1):e78082. doi:10.7759/cureus.78082.
9. Fabrizi E, et al. The Italian guideline on diagnosis and treatment of dementia and mild cognitive impairment. *Age Ageing*. 2024;53(11):afae250. doi:10.1093/ageing/afae250.

10. Ismail Z, et al. Recommendations of the 5th Canadian Consensus Conference on the diagnosis and treatment of dementia. *Alzheimers Dement.* 2020;16(8):1180-1195. doi:10.1002/alz.12105.
11. Palmqvist S, et al. Clinical practice guideline for blood-based biomarkers in the diagnostic workup of Alzheimer's disease within specialized care settings: a report from the Alzheimer's Association. *Alzheimers Dement.* 2025;21(Suppl 7):e108894. doi:10.1002/alz70861_108894.