REVISTA ENFERMAGEM ATUAL IN DERME

USO DA INTELIGÊNCIA ARTIFICIAL NA AVALIAÇÃO DE SERVIÇOS EM SAÚDE: PROTOCOLO DE REVISÃO DE ESCOPO

USE OF ARTIFICIAL INTELLIGENCE IN THE EVALUATION OF HEALTH SERVICES: SCOPING REVIEW PROTOCOL

USO DE LA INTELIGENCIA ARTIFICIAL EN LA EVALUACIÓN DE LOS SERVICIOS DE SALUD: PROTOCOLO DE REVISIÓN DE ALCANCE

[1]Olivia Maria Villefort França
[2]Ana Carla Dantas Cavalcanti
[3]Flávio Luiz Seixas
[4]Lucas Souza de Oliveira

[1]Universidade Federal Fluminense, Rio de Janeiro, Brasil. ORCID: https://orcid.org/0009-0000-6112-6764
[2]Universidade Federal Fluminense, Rio de Janeiro, Brasil. ORCID: https://orcid.org/0000-0003-3531-4694
[3]Universidade Federal Fluminense, Rio de Janeiro, Brasil. ORCID: https://orcid.org/0000-0002-7160-0818
[4]Universidade Federal Fluminense, Rio de Janeiro, Brasil. ORCID: https://orcid.org/0009-0008-9497-7416

Autor correspondente
Olivia Maria Villefort França
Av. Deputado Cristovam Chiaradia, Nº 837, Apto 203, Brasil. Belo Horizonte, Minas Gerais. CEP: 30.575-815
Telefone: +5531 99614-6159 - E-mail: ofranca@id.uff.br

Submissão: 04-08-2025
Aprovado: 17-03-2026

RESUMO
Introdução: O uso da Inteligência Artificial na saúde tem avançado de forma expressiva, especialmente com os Modelos de Linguagem de Larga Escala, já aplicados em tarefas como diagnóstico, triagem e comunicação clínica. Apesar dos avanços, observa-se uma lacuna na literatura quanto à aplicação desses modelos como ferramentas de avaliação e validação de tecnologias em saúde. **Objetivo:** Mapear as evidências disponíveis sobre o uso de Modelos de Linguagem de Larga Escala como avaliadores em estudos voltados à avaliação de tecnologias em saúde. **Métodos:** Trata-se de um protocolo de revisão de escopo conduzido conforme a metodologia do *Joanna Briggs Institute* e as diretrizes do checklist PRISMA-ScR. A pergunta de pesquisa foi elaborada com base na estratégia PCC: População (serviços de saúde), Conceito (inteligência artificial) e Contexto (tecnologias em saúde). As buscas serão realizadas nas bases MEDLINE via PubMed, LILACS via BVS, Web of Science, Scopus e Embase via Portal CAPES, com estratégias ajustadas à lógica e aos descritores específicos de cada base. A triagem dos estudos será realizada por dois revisores independentes, utilizando o software Rayyan. Os dados extraídos serão organizados em planilhas do Excel e analisados com o apoio do software IRaMuTeQ.
Palavras-chave: Inteligência Artificial; Tecnologias em Saúde; Serviços de Saúde.

ABSTRACT
Introduction: The use of Artificial Intelligence (AI) in healthcare has grown significantly, particularly with the development of Large Language Models (LLMs), already applied in tasks such as diagnosis, triage, and clinical communication. Despite these advancements, there is a gap in the literature regarding the use of such models as tools for the evaluation and validation of health technologies. **Objective:** To map the available evidence on the use of Large Language Models as evaluators in studies aimed at assessing health technologies. **Methods:** This is a scoping review protocol conducted in accordance with the Joanna Briggs Institute methodology and the PRISMA-ScR checklist. The research question was formulated using the PCC framework: Population (health services), Concept (artificial intelligence), and Context (health technologies). Searches will be conducted in the MEDLINE database via PubMed, LILACS via BVS, Web of Science, Scopus, and Embase via the CAPES Portal, with strategies tailored to the indexing logic and controlled vocabularies of each database. Study selection will be performed independently by two reviewers using the Rayyan software. Extracted data will be organized in Excel spreadsheets and analyzed with the support of the IRaMuTeQ software.
Keywords: Artificial Intelligence; Health Technologies; Health Services.

RESUMEN
Introducción: El uso de la Inteligencia Artificial (IA) en la salud ha avanzado considerablemente, especialmente con los Modelos de Lenguaje de Gran Escala (LLMs), ya aplicados en tareas como diagnóstico, triaje y comunicación clínica. A pesar de estos avances, se observa una brecha en la literatura respecto al uso de dichos modelos como herramientas de evaluación y validación de tecnologías en salud. **Objetivo:** Mapear las evidencias disponibles sobre el uso de Modelos de Lenguaje de Gran Escala como evaluadores en estudios dirigidos a la evaluación de tecnologías en salud. **Métodos:** Se trata de un protocolo de revisión de alcance realizado según la metodología del Joanna Briggs Institute y las directrices del checklist PRISMA-ScR. La pregunta de investigación fue formulada con base en la estrategia PCC: Población (servicios de salud), Concepto (inteligencia artificial) y Contexto (tecnologías en salud). Las búsquedas se realizarán en las bases MEDLINE vía PubMed, LILACS vía BVS, Web of Science, Scopus y Embase a través del Portal CAPES, con estrategias ajustadas a la lógica de indexación y a los descriptores específicos de cada base. La selección de los estudios será realizada por dos revisores independientes mediante el software Rayyan. Los datos extraídos serán organizados en hojas de cálculo de Excel y analizados con el apoyo del software IRaMuTeQ.
Palabras clave: Inteligencia Artificial; Tecnologías en Salud; Servicios de Salud.

REVISTA ENFERMAGEM ATUAL IN DERME

INTRODUÇÃO

A aplicação da Inteligência Artificial (IA) na área da saúde tem avançado de forma expressiva nos últimos anos, especialmente com o desenvolvimento dos Modelos de Linguagem de Larga Escala (LLMs, do inglês Large Language Models), que vêm se consolidando como ferramentas promissoras em diferentes dimensões do cuidado[1]. Entre esses avanços, destaca-se a IA generativa, uma vertente capaz de criar textos, imagens, sons ou outros formatos de conteúdo a partir de dados existentes, utilizando padrões aprendidos durante o treinamento.

No contexto das LLMs, a IA generativa é empregada para produzir respostas coerentes e contextualizadas, simulando interações humanas e permitindo a geração de informações clínicas personalizadas. Esses modelos têm demonstrado elevado potencial para apoiar tarefas complexas, como a formulação de hipóteses diagnósticas, a triagem de sintomas, a tomada de decisão clínica e a comunicação entre profissionais de saúde e pacientes, atividades que tradicionalmente exigem sensibilidade contextual e julgamento clínico qualificado[2,3].

Embora os avanços tecnológicos tenham ampliado o interesse e a incorporação dos LLMs na prática clínica, ainda são escassos os estudos que examinam de forma crítica sua aplicação em processos estruturados de avaliação e validação de tecnologias, intervenções ou modelos de cuidado[4]. Essa lacuna torna-se ainda mais relevante diante da crescente demanda por métodos escaláveis, objetivos e cientificamente robustos, capazes de sustentar a validação e a integração segura de soluções digitais e abordagens clínicas inovadoras[1,5].

Entre as propostas emergentes, destaca-se o conceito de *LLM-as-a-Judge*, que discute o uso de modelos de linguagem como avaliadores automatizados de saídas textuais em tarefas que envolvem julgamento complexo. Essa abordagem propõe que os LLMs possam complementar ou mesmo substituir avaliadores humanos em contextos como a revisão por pares de conteúdos acadêmicos, a classificação de respostas clínicas e a análise de textos gerados por IA, oferecendo vantagens em termos de escalabilidade, consistência e custo-efetividade[6].

Além dessas possibilidades de aplicação, o diferencial do paradigma *LLM-as-a-Judge* reside em sua capacidade de aliar amplitude operacional à sensibilidade semântica, resultado do treinamento intensivo em linguagem natural. Essa abordagem busca superar deficiências tanto das métricas automatizadas convencionais, geralmente restritas a análises superficiais, quanto das avaliações humanas, frequentemente marcadas por variabilidade, alto custo e baixa reprodutibilidade[6].

No cenário internacional, um dos marcos mais relevantes foi o lançamento do *HealthBench*, pela *OpenAI*, em maio de 2025. Trata-se de um benchmark público composto por 5.000 diálogos clínicos simulados, validados por 262 médicos de 60 países com 48.562 critérios clínicos rubricados, destinado a avaliar modelos de linguagem em cenários realistas. O conjunto aborda múltiplas especialidades e contextos assistenciais, permitindo mensurações criteriosas de atributos como acurácia diagnóstica, segurança, empatia na comunicação e capacidade

REVISTA ENFERMAGEM ATUAL IN DERME

de raciocínio clínico. A disponibilização aberta desses dados e critérios marca um avanço significativo em direção a uma avaliação de IA na saúde mais ética, transparente e tecnicamente robusta[2].

Iniciativas como essa sinalizam uma transformação relevante na maneira como se estruturam os processos de avaliação e incorporação da inteligência artificial no campo da saúde. A compreensão crítica de suas capacidades, e especialmente de suas limitações, exige mais do que indicadores técnicos; demanda também referenciais clínicos e éticos que garantam um uso seguro, equitativo e verdadeiramente centrado no paciente[7].

Uma busca exploratória nas bases de dados da MEDLINE/PubMed, realizada em julho de 2025, não identificou revisões sistemáticas ou de escopo recentes ou em andamento que abordem de forma mais aprofundada o uso de LLMs, bem como outros modelos de IA Generativa, como ferramentas avaliadoras nos serviços saúde. Diante dessa lacuna, torna-se urgente mapear e sintetizar as evidências disponíveis sobre essa aplicação, a fim de contribuir para o amadurecimento conceitual e metodológico da área. A sistematização dessas evidências pode oferecer subsídios relevantes para o desenvolvimento de diretrizes seguras, efetivas e alinhadas às boas práticas científicas, favorecendo uma implementação ética e responsável dessas tecnologias na saúde.

MÉTODOS

A presente revisão de escopo será conduzida conforme as diretrizes metodológicas do *Joanna Briggs Institute* (JBI), respeitando as etapas sistemáticas recomendadas[8]. O processo envolve: definição e alinhamento do objetivo e da pergunta de pesquisa; delimitação dos critérios de inclusão em consonância com esses elementos; detalhamento da estratégia de busca, seleção e extração dos dados, bem como da forma de apresentação das evidências; realização da busca nas fontes selecionadas; triagem e seleção dos estudos; extração e análise dos dados; organização e apresentação dos achados; além da síntese final, considerando o objetivo proposto, as principais conclusões e suas implicações para a prática e para futuras pesquisas[9].

O protocolo desta revisão foi devidamente registrado na plataforma *Open Science Framework* (OSF) e está disponível para acesso público por meio do DOI: https://doi.org/10.17605/OSF.IO/D4UVF.

Pergunta da revisão

A formulação da pergunta de pesquisa seguiu a estratégia metodológica baseada no acrônimo PCC, que contempla os elementos: População, Conceito e Contexto. A questão norteadora definida foi: Como a Inteligência Artificial tem sido validada em estudos realizados no contexto dos serviços de saúde, com destaque para LLMs e outros modelos de IA Generativa? Nesse enquadramento, a População (P) corresponde à Inteligência Artificial; o

REVISTA ENFERMAGEM ATUAL IN DERME

Conceito (C) refere-se aos estudos de validação; e o Contexto (C) diz respeito aos serviços de saúde.

Critério de elegibilidade

Dada a natureza exploratória da revisão de escopo, serão adotados critérios de inclusão amplos e não serão considerados recortes temporais ou idiomáticos. Estudos duplicados serão identificados e contabilizados apenas uma vez.

Serão considerados elegíveis diferentes delineamentos metodológicos, incluindo estudos quantitativos, qualitativos, mistos, revisões, artigos de opinião e relatos técnicos.

Estratégia de busca

A formulação da estratégia de busca seguiu um processo em três etapas interdependentes. Inicialmente, foram identificados termos relevantes nos títulos, resumos e descritores de artigos previamente selecionados. Em seguida, esses termos foram organizados e testados com o objetivo de ajustar a sensibilidade e a especificidade da estratégia. Na terceira etapa, a expressão final foi aplicada de forma sistemática, respeitando os critérios de elegibilidade estabelecidos para a revisão.

A busca preliminar foi realizada na base MEDLINE via PubMed em 25 de julho de 2025 a partir da combinação de descritores dos vocabulários controlados (*Medical Subject Headings* – MeSH) com palavras-chave livres, conforme descrito na figura 1.

Figura 1 - Estratégia de busca para recuperação das publicações nas bases de dados. Rio de Janeiro, RJ, Brasil, 2023

Busca	Estratégia	Resultados
#1	("Artificial Intelligence"[MeSH Terms] OR "Artificial Intelligence"[All Fields])	310.385
#2	("Benchmarking"[MeSH Terms] OR "Validation Studies as Topic"[MeSH Terms] OR "Performance Evaluation"[All Fields] OR "Model Evaluation"[All Fields])	34.385
#3	("Health Services"[MeSH Terms] OR "Health Care"[All Fields] OR "Medical Informatics Applications"[MeSH Terms])	3.791.975
#4	#1 AND #2	2,818
#5	#4 AND #3	762

As buscas serão conduzidas nas seguintes bases de dados científicas: MEDLINE, por meio da plataforma PubMed; LILACS, acessada via Biblioteca Virtual em Saúde (BVS); Web of Science (WoS); Scopus; e Embase, esta última acessada pelo Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). As estratégias de busca serão devidamente adaptadas à lógica de indexação, aos vocabulários controlados e aos operadores específicos de cada base, a fim de assegurar sensibilidade, abrangência e precisão na identificação das evidências pertinentes à questão da revisão.

REVISTA ENFERMAGEM ATUAL IN DERME

Seleção da fonte de evidências

Os resultados obtidos nas buscas serão transferidos para o software *EndNote Web* para a identificação e remoção de duplicatas. Em seguida, os registros únicos serão importados e triados na plataforma Rayyan. O processo de seleção será conduzido por dois revisores independentes, previamente capacitados, que realizarão a triagem dos títulos e resumos de forma individual e cega, utilizando dispositivos distintos para garantir a imparcialidade. Os estudos que atenderem aos critérios de elegibilidade serão selecionados para leitura na íntegra. Aqueles que não responderem à pergunta da revisão serão excluídos nesta etapa.

Eventuais divergências entre os revisores durante a seleção serão resolvidas por um terceiro avaliador, com experiência no tema, que emitirá o parecer final. Os registros excluídos e os motivos de sua exclusão serão documentados no relatório da revisão.

Para assegurar a transparência e a rastreabilidade do processo, os resultados da triagem serão apresentados por meio de um fluxograma conforme as diretrizes do PRISMA-ScR (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews*[10].

Extração de dados

Os estudos que atenderem aos critérios de elegibilidade serão acessados em texto completo e analisados integralmente por um revisor independente. A extração das informações será realizada por meio de uma planilha elaborada no Microsoft Excel, contemplando autores, ano, objetivos, método, principais resultados e construída especificamente para esta revisão, com base nas orientações do *JBI Manual for Evidence Synthesis*.

Análise e apresentação das evidências

As informações extraídas serão organizadas em quadros sintéticos e, também representadas visualmente por meio de uma nuvem de palavras, gerada com o auxílio do software *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires* (IRAMUTEQ). Os achados serão ainda discutidos em um resumo descritivo, articulando os resultados com os objetivos e a pergunta da revisão.

CONSIDERAÇÕES FINAIS

Espera-se mapear as abordagens metodológicas e identificar como as LLMs e outros tipos de IA Generativa tem sido validadas em estudos realizados no contexto dos serviços de saúde. Este protocolo visa proporcionar rigor metodológico e transparência ao processo de revisão de escopo sobre o tema, oferecendo subsídios relevantes para pesquisadores, profissionais de saúde e formuladores de políticas públicas.

REVISTA ENFERMAGEM ATUAL IN DERME

REFERÊNCIAS

1. Maity S, Saikia MJ. Large language models in healthcare and medical applications: a review. Bioengineering (Basel). 2025 Jun 10;12(6):631. doi:10.3390/bioengineering12060631. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12189880

2. Arora R, Dai A, Zhang Y, et al. DoctorGPT: a clinical large language model for reasoning and generation [preprint]. arXiv. 2025. Disponível em: https://arxiv.org/abs/2403.01859

3. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res. 2025;27:e59069. doi:10.2196/59069. Disponível em: https://www.jmir.org/2025/1/e59069

4. Morone G, De Angelis L, Martino Cinnera A, Carbonetti R, Bisirri A, Ciancarelli I, et al. Artificial intelligence in clinical medicine: a state of the art overview of systematic reviews with methodological recommendations for improved reporting. Front Digit Health. 2025;7:1550731. doi:10.3389/fdgth.2025.1550731. Disponível em: https://www.frontiersin.org/articles/10.3389/fdgth.2025.1550731/full

5. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA. Digital health solutions and public health: a call to action. J Med Internet Res. 2023;25:e46992. doi:10.2196/46992. Disponível em: https://www.jmir.org/2023/1/e46992

6. Gu J, Jiang X, Shi Z, Tan H, Zhai X, et al. A survey on LLM-as-a-judge [preprint]. arXiv. 2025 [citado 2025 maio 27]. Disponível em: https://arxiv.org/abs/2411.15594

7. Singh MP, Keche YN. Ethical integration of artificial intelligence in healthcare: narrative review of global challenges and strategic solutions. Cureus. 2025 May 25;17(5):e84804. doi:10.7759/cureus.84804. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12195640

8. Joanna Briggs Institute. Template for scoping reviews protocols [Internet]. Adelaide: JBI; 2020 [citado 2022 jun 8]. Disponível em: https://jbi.global/scoping-review-network/resources

9. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping reviews (2020). In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, eds. JBI manual for evidence synthesis [Internet]. Adelaide: JBI; 2024 [citado 2025 Maio 27]. Disponível em: https://synthesismanual.jbi.global

10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467–73. doi:10.7326/M18-0850

Fomento e Agradecimento

A pesquisa não recebeu financiamento.

Declaração de conflito de interesses

Nada a declarar.

Declaração de disponibilidade de dados

Não foram gerados bancos de dados neste estudo. As informações apresentadas estão descritas no corpo do artigo.

Critérios de autoria

Olivia Maria Villefort França: Concepção e planejamento do estudo; Redação, revisão crítica e aprovação final

Ana Carla Dantas Cavalcanti: Revisão crítica e aprovação final

Flávio Luiz Seixas: Revisão crítica e aprovação final

Lucas Souza de Oliveira: Revisão crítica e aprovação final

REVISTA ENFERMAGEM ATUAL IN DERME

Editor Científico: Ítalo Arão Pereira Ribeiro.
Orcid: https://orcid.org/0000-0003-0778-1447