SCOPE REVIEW PROTOCOL

REVISTA ENFERMAGEM ATUAL IN DERME

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

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

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

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

Submission: 04-08-2025
Approval: 17-03-2026

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

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.

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.

INTRODUCTION

The application of Artificial Intelligence (AI) in healthcare has advanced significantly in recent years, especially with the development of Large Language Models (LLMs), which are becoming established as promising tools in different dimensions of care [1]. Among these advances, generative AI stands out, a branch capable of creating texts, images, sounds, or other content formats from existing data, using patterns learned during training.

In the context of LLMs, generative AI is used to produce coherent and contextualized responses, simulating human interactions and allowing the generation of personalized clinical information. These models have demonstrated high potential to support complex tasks, such as the formulation of diagnostic hypotheses, symptom screening, clinical decision-making, and communication between healthcare professionals and patients, activities that traditionally require contextual sensitivity and qualified clinical judgment [2,3].

Although technological advances have broadened the interest in and incorporation of LLMs in clinical practice, studies that critically examine their application in structured processes for evaluating and validating technologies, interventions, or care models are still scarce [4]. This gap becomes even more relevant given the growing demand for scalable, objective, and scientifically robust methods capable of supporting the validation and safe integration of digital solutions and innovative clinical approaches [1,5].

Among emerging proposals, the concept of LLM-as-a-Judge stands out, which discusses the use of language models as automated evaluators of textual outputs in tasks involving complex judgment. This approach proposes that LLMs can complement or even replace human evaluators in contexts such as peer review of academic content, classification of clinical responses, and analysis of AI-generated texts, offering advantages in terms of scalability, consistency, and cost-effectiveness [6].

Beyond these application possibilities, the differentiating factor of the LLMasaJudge paradigm lies in its ability to combine operational breadth with semantic sensitivity, a result of intensive training in natural language. This approach seeks to overcome deficiencies in both conventional automated metrics, generally restricted to superficial analyses, and human assessments, frequently marked by variability, high cost, and low reproducibility [6].

In the international arena, one of the most relevant milestones was the launch of HealthBench by OpenAI in May 2025. This is a public benchmark composed of 5,000 simulated clinical dialogues, validated by 262 physicians from 60 countries with 48,562 rubricated clinical criteria, designed to evaluate language models in realistic scenarios. The set addresses multiple specialties and care contexts, allowing for rigorous measurements of attributes such as diagnostic accuracy, safety, empathy in communication, and clinical reasoning ability. The open availability of this data and criteria marks a significant step towards a more ethical, transparent, and technically robust evaluation of AI in healthcare [2].

Initiatives like this signal a relevant transformation in how the processes of evaluating and incorporating artificial intelligence in the healthcare field are structured. A critical understanding of its capabilities, and especially its limitations, requires more than technical indicators; it also demands clinical and ethical frameworks that guarantee safe, equitable, and truly patient-centered use [7].

An exploratory search of the MEDLINE/PubMed databases, conducted in July 2025, did not identify recent or ongoing systematic or scoping reviews that address in more depth the use of LLMs, as well as other Generative AI models, as evaluation tools in healthcare services. Given this gap, it becomes urgent to map and synthesize the available evidence on this application, in order to contribute to the conceptual and methodological maturation of the area. Systematizing this evidence can provide relevant input for the development of safe, effective guidelines aligned with good scientific practices, promoting the ethical and responsible implementation of these technologies in healthcare.

METHODS

This scoping review will be conducted according to the methodological guidelines of the Joanna Briggs Institute (JBI), respecting the recommended systematic steps[8]. The process involves: defining and aligning the objective and research question; delimiting the inclusion criteria in accordance with these elements; detailing the search strategy, data selection and extraction, as well as the form of presentation of the evidence; conducting the search in the selected sources; screening and selection of studies; data extraction and analysis; organization and presentation of findings; in addition to the final synthesis, considering the proposed objective, the main conclusions and their implications for practice and future research[9].

The protocol of this review was duly registered on the Open Science Framework (OSF) platform and is available for public access through the DOI: https://doi.org/10.17605/OSF.IO/D4UVF·

Review Question

The formulation of the research question followed the methodological strategy based on the PCC acronym, which includes the elements: Population, Concept and Context. The guiding question defined was: How has Artificial Intelligence been validated in studies conducted in the context of healthcare services, with emphasis on LLMs and other Generative AI models? In this framework, the Population (P) corresponds to Artificial Intelligence; the Concept (C) refers to validation studies; and the Context (C) refers to healthcare services.

Eligibility Criteria

Given the exploratory nature of the scoping review, broad inclusion criteria will be adopted, and temporal or idiomatic delimitations will not be considered. Duplicate studies will be identified and counted only once.

Different methodological designs will be considered eligible, including quantitative, qualitative, mixed studies, reviews, opinion articles, and technical reports.

Search Strategy

The formulation of the search strategy followed a three-step interdependent process. Initially, relevant terms were identified in the titles, abstracts, and descriptors of previously selected articles. Next, these terms were organized and tested with the aim of adjusting the sensitivity and specificity of the strategy. In the third stage, the final expression was systematically applied, respecting the eligibility criteria established for the review.

The preliminary search was conducted in the MEDLINE database via PubMed on July 25, 2025, using a combination of descriptors from controlled vocabularies (Medical Subject Headings – MeSH) with free keywords, as described in Figure 1.

Figure 1 - Search strategy for retrieving publications in databases. Rio de Janeiro, RJ, Brazil, 2023

Search	Strategy	Results
#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

The searches will be conducted in the following scientific databases: MEDLINE, through the PubMed platform; LILACS, accessed via the Virtual Health Library (BVS); Web of Science (WoS); Scopus; and Embase, the latter accessed through the CAPES (Coordination for the Improvement of Higher Education Personnel) Periodicals Portal. The search strategies will be duly adapted to the indexing logic, controlled vocabularies, and specific operators of each database, in order to ensure sensitivity, comprehensiveness, and precision in identifying evidence relevant to the review question.

Selection of the source of evidence

The results obtained in the searches will be transferred to the EndNote Web software for the identification and removal of duplicates. Then, the unique records will be imported and sorted on the Rayyan platform. The selection process will be conducted by two independent, previously trained reviewers, who will individually and blindly screen the titles and abstracts, using different devices to ensure impartiality. Studies that meet the eligibility criteria will be selected for full-text reading. Those that do not answer the review question will be excluded at this stage.

Any disagreements between reviewers during the selection process will be resolved by a third reviewer with expertise in the subject, who will issue the final opinion. The excluded records and the reasons for their exclusion will be documented in the review report.

To ensure transparency and traceability of the process, the screening results will be presented using a flowchart in accordance with the PRISMA-ScR guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews[10].

Data Extraction

Studies that meet the eligibility criteria will be accessed in full text and fully analyzed by an independent reviewer. Information extraction will be performed using a spreadsheet created in Microsoft Excel, including authors, year, objectives, method, main results, and specifically constructed for this review, based on the guidelines of the JBI Manual for Evidence Synthesis.

Analysis and Presentation of Evidence

The extracted information will be organized into summary tables and also visually represented by a word cloud, generated using the Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (IRAMUTEQ) software. The findings will also be discussed in a descriptive summary, linking the results to the objectives and the research question. of the review.

FINAL CONSIDERATIONS

The aim is to map methodological approaches and identify how LLMs and other types of Generative AI have been validated in studies conducted in the context of health services. This protocol aims to provide methodological rigor and transparency to the scoping review process on the subject, offering relevant support for researchers, health professionals, and public policy makers.

REFERENCES

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

Funding and Acknowledgments

This research received no funding.

Conflict of Interest Statement
Nothing to declare.

Data Availability Statement No databases were generated in this study. The information presented is described in the body of the article.

Authorship Criteria

Olivia Maria Villefort França: Study conception and planning; Writing, critical review, and final approval

Ana Carla Dantas Cavalcanti: Critical review and final approval

Flávio Luiz Seixas: Critical review and final approval

Lucas Souza de Oliveira: Critical review and final approval

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