Journal of Women Health Care and Gynecology
2025 Volume 5, Issue 2
DOI: 10.59657/2993-0871.brs.25.079

BioRes Scientia
Value your Research

Review Article

Open Access

Artificial Intelligence and Large Language Models in Medicine: A Critical Reflection on Current State, Challenges, and Future Implications

Canio Martinelli[1,2]*, Davide Arrigo[3] , Vincenzo De Meo[4], Antonio Giordano[1,5], Alfredo Ercoli[2]

[1]Sbarro Institute for Cancer Research and Molecular Medicine and Center of Biotechnology, College of Science and Technology, Temple University, 1900 N 12th St, Philadelphia, PA 19122, United States. [2]Department of Human Pathology of Adult and Childhood "Gaetano Barresi", Unit of Obstetrics and Gynecology, University of Messina, Via Consolare Valeria 1, 98124, Messina, Italy. [3]Dipartimento per la Salute della Donna e del Bambino e della Salute Pubblica, Fondazione Policlinico Universitario A Gemelli, IRCCS, UOC Ginecologia Oncologica, Rome 00168, Italy; Università Cattolica del Sacro Cuore, Istituto di Ginecologia e Ostetricia, Rome, Italy. [4]Department of Clinical and Experimental Medicine, Unit of Anesthesia and Intensive Care, Maternal-Infant Service, University of Messina, Piazza Pugliatti, 1 - 98122 Messina, Italy. [5]Department of Medical Biotechnology, University of Siena, via Aldo Moro 2, 53100, Siena, Italy.

**Corresponding author: Canio Martinelli.*

Abstract

Artificial Intelligence (AI) and Large Language Models (LLMs) are rapidly transforming healthcare delivery and medical practice. This reflection paper critically examines the current landscape, applications, and implications of these technologies in medicine. Recent advances in LLMs, exemplified by models like GPT-4, Claude 3, and MedPaLM-2, have demonstrated remarkable capabilities in medical knowledge assessment, achieving performance levels that surpass average medical students in standardized examinations. While these technologies show promise in various domains, including medical imaging analysis, clinical decision support, and medical education, significant challenges persist regarding their implementation and validation. The paper explores critical concerns about output reliability, the prevalence of "hallucinations," and the need for rigorous validation processes in clinical settings. Particular attention is given to emerging applications in surgical specialties, where AI integration faces unique challenges due to procedural heterogeneity and the need for real-time adaptation. The discussion extends to broader implications for healthcare delivery, including the potential for reducing administrative burden and the importance of maintaining human oversight in clinical decision-making. Through critical analysis, this paper reflects on the balance between technological advancement and clinical responsibility, emphasizing the need for thoughtful integration of AI tools while preserving the essential role of human judgment in medical practice.

Keywords: artificial intelligence; delivery; LLMs; medical education

Introduction

Artificial Intelligence (AI) has emerged as a transformative force in healthcare, fundamentally reshaping approaches to medical practice, education, and research. As technology increasingly permeates clinical settings, understanding AI's foundational concepts and their implications for healthcare delivery becomes crucial. AI can be broadly conceptualized as "technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy [1]." Within the AI ecosystem, distinct but interrelated concepts have evolved. Machine Learning (ML) encompasses the development of algorithmic models trained to make predictions or decisions based on data patterns, while Deep Learning (DL) employs multilayered neural networks to approximate complex human cognitive processes. These technologies serve as the foundation for specialized applications in Computer Vision (CV) and Natural Language Processing (NLP), enabling machines to interpret visual inputs and interact with human language, respectively [2-4]. The emergence of Generative AI (GenAI) represents a significant advancement, particularly through Large Language Models (LLMs) such as Open AI's GPT-4, Google's Gemini 1.5, and Anthropic's Claude 3. These models exemplify the concept of "foundation models," which are pre-trained on extensive datasets via self-supervised learning (SSL), enabling broad applicability across various tasks without task-specific

retraining. In the medical domain, specialized models like Google DeepMind's MedPaLM-2 and NVIDIA's BioNeMo have been developed to address healthcare-specific applications, from clinical decision support to molecular biology research [2-6]. The integration of AI in healthcare presents both unprecedented opportunities and significant challenges. While these technologies demonstrate remarkable capabilities in medical knowledge assessment [7-9] and clinical applications [10,11-14], concerns persist regarding output reliability, the potential for "hallucinations," and the need for rigorous validation in clinical settings [15,16]. The high linguistic proficiency of modern LLMs can mask inaccuracies, necessitating careful oversight and expert verification of their outputs. This reflection paper aims to critically examine the current landscape of AI and LLMs in medicine, with particular attention to their applications, limitations, and implications for future medical practice. Through analysis of existing literature and emerging evidence, we explore the balance between technological advancement and clinical responsibility, considering both the transformative potential of these technologies and the essential role of human judgment in medical decision-making.

Current Landscape of AI and LLMs in Medicine

The evolution of AI in medicine represents a paradigm shift in healthcare delivery, characterized by increasingly sophisticated models that aim to mirror the complexity of clinical decision-making. Historical perspectives suggest that the foundation for AI in medicine was conceptualized as early as 1959, when Brodman and colleagues proposed that "the making of correct diagnostic interpretations of symptoms can be a process in all aspects logical and so completely defined that it can be carried out by a machine [17]." This early vision has materialized through contemporary developments in AI technologies, particularly in their ability to analyze unstructured data and uncover hidden relationships that may parallel or exceed human intuition [3,10,18].

Technical Architecture and Capabilities

Modern LLMs operate through a sophisticated three-stage process: initial training on extensive multimodal datasets to build "foundational models," subsequent tuning for specific tasks often utilizing reinforcement learning with human feedback (RLHF), and finally, the generation phase where outputs are produced and iteratively refined. A key innovation underlying these systems is the "Transformer" architecture, which enables differential weighting of input data importance, fundamentally enhancing the models' ability to process and generate context-appropriate responses [2-4].

Current Leading Technologies

The landscape of medical AI is dominated by several key platforms:

General-Purpose LLMs

- OpenAI's GPT-4
- Google's Gemini 1.5
- Anthropic's Claude 3

Domain-Specific Medical Models

- Google DeepMind's MedPaLM-2: Specifically fine-tuned for healthcare applications
- NVIDIA's BioNeMo: Specialized for genomics and molecular biology research.
- IBM Watson Assistant for Health: Focused on clinical environment integration [2-6].

Performance Metrics and Validation

Recent evaluations of LLMs' medical knowledge have yielded promising results, particularly in standardized assessment contexts. Studies utilizing datasets from the United States Medical Licensing Examination (USMLE) and related medical examination questions have demonstrated that models like GPT-4 consistently achieve average scores exceeding 80%, surpassing earlier versions and approaching or exceeding average medical student performance. These results were achieved using standardized zero-shot prompting strategies, without requiring advanced techniques such as chain-of-thought prompting or retrieval-augmented generation [7-9].

Integration Challenges

Despite impressive performance metrics, several critical challenges persist:

Accuracy and Reliability

- High prevalence of "hallucinations" in medical contexts
- Need for vigilant expert review and proofreading
- Challenges in verifying the accuracy of referenced information [19].

Implementation Barriers

- Lack of standardized validation methodologies
- Complexities in integrating AI tools into existing clinical workflows

- Requirements for robust data preprocessing and standardization [15,20].

Bias and Representativeness

- Concerns regarding dataset diversity and real-world population representation
- Potential for embedded biases related to demographics and socioeconomic factors
- Need for enhanced understanding of bias introduction during model training [21-24].

Clinical Applications and Impact

How has the integration of AI and LLMs transformed medical practice across different specialties, and what implications does this hold for the future of healthcare? The implementation landscape reveals a complex interplay between technological capability and clinical utility, with varying degrees of validation and acceptance across medical domains. The field of medical imaging, particularly radiology, stands at the forefront of AI adoption, boasting the highest number of FDA-approved AI-enabled devices. The introduction of vision-capable LLMs like GPT-4V has expanded diagnostic possibilities, yet this advancement comes with notable caveats. Recent studies have revealed concerning diagnostic accuracy rates, with hallucination frequencies exceeding 40% in some contexts, highlighting the critical need for careful implementation and validation protocols [15,16]. In oncology, the impact of deep learning algorithms has been particularly noteworthy. These systems have demonstrated capabilities that not only match but occasionally surpass human expertise in several crucial areas. From the precise identification of tumors in imaging studies to the intricate interpretation of genetic data, AI systems have shown remarkable versatility. Perhaps most promising is their application in drug discovery, where they accelerate the development of new therapeutic compounds through sophisticated analysis of both computational and experimental data. The technology's ability to analyze histopathological samples with high precision represents another significant advancement, offering enhanced capability in distinguishing between benign and malignant cells [11-14,20]. The transformation of research methodologies and clinical trials through AI integration raises intriguing possibilities for the future of medical investigation. How might the automation of participant recruitment and the enhancement of data analysis reshape our approach to clinical research? The emergence of "synthetic patient" models for trial simulation suggests a paradigm shift in how we conceptualize and conduct medical research [17,20,25]. This development, while promising, prompts important questions about the validity and generalizability of such approaches. Specialty-specific applications have demonstrated particularly noteworthy developments in Obstetrics and Gynecology. The technology's ability to support complex decision-making in fertility treatment, enhance resident education, and facilitate patient counseling represents a significant advancement in clinical practice [26-29]. These applications extend across various subspecialties, suggesting a broader pattern of utility that may be applicable to other medical fields [30-34]. The surgical domain presents unique challenges and opportunities for AI integration. While adoption has been slower compared to other specialties, innovative applications have emerged in minimally invasive procedures. For instance, the implementation of computer vision analysis in cholecystectomy has enhanced the identification of safe dissection zones, while AI-assisted navigation in colorectal surgery has improved nerve preservation outcomes [35-37]. The impact of these technologies on medical education deserves particular attention. Models like GPT-4 have achieved remarkable performance on standardized medical examinations, consistently scoring above 80% and often surpassing average medical student performance [2-5,7-9]. The ability of these systems to provide detailed explanations and identify errors suggests potential applications in medical training that extend beyond simple knowledge testing.

Critical Analysis of Implementation Challenges

The implementation of AI and LLMs in medical practice presents a complex web of technical, ethical, and practical challenges that warrant careful examination. These challenges fundamentally influence the trajectory of AI integration in healthcare and shape the evolving relationship between technological advancement and clinical practice.

Technical Reliability and Validation

A primary concern in the implementation of AI systems, particularly LLMs, centers on output reliability and validation methodology. The challenge extends beyond simple accuracy metrics to the fundamental question of how to verify responses in contexts where the complexity of medical knowledge may exceed even expert capabilities. This validation

burden ultimately falls to clinicians, yet the depth and breadth of knowledge required for comprehensive verification presents a significant challenge [6,8]. The situation is further complicated by the unclear underpinning of references in LLM outputs, raising substantial concerns about the reliability and integrity of medical content generation [19]. The development of standardized evaluation protocols represents another crucial challenge. While traditional clinical interventions follow well-established validation pathways, AI tools present unique complications in assessment methodology. The medical community currently faces a notable absence of defined standards for description, evaluation, and validation of AI interventions [11,17,20,38]. This gap has led to the emergence of a three-stage evaluation framework encompassing technical performance, usability assessment, and health impact analysis. However, the rapid evolution of AI technologies often outpaces the development of evaluation methodologies, creating a persistent challenge in validation efforts [39-42].

Explain ability and Transparency

The increasing complexity of AI systems has given rise to significant challenges in interpreting their decision-making processes. The emergence of Explainable Artificial Intelligence (XAI) represents an attempt to address this opacity, yet fundamental challenges persist. The intricate architectures of deep learning models, involving millions of parameters and numerous nonlinear transformations, make tracing decision pathways from input to output exceptionally difficult [20,43,44]. This lack of transparency raises critical questions about the integration of these technologies in clinical settings where understanding the basis for decisions is crucial for patient care.

Bias and Generalizability

A fundamental challenge lies in understanding how AI models define and interpret medical "norm values" and how they contextualize their inferences within real-world scenarios [3,6,17,20]. The representation of diverse patient populations in training datasets remains a critical concern, as limitations in dataset diversity can lead to biased outputs and potentially harmful recommendations for certain demographic groups [21-24]. While technical approaches for identifying and mitigating biases in supervised machine-learning systems have advanced, significant challenges persist in defining and standardizing measures of fairness. The inherent biases in LLMs, particularly regarding race, gender, and socioeconomic status, remain poorly understood, especially concerning how these biases are introduced during training and manifest in model outputs [3,23,24].

Integration and Professional Impact

The integration of AI technologies into clinical workflows raises questions about the changing nature of medical practice. How will the increasing capability of AI systems affect the role and perceived value of human expertise? Studies indicate shifting perceptions among medical professionals, particularly evident in specialties like radiology, where career choices are already being influenced by the perceived impact of AI [45]. While younger professionals often view AI as a complementary tool, the broader medical community maintains skepticism about AI's role in clinical decision-making, emphasizing the continued importance of human judgment [46,47].

Future Directions and Implications

The evolving landscape of AI and LLMs in medicine necessitates careful consideration of future research priorities, implementation strategies, and systemic adaptations. This section examines key areas requiring attention and development to optimize the integration of these technologies into medical practice.

Research and Validation Priorities

Future research must address the fundamental challenge of establishing standardized validation methodologies for AI interventions in healthcare. This need is particularly acute given the rapid evolution of AI capabilities and the current lack of established frameworks for evaluation [40,41,42]. The development of pragmatic trials, which can effectively measure AI effectiveness in real-world environments while maintaining methodological rigor, represents a crucial next step in generating actionable evidence for clinical implementation. The advancement of Explainable Artificial Intelligence (XAI) methodologies emerges as another critical research priority. Current limitations in understanding AI decision-making processes necessitate innovative approaches to enhance transparency and interpretability. The development of self-explanatory systems that eliminate the need for post-hoc analyses could significantly improve the integration of AI tools in clinical practice [43,44]. This advancement would not only enhance clinician confidence in AI-generated recommendations but also facilitate more effective oversight and quality assurance.

Clinical Integration and Workflow Optimization

The optimization of AI integration into clinical workflows requires careful consideration of both technical and human factors. The evidence suggests that successful implementation will depend on developing sophisticated interfaces between AI systems and existing healthcare infrastructure, particularly in areas such as electronic health records integration and real-time clinical decision support [4,5,48-51]. The surgical domain presents unique challenges and opportunities, particularly in developing real-time AI assistance that can adapt to the dynamic nature of surgical procedures while maintaining reliability and safety [50,51].

Educational Implications and Professional next

The demonstrated capabilities of LLMs in medical knowledge assessment necessitate a reevaluation of medical education approaches [7-9]. Future educational frameworks must evolve to incorporate AI literacy while maintaining focus on critical thinking and clinical reasoning skills. This evolution requires careful consideration of how to leverage AI tools as educational supplements while ensuring the development of robust clinical judgment. The impact on professional roles and specialization patterns warrants particular attention. Current evidence indicating shifts in specialty choice preferences, particularly in imaging-intensive fields, [45] suggests the need for proactive strategies to address concerns about professional displacement while emphasizing the complementary nature of AI assistance [46,47].

Ethical Considerations and Bias Mitigation

Future developments must prioritize addressing systemic biases and ensuring equitable access to AI-enhanced healthcare. The current limitations in dataset diversity and representativeness require systematic approaches to data collection and model training that better reflect real-world patient populations [21-24]. Additionally, the development of standardized fairness metrics and bias detection methodologies represents a crucial area for future research.

Policy and Regulatory Framework Development

The advancement of AI in medicine necessitates the development of comprehensive regulatory frameworks that can effectively balance innovation with patient safety. Future policy development must address questions of liability, data privacy, and quality assurance while maintaining sufficient flexibility to accommodate rapid technological advancement [38,41].

Concluding Reflections

This critical reflection has examined the multifaceted landscape of AI and LLMs in medicine, revealing both transformative potential and significant implementation challenges that warrant careful consideration. The evidence presented demonstrates that these technologies are reshaping multiple aspects of healthcare delivery, from clinical decision-making to medical education and research methodologies. The performance metrics of current LLM systems, particularly in standardized medical knowledge assessment [2-5,7-9], suggest capabilities that approach or exceed human performance in specific domains. However, this technical prowess must be contextualized within the broader landscape of clinical practice, where the complexity of patient care extends beyond pure knowledge application. The prevalence of "hallucinations" and accuracy concerns in medical applications [15,16] underscores the critical importance of maintaining robust human oversight in clinical decision-making. The integration of AI technologies in surgical specialties [50-53] provides a compelling case study in both the potential and limitations of current AI applications. While promising results have been demonstrated in specific procedures [35-37], the challenges of real-time adaptation and procedural heterogeneity highlight the continuing need for careful validation and implementation strategies. These findings suggest that the optimal path forward likely involves viewing AI tools as augmentative rather than replacement technologies, enhancing rather than superseding human clinical judgment. The evidence regarding bias and fairness in AI systems raises crucial questions about equity in healthcare delivery. The demonstrated limitations in dataset representativeness and the potential for embedded biases suggest that future development must prioritize inclusive data collection and systematic bias mitigation strategies. Paradoxically, while current AI systems may perpetuate certain healthcare disparities, they also hold potential for reducing others through improved access to medical expertise and decision support in underserved areas [3]. The evolving landscape of medical education and professional development requires particular attention. The impact of AI

integration on specialty choice and professional attitudes [46,47]. Suggests the need for proactive strategies to address concerns about professional displacement while emphasizing the complementary nature of AI assistance. Educational frameworks must evolve to incorporate AI literacy while maintaining focus on core clinical competencies. Looking forward, the successful integration of AI in medicine will require careful balance between technological advancement and clinical responsibility. The development of robust validation methodologies [40-42] and enhancement of explainable AI systems [43,44] emerge as critical priorities. Future research must address not only technical capabilities but also the broader implications for healthcare delivery, professional development, and patient outcomes.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no actual or potential conflicts of interest with respect to the research, authorship, and/or publication of this reflection paper. No financial or personal relationships that could inappropriately influence or bias the content of the paper exist.

Data Availability Statement

This reflection paper synthesizes publicly available literature and published research. No novel datasets were generated or analyzed during the development of this manuscript. All cited works are available through standard academic databases and repositories, with specific references provided in the bibliography.

Author Contributions

C. Martinelli: Conceptualization, Methodology, Writing - Original Draft Development of the paper's framework and structure; formulation of the critical analysis approach for examining AI and LLMs in medicine; preparation of the initial manuscript with focus on technological aspects and clinical implications.

D. Arrigo: Investigation, Data Curation, Writing - Review & Editing Comprehensive literature synthesis; systematic organization of evidence regarding AI applications in medicine; critical revision of technical content and validation of clinical applications.

V. De Meo: Writing - Original Draft, Validation, Writing - Review & Editing Initial manuscript development focusing on clinical implementations; verification of medical accuracy; critical revision of content regarding healthcare applications and challenges.

A. Giordano: Supervision, Project administration Overall guidance and oversight of the manuscript development; coordination of author contributions; ensuring academic rigor and coherence throughout the paper.

A. Ercoli: Methodology, Writing - Review & Editing Refinement of analytical approach; critical revision focusing on future implications and research directions; integration of clinical and technological perspectives.

All authors have read and agreed to the published version of the manuscript. Each author's contribution was essential to the development and completion of this reflection paper.

References

1. What Is Artificial Intelligence (AI)? *IBM*.
2. Bahir D, Zur O, Attal L, et al. (2024). Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. *Graefes Arch Clin Exp Ophthalmol*.
3. Nori H, King N, McKinney SM, Carignan D, Horvitz E. (2023). Capabilities of GPT-4 on Medical Challenge Problems.
4. Singhal K, Azizi S, Tu T, et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972):172-180.
5. Clusmann J, Kolbinger FR, Muti HS, et al. (2023). The future landscape of large language models in medicine. *Commun Med*, 3(1):141.
6. Lee P, Bubeck S, Petro J. (2023). Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med*, 388(13):1233-1239.
7. Gilson A, Safranek CW, Huang T, et al. (2023). How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*, 9:e45312.
8. Meyer A, Riese J, Streichert T. (2024). Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the

Written German Medical Licensing Examination: Observational Study. *JMIR Med Educ,* 10:e50965.
9. Liu M, Okuhara T, Chang X, et al. (2024). Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res,* 26:e60807.
10. Varghese C, Harrison EM, O'Grady G, Topol EJ. (2024). Artificial intelligence in surgery. *Nat Med*, 30(5):1257-1268.
11. Kather JN. (2023). Artificial intelligence in oncology: chances and pitfalls. *J Cancer Res Clin Oncol,* 149(10):7995-7996.
12. Unger M, Kather JN. (2024). A systematic analysis of deep learning in genomics and histopathology for precision oncology. *Med Genomics,* 17(1):48.
13. Verlingue L, Boyer C, Olgiati L, Brutti Mairesse C, Morel D, Blay JY. (2024). Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice. *The Lancet Regional Health – Europe,* 46:101064.
14. Gentile F, Malara N. (2024). Artificial intelligence for cancer screening and surveillance. *ESMO Real World Data and Digital Oncology,* 5:100046.
15. Horiuchi D, Tatekawa H, Oura T, et al. (2024). ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol.*
16. Brin D, Sorin V, Barash Y, et al. (2024). Assessing GPT-4 multimodal performance in radiological image analysis. Eur Radiol.
17. Haug CJ, Drazen JM. (2023). Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. Drazen JM, Kohane IS, Leong TY, eds. *N Engl J Med,* 388(13):1201-1208.
18. Truhn D, Reis-Filho JS, Kather JN. (2023). Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat Med,* 29(12):2983-2984.
19. Ghanem D, Zhu AR, Kagabo W, Osgood G, Shafiq B. (2024). ChatGPT-4 Knows Its A B C D E but Cannot Cite Its Source. 9(3):e24.00099.
20. Perez-Lopez R, Ghaffari Laleh N, Mahmood F, Kather JN. (2024). A guide to artificial intelligence for cancer researchers. 24(6):427-441.
21. Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D. (2023). Threats by artificial intelligence to human health and human existence. *BMJ Glob Health,* 8(5):e010435.
22. Vandemeulebroucke T. (2024). The ethics of artificial intelligence systems in healthcare and medicine: from a local to a global perspective, and back. *Pflugers Arch - Eur J Physiol.*
23. Haltaufderheide J, Ranisch R. (2024). The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *NPJ Digit Med,* 7(1):183.
24. Balagopalan A, Baldini I, Celi LA, et al. (2024). Machine learning for healthcare that matters: Reorienting from technical novelty to equitable impact. *PLOS Digit Health,* 3(4):e0000474.
25. Wang H, Fu T, Du Y, et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature,* 620(7972):47-60.
26. Beilby K, Hammarberg K. (2024). ChatGPT: a reliable fertility decision-making tool? *Human Reproduction,* 39(3):443-447.
27. Eoh KJ, Kwon GY, Lee EJ, et al. (2024). Efficacy of large language models and their potential in Obstetrics and Gynecology education.
28. Psilopatis I, Bader S, Krueckel A, Kehl S, Beckmann MW, Emons J. (2024). Can Chat-GPT read and understand guidelines? An example using the S2k guideline intrauterine growth restriction of the German Society for Gynecology and Obstetrics. *Gynecol Obstet,* 310(5):2425-2437.
29. Solmonovich RL, Kouba I, Quezada O, et al. (2024). Artificial intelligence generates proficient Spanish obstetrics and gynecology counseling templates. *AJOG Glob Rep,* 4(4):100400.
30. Malani SN, Shrivastava D, Raka MS. (2023). A Comprehensive Review of the Role of Artificial Intelligence in Obstetrics and Gynecology. *Cureus.*
31. Safiullina ER, Rychkova EI, Mayorova IV, et al. (2023). Application of digital methods and artificial intelligence capabilities for diagnostics in obstetrics and gynecology. CM, (27):111-117.
32. Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorghiou AT, Jouannic JM. (2022). Contributions of Artificial Intelligence Reported in Obstetrics and Gynecology Journals: Systematic Review. *J Med Internet Res*, 24(4):e35465.
33. Brandão M, Mendes F, Martins M, et al. (2024). Revolutionizing Women's Health: A Comprehensive Review of Artificial Intelligence Advancements in Gynecology. *JCM,* 13(4):1061.
34. Mysona DP, Kapp DS, Rohatgi A, et al. (2021). Applying Artificial Intelligence to Gynecologic Oncology: A Review. *Obstet Gynec,* 76(5):292-301.

35. Madani A, Namazi B, Altieri MS, et al. (2022). Artificial Intelligence for Intraoperative Guidance: Using Semantic Segmentation to Identify Surgical Anatomy During Laparoscopic Cholecystectomy. *Annals of Surge,* 276(2):363-369.
36. Laplante S, Namazi B, Kiani P, et al. (2023). Validation of an artificial intelligence platform for the guidance of safe laparoscopic cholecystectomy. *Surg Endosc,* 37(3):2260-2268.
37. Ryu S, Goto K, Imaizumi Y, Nakabayashi Y. (2024). Laparoscopic Colorectal Surgery with Anatomical Recognition with Artificial Intelligence Assistance for Nerves and Dissection Layers. *Ann Surg Oncol,* 31(3):1690-1691.
38. Freyer O, Wiest IC, Kather JN, Gilbert S. (2024). A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit Heal,* 6(9):e662-e672.
39. Bongurala AR, Save D, Virmani A, Kashyap R. (2024). Transforming Health Care with Artificial Intelligence: Redefining Medical Documentation. *Clinic Proceedings: Digital Health*, 2(3):342-347.
40. Rajagopal A, Ayanian S, Ryu AJ, et al. (2024). Machine Learning Operations in Health Care: A Scoping Review. *Mayo Clinic Proceedings: Digital Health,* 2(3):421-437.
41. Jin MF, Noseworthy PA, Yao X. (2024). Assessing Artificial Intelligence Solution Effectiveness: The Role of Pragmatic Trials. *Mayo Clinic Proceedings: Digital Health,* 2(4):499-510.
42. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. (2024). Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *The Lancet Digital Health,* 6(5):e367-e373.
43. Sadeghi Z, Alizadehsani R, Cifci MA, et al. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering,* 118:109370.
44. Beger J. (2024). The crucial role of explainability in healthcare AI. *European Journal of Radiology,* 176:111507.
45. Reeder K, Lee H. (2022). Impact of artificial intelligence on US medical students' choice of radiology. *Clin Imaging,* 81:67-71.
46. Cè M, Ibba S, Cellina M, et al. (2024). Radiologists' perceptions on AI integration: An in-depth survey study. *European Journal of Radiology,* 177:111590.
47. Jungmann F, Jorg T, Hahn F, et al. (2021). Attitudes Toward Artificial Intelligence Among Radiologists, IT Specialists, and Industry. *Academic Radiology,* 28(6):834-840.
48. Kather JN, Ferber D, Wiest IC, Gilbert S, Truhn D. (2024). Large language models could make natural language again the universal interface of healthcare. *Nat Med.*
49. Perivolaris A, Adams-McGavin C, Madan Y, et al. (2024). Quality of interaction between clinicians and artificial intelligence systems. A systematic review. *Future Healthcare Journal,* 11(3):100172.
50. Huang J, Yang DM, Rong R, et al. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digit Med,* 7(1):106.
51. Alkhalaf M, Yu P, Yin M, Deng C. (2024). Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics,* 156:104662.
52. Knudsen JE, Ghaffar U, Ma R, Hung AJ. (2024). Clinical applications of artificial intelligence in robotic surgery. *J Robotic Surg,* 18(1):102.
53. Zhang C, Hallbeck MS, Salehinejad H, Thiels C. (2024). The integration of artificial intelligence in robotic surgery: A narrative review. *Surgery,* S0039606024000680.

Cite this article: Martinelli C, Arrigo D, Meo V D, Giordano A, Ercoli A. (2025). Artificial Intelligence and Large Language Models in Medicine: A Critical Reflection on Current State, Challenges, and Future Implications. *Journal of Women Health Care and Gynecology*, BioRes Scientia Publishers. 5(2):1-8. DOI: 10.59657/2993-0871.brs.25.079

Article History: **Received:** December 23, 2024 | **Accepted:** January 17, 2025 | **Published:** January 25, 2025