

REVISÃO DE LITERATURA IDENTIFICADO

INTELIGÊNCIA ARTIFICIAL NA AVALIAÇÃO E NO FEEDBACK DE HABILIDADES CLÍNICAS EM EXAMES CLÍNICOS OBJETIVOS ESTRUTURADOS (OSCE): UMA REVISÃO INTEGRATIVA

ARTIFICIAL INTELLIGENCE IN THE ASSESSMENT AND FEEDBACK OF CLINICAL SKILLS IN OBJECTIVE STRUCTURED CLINICAL EXAMINATIONS (OSCE): AN INTEGRATIVE REVIEW

Gabriel Henrique Carvalho Leite Romeiro^{1*}, Vitor Comassetto Paes¹, Larissa Leite de Souza Leão¹, Luciano Timbó Barbosa², Jessyca Lane Fausto Lira², Reinaldo dos Santos Moura²

¹Acadêmicos de Medicina, Afya Centro Universitário de Maceió (UNIMA), Maceió, Alagoas, Brasil; ²Docente do Curso de Medicina, Afya Centro Universitário de Maceió (UNIMA), Maceió, Alagoas, Brasil.

*Autor principal. E-mail: gabriel.romeiro@alunos.afya.com.br

Eixo: 03 - Tecnologia e Inovação

RESUMO

O Exame Clínico Objetivo Estruturado (OSCE) é padrão-ouro para avaliação de habilidades clínicas na formação médica, mas enfrenta limitações de custo, escalabilidade e variabilidade entre avaliadores. O avanço de IA generativa, processamento de linguagem natural (PLN) e visão computacional tem motivado seu uso experimental como pacientes simulados, avaliadores ou geradores de feedback. Objetivo: sintetizar as evidências sobre aplicações de IA na avaliação e no feedback de habilidades clínicas em OSCE e simulação estruturada. Método: revisão integrativa em bases indexadas (PubMed/MEDLINE, Europe PMC, periódicos JMIR), priorizando estudos publicados entre 2016 e 2025. Resultados: foram incluídos 17 estudos em quatro eixos: IA generativa como paciente simulado e avaliador; PLN para feedback de anamnese; reconhecimento de fala e comunicação não verbal; e visão computacional para habilidades procedurais. Modelos de linguagem de grande porte alcançaram pontuações comparáveis ou superiores às de candidatos humanos em domínios cognitivos, enquanto PLN e reconhecimento de fala mostraram-se úteis para feedback de anamnese, com validação heterogênea. Visão computacional demonstrou desempenho promissor em habilidades procedurais, com pouca transposição para estações clínicas gerais. Conclusão: a IA apresenta potencial crescente para apoiar avaliação e feedback, com evidência mais consistente em domínios cognitivos e técnicos do que em competências relacionais, sendo necessária validação prospectiva multicêntrica antes de incorporação a avaliações somativas de alto risco.

Palavras-chave: Inteligência Artificial; Avaliação Educacional; Competência Clínica; Educação Médica.

ABSTRACT

The Objective Structured Clinical Examination (OSCE) is the gold standard for assessing clinical skills in medical education, but faces limitations of cost, scalability, and inter-rater variability. The advance of generative AI, natural

language processing (NLP), and computer vision has motivated its experimental use as simulated patients, evaluators, or feedback generators. Objective: to synthesize the evidence on AI applications in the assessment and feedback of clinical skills in OSCE and structured simulation. Method: integrative review in indexed databases (PubMed/MEDLINE, Europe PMC, JMIR journals), prioritizing studies published between 2016 and 2025. Results: 17 studies were included across four domains: generative AI as simulated patient and evaluator; NLP for history-taking feedback; speech recognition and non-verbal communication; and computer vision for procedural skills. Large language models achieved scores comparable to or higher than human candidates in cognitive domains, while NLP and speech recognition proved useful for structured history-taking feedback, albeit with heterogeneous validation. Computer vision showed promising performance in procedural skills, with little transposition to general clinical stations. Conclusion: AI shows growing potential to support assessment and feedback, with more consistent evidence in cognitive and technical domains than in relational competencies, requiring multicenter prospective validation before incorporation into high-stakes summative assessments.

Keywords: Artificial Intelligence; Educational Measurement; Clinical Competence; Medical Education.

1. INTRODUÇÃO

O Exame Clínico Objetivo Estruturado (OSCE) consolidou-se como principal instrumento de avaliação de competências clínicas na formação médica, por permitir observação padronizada de anamnese, exame físico, raciocínio clínico e comunicação em estações simuladas¹. Apesar da ampla adoção, o método é intensivo em recursos e exige treinamento extenso, estando sujeito a variabilidade interobservador que compromete a confiabilidade da nota final¹.

Paralelamente, o avanço de modelos de IA generativa baseados em grandes modelos de linguagem (LLM) ampliou as possibilidades de automatizar etapas do processo avaliativo em educação médica². Sistemas dessa natureza já foram testados como pacientes simulados conversacionais capazes de gerar feedback automatizado sobre a anamnese³.

Ainda assim, a literatura é dispersa entre subcampos que raramente dialogam entre si: LLM generativos como avaliadores, PLN aplicado a feedback, reconhecimento de fala, e visão computacional para habilidades procedurais. Não há síntese que integre essas quatro frentes no contexto de OSCE.

2. OBJETIVO

Sintetizar as evidências disponíveis sobre o uso de IA generativa, PLN, reconhecimento de fala e visão computacional na avaliação e no feedback de habilidades clínicas em OSCE, identificando o grau de concordância com avaliadores humanos, as aplicações mais consolidadas e as principais limitações metodológicas e éticas relatadas.

3. MÉTODO

Trata-se de revisão integrativa da literatura. A busca foi realizada nas bases PubMed/MEDLINE e Europe PMC, complementada por periódicos da família JMIR, combinando os descritores “artificial intelligence”, “ChatGPT”, “large language model”, “natural language processing”, “computer vision”, “speech recognition”, associados a “objective structured clinical examination”, “clinical skills assessment” e “medical education”.

Foram incluídos estudos primários e revisões sistemáticas, publicados entre 2016 e 2025, sobre IA na avaliação de desempenho clínico ou geração de feedback em OSCE ou simulação equivalente. Excluídos: IA aplicada só ao diagnóstico, editoriais e resumos sem texto completo.

Todas as referências foram verificadas quanto a autoria, periódico, ano e DOI/PMID. Não foi aplicado PRISMA formal nem escore de qualidade metodológica, limitação detalhada na seção de limitações. Como se trata de revisão de literatura, sem coleta direta de dados, o presente trabalho está dispensado de Comitê de Ética em Pesquisa, conforme item 4.3 do edital do 6º AFYA Global Meeting.

4. RESULTADOS

Inteligência artificial generativa como paciente simulado e avaliador

O estudo de maior repercussão submeteu o ChatGPT a uma simulação do exame de admissão ao Royal College of Obstetricians and Gynaecologists: o sistema obteve pontuação média de 77,2%, superior à média histórica de 73,7% de candidatos humanos⁴. O uso do ChatGPT no apoio à construção de estações de OSCE sugere ganhos de eficiência para docentes².

Um estudo comparou avaliação humana em tempo real, por vídeo, e por dois modelos de IA em quatro habilidades clínicas, observando notas sistematicamente mais altas da IA, levantando preocupação sobre calibração¹. Outro testou a validade de IA generativa na avaliação de estudantes⁵, e um terceiro avaliou o ChatGPT como paciente simulado, com viabilidade técnica mas variabilidade entre execuções⁶. Estudo qualitativo identificou boa verossimilhança conversacional do ChatGPT, mas limitações em manter consistência clínica em diálogos longos⁷.

Processamento de linguagem natural para feedback de anamnese e prontuário

Um grupo desenvolveu modelo preditivo baseado em PLN a partir de feedback narrativo de docentes⁸, incorporado a painel visual de desempenho do estudante⁹. Outro estudo relatou boa aceitabilidade de feedback automatizado, com preferência por manter revisão humana complementar¹⁰. Estudo japonês encontrou benefício modesto de transcrições por reconhecimento de fala como autorrevisão¹¹.

Reconhecimento de fala e comunicação não verbal

Pesquisadores testaram plataforma de telessaúde com feedback automatizado sobre comportamento não verbal em consultas simuladas, relatando aumento da consciência dos estudantes¹². Outro estudo avaliou pacientes simulados virtuais com pontuação por reconhecimento de fala, relatando acurácia satisfatória como apoio à avaliação¹³.

Visão computacional na avaliação de habilidades procedurais

Um dos primeiros sistemas de avaliação automatizada de habilidades cirúrgicas a partir de vídeo foi proposto para escolas médicas¹⁴. Estudo posterior avaliou desempenho cirúrgico a partir de vídeos de console robótico com boa acurácia¹⁵. Duas revisões sistemáticas consolidaram essa evidência^{16,17}: modelos de ML já alcançam desempenho comparável a avaliadores experientes, mas com heterogeneidade metodológica.

Síntese e implicações

Em conjunto, sistemas de IA apresentam desempenho comparável, e por vezes superior, ao humano em domínios cognitivos e técnico-procedurais, mas a evidência é mais limitada em competências relacionais e comunicação não verbal. A tendência de a IA atribuir notas mais altas¹ é achado relevante para calibração em avaliações somativas.

Do ponto de vista ético, nenhum estudo avaliou em profundidade viés algorítmico ou risco de estudantes aprenderem a “otimizar” respostas para a IA em vez de desenvolver a habilidade clínica subjacente.

Limitações

A busca não seguiu protocolo PRISMA formal nem avaliação pareada de elegibilidade. A maioria dos estudos tem desenho observacional de pequeno porte, unicêntrico. A heterogeneidade de desfechos entre os quatro subcampos impediu síntese quantitativa.

5. CONCLUSÃO

A inteligência artificial vem sendo aplicada de forma crescente à avaliação e ao feedback de habilidades clínicas em OSCE, com evidência mais consistente em domínios cognitivos e técnico-procedurais do que em competências relacionais.

Os achados indicam potencial real como apoio complementar, sobretudo para feedback formativo, mas não sustentam adoção isolada em avaliações somativas de alto risco sem supervisão humana e validação prospectiva multicêntrica.

REFERÊNCIAS

1. Tekin M, Yurdal MO, Toraman Ç, Korkmaz G, Uysal İ. Is AI the future of evaluation in medical education? AI vs. human evaluation in objective structured clinical examination. *BMC Med Educ.* 2025;25:641.
2. Misra SM, Suresh S. Artificial Intelligence and Objective Structured Clinical Examinations: Using ChatGPT to Revolutionize Clinical Skills Assessment in Medical Education. *J Med Educ Curric Dev.* 2024;11:23821205241263475.
3. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A Language Model-Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study. *JMIR Med Educ.* 2024;10:e59213.
4. Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am J Obstet Gynecol.* 2023;229(2):172.e1-172.e12.
5. Yokose M, Hirosawa T, Sakamoto T, Kawamura R, Suzuki Y, Harada Y, et al. The Validity of Generative Artificial Intelligence in Evaluating Medical Students in Objective Structured Clinical Examination: Experimental Study. *JMIR Form Res.* 2025;9:e79465.
6. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, et al. Application of Large Language Models in Medical Training Evaluation, Using ChatGPT as a Standardized Patient: Multimetric Assessment. *J Med Internet Res.* 2025;27:e59435.
7. Cross J, Kayalackakom T, Robinson RE, Vaughans A, Sebastian R, Hood R, et al. Assessing ChatGPT's Capability as a New Age Standardized Patient: Qualitative Study. *JMIR Med Educ.* 2025;11:e63353.
8. Maimone C, Dolan BM, Green MM, Sanguino SM, Garcia PM, O'Brien CL. Utilizing Natural Language Processing of Narrative Feedback to Develop a Predictive Model of Pre-Clerkship Performance: Lessons Learned. *Perspect Med Educ.* 2023;12(1):141-148.
9. Maimone C, Dolan BM, Green MM, Sanguino SM, O'Brien CL. Using Natural Language Processing to Visualize Narrative Feedback in a Medical Student Performance Dashboard. *Acad Med.* 2024;99(10):1094-1098.
10. Bansal SK, Yadav M, Zhou J, Ebert-Allen RA, Klute RM, Bond WF, et al. Medical Students' Perception of Automated Note Feedback After Simulated Encounters. *Clin Teach.* 2025;22(6):e70273.

11. Shikino K, Tsukamoto T, Noda K, Ohira Y, Yokokawa D, Hirose Y, et al. Do clinical interview transcripts generated by speech recognition software improve clinical reasoning performance in mock patient encounters? A prospective observational study. *BMC Med Educ.* 2023;23(1):272.
12. Liu C, Lim RL, McCabe KL, Taylor S, Calvo RA. A Web-Based Telehealth Training Platform Incorporating Automated Nonverbal Behavior Feedback for Teaching Communication Skills to Medical Students: A Randomized Crossover Study. *J Med Internet Res.* 2016;18(9):e246.
13. Zhang X, Zeng D, Wang X, Fu Y, Han Y, He M, et al. Analysis of virtual standardized patients for assessing clinical fundamental skills of medical students: a prospective study. *BMC Med Educ.* 2024;24(1):981.
14. Zia A, Sharma Y, Bettadapura V, Sarin EL, Ploetz T, Clements MA, et al. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *Int J Comput Assist Radiol Surg.* 2016;11(9):1623-1636.
15. Baghdadi A, Hussein AA, Ahmed Y, Cavuoto LA, Guru KA. A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos. *Int J Comput Assist Radiol Surg.* 2019;14(4):697-707.
16. Lam K, Chen J, Wang Z, Iqbal FM, Darzi A, Lo B, et al. Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digit Med.* 2022;5(1):24.
17. Pedrett R, Mascagni P, Beldi G, Padoy N, Lavanchy JL. Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review. *Surg Endosc.* 2023;37(10):7412-7424.