Journal of Healthcare Leadership

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

# Generative Artificial Intelligence in Healthcare: Automation Bias, Deskilling, and Cognitive Implications – A Systematic Review

Fahad M Al-Anezi

Department Management Information Systems, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

Correspondence: Fahad M Al-Anezi, Department Management Information Systems, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, Email fmoalanezi@iau.edu.sa



**Background:** Generative artificial intelligence (GenAI) is rapidly transforming eHealthcare, promising substantial gains in diagnostic accuracy, workflow efficiency, and decision support, yet raising concerns about automation bias and clinician deskilling that may erode core diagnostic expertise and professional judgment over time.

**Aim and Objectives:** This systematic review investigates the dual impact of GenAI in eHealthcare, focusing on how it enhances clinical efficiency and decision support while potentially diminishing clinicians' diagnostic expertise and professional judgment.

**Methods:** A systematic search was conducted across PubMed, Scopus, Web of Science, and Google Scholar. A total of 11,269 records were identified, and 29 studies met inclusion criteria following PRISMA 2020 guidelines. Studies were synthesized using a narrative approach due to heterogeneity in design, clinical domains, and AI system maturity.

**Results:** Across the included studies, recurrent themes included automation bias (reported in 10 studies), concerns regarding clinician deskilling (9 studies), and impacts on diagnostic reasoning (9 studies). Evidence was predominantly observational, experimental simulation-based, or conceptual in nature. Findings reveal that GenAI significantly improves diagnostic accuracy, workflow efficiency, and decision quality across radiology, telehealth, and education domains. However, overreliance introduces risks of automation bias, cognitive deskilling, and loss of interpretive autonomy. The review identifies key mitigation strategies, including human-in-the-loop frameworks, explainable AI (XAI), ethical governance, and continuous clinician reskilling. Theoretically, the results redefine the clinician–AI relationship as a dynamic cognitive partnership, while practically they emphasize responsible integration through education and regulation.

**Conclusion:** Sustainable adoption of GenAI demands balanced implementation—leveraging its analytical capabilities without compromising human judgment, clinical reasoning, or professional accountability.

**Keywords:** clinical decision-making, cognitive load, human–AI interaction, diagnostic reasoning, professional competence, large language models, healthcare governance, human-in-the-loop systems

## Introduction

Generative AI (GenAI) in healthcare refers to intelligent systems that can autonomously generate outputs such as clinical notes, patient summaries, algorithmically derived treatment suggestions, and simulated patient scenarios.[1] Unlike conventional AI models that classify or predict outcomes based solely on input data, GenAI can produce original, adaptive responses that align with complex clinical contexts. Electronic healthcare (eHealthcare) serves as the broader ecosystem where these tools operate, integrating digital health technologies to facilitate remote consultations, interoperable health records, and AI-powered decision support.[2] For clarity, this review distinguishes generative AI (eg, large language models capable of producing novel text or multimodal outputs) from traditional predictive or discriminative AI systems, which classify, detect, or forecast outcomes without generating new content. The cognitive and workflow implications of generative systems may differ substantively from earlier AI paradigms Crucial to this discussion are concepts such as *diagnostic expertise*—the clinician's capacity to synthesize patient history, symptoms, and evidence to

https://doi.org/10.2147/JHL.S590498
Received: 20 December 2025
Accepted: 24 May 2026
Published: 9 July 2026

reach accurate conclusions—and *professional judgment*, which combines technical knowledge with ethical and contextual considerations in patient care. Another relevant term, *automation bias*, refers to the tendency to over trust machine outputs while undervaluing human reasoning, a cognitive vulnerability that may become more prevalent with GenAI dependence.[3,4]

GenAI has emerged as one of the most disruptive innovations in healthcare, particularly within the evolving domain of eHealthcare. As a subset of advanced artificial intelligence technologies, GenAI is capable of generating novel and contextually relevant outputs—ranging from diagnostic reports to patient-specific treatment recommendations—based on complex patterns learned from large and diverse datasets.[5] eHealthcare, which encompasses telemedicine, mobile health (mHealth) applications, electronic health records (EHR), and AI-assisted clinical decision support systems, has experienced a rapid infusion of GenAI-driven solutions aimed at improving clinical efficiency, diagnostic accuracy, and workflow management. The integration of GenAI into these systems holds enormous potential to redefine how practitioners deliver care, optimize resources, and make data-driven clinical decisions.[4,6]

The promise of GenAI in eHealthcare is most visible in areas such as clinical workflow optimization, automated analysis of medical imagery, predictive analytics for disease progression, and improved patient risk stratification. These capabilities can reduce administrative burden, provide early detection signals, offer comprehensive decision support, and deliver evidence-based treatment suggestions within seconds.[7–9] For example, radiological imaging analysis augmented by GenAI can detect subtle anomalies that human eyes may miss, and predictive algorithms can forecast patient deterioration before clinical symptoms escalate. Such tools can significantly increase throughput, shorten diagnostic timelines, and enhance the precision of clinical decisions.[10] In resource-constrained healthcare environments, these efficiencies can directly translate into improved patient safety, better resource allocation, and reduced costs.[11,12]

Yet, alongside this transformative potential, there is mounting evidence that GenAI may inadvertently undermine clinicians' core competencies. Over time, healthcare professionals accustomed to AI-generated outputs may begin to engage less deeply in independent problem-solving or critical diagnostic reasoning. This erosion of skillsets poses a significant professional risk, especially in situations where AI tools are absent, malfunctioning, or produce biased outputs.[13,14] Furthermore, the cognitive ease offered by automation can cultivate overdependence, leading to a detachment from the meticulous verification processes that safeguard clinical accuracy. Such trends can diminish practitioner autonomy and weaken resilience in non-digitally supported practice contexts, ultimately affecting the quality of care.[15] Thurzo & Varga[16] highlighted how generative AI is reshaping the scientific synthesis process itself, influencing literature generation, review methodologies, and research workflows. This recursive effect—where AI tools both shape and are evaluated within scientific production—introduces new epistemic considerations for systematic review methodology.

The significance of studying GenAI's role in eHealthcare lies in its dual nature: while it clearly offers efficiency gains and advanced decision-support capabilities, the same technology may inadvertently contribute to a decline in clinicians' foundational medical knowledge, diagnostic expertise, and professional judgment. This paradox—simultaneous advancement and erosion—has created an important debate among healthcare researchers, policymakers, and practitioners. On one hand, GenAI promises faster data processing, real-time analytics, and improved diagnostic completeness.[11,12] On the other, there is growing concern that sustained reliance on machine-generated recommendations could foster automation bias, decrease cognitive engagement, and weaken the intellectual rigor that underpins effective medical practice.[4,15]

Existing scholarly literature[17–22] on GenAI's application in healthcare often focuses almost exclusively on technical performance indicators such as accuracy rates, time efficiency, and comparative error reduction against human benchmarks. Far fewer studies[13,14] have explored the long-term cognitive and professional implications for clinicians operating in AI-integrated environments. There is a lack of comprehensive evidence examining both sides of the equation, the measurable benefits and the subtle but consequential risks, within a single analytical framework. Additionally, longitudinal data tracking the evolution of clinician skills over extended periods of GenAI use is scarce, leaving major questions unanswered about sustainability and knowledge retention. While some researchers have acknowledged these risks in passing, systematic exploration of mitigation strategies is still limited, resulting in fragmented guidelines for safe and balanced integration.[3,23]

This systematic review directly addresses these gaps by synthesizing peer-reviewed evidence published between 2015 and 2025 on GenAI's dual impact in eHealthcare. The aim is to critically evaluate how GenAI simultaneously enhances efficiency and decision quality while potentially undermining clinician expertise and judgment. By examining diverse clinical contexts, technological implementations, and reported outcomes, the study will produce an integrative understanding that can inform policy recommendations, technology design considerations, and targeted professional training initiatives aimed at preserving essential medical competencies.[23] Despite rapid adoption, gaps remain in understanding how generative AI may influence clinician cognitive processes, professional skill maintenance, and diagnostic reasoning. This review is intended to inform clinicians, healthcare educators, system leaders, and policymakers involved in AI governance and implementation. Accordingly, this review addresses the following research questions:

RQ1: What efficiency and decision-support enhancements have been demonstrated by GenAI in eHealthcare clinical practice?

RQ2: What evidence exists for negative impacts of GenAI reliance on clinician diagnostic expertise, foundational medical knowledge, and professional judgment?

RQ3: What recommendations and mitigation strategies have been proposed to balance GenAI's benefits with preservation of clinician expertise in practice?

# Methodology

This study conducts a thorough systematic literature review to analyze the integration, advantages, and challenges of GenAI in eHealthcare clinical practice. The review critically examines the dualistic effects of GenAI, highlighting its capacity to improve efficiency and clinical decision support while simultaneously undermining clinician expertise, autonomy, and diagnostic judgment. The methodology adheres to the structured, transparent, and reproducible principles established by Tranfield et al[24] and Denicol,[25] thereby guaranteeing methodological rigor and reducing bias in the sourcing, screening, extraction, and synthesis of pertinent studies. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[26] makes sure that the review is reliable and can be repeated in both medical and technological fields.

## Data Sources and Search Strategy

The literature search was conducted across six electronic databases: PubMed, Scopus, Web of Science, ScienceDirect, IEEE Xplore, and Google Scholar. Searches covered publications from January 2015 to June 2025 to capture both early AI developments and recent generative AI applications. The search strategy combined controlled vocabulary terms and free-text keywords related to generative artificial intelligence and healthcare applications, using Boolean operators (AND, OR, NOT). Search strings included:

PubMed/Scopus/Web of Science:

- ("Generative Artificial Intelligence" OR "GenAI" OR "Large Language Models" OR "LLMs" OR "ChatGPT" OR "GPT-4" OR "Med-PaLM" OR "BioGPT") AND ("Healthcare" OR "eHealthcare" OR "Clinical Decision Support" OR "Diagnostic Reasoning" OR "Clinical Workflow" OR "Medical Documentation" OR "Telemedicine" OR "eHealth") AND ("Automation Bias" OR "Deskilling" OR "Cognitive Offloading" OR "Overreliance" OR "Diagnostic Expertise" OR "Professional Judgment")

IEEE Xplore/ScienceDirect

- ("Generative AI" OR "LLM" OR "ChatGPT") AND ("healthcare" OR "clinical" OR "medical") AND ("bias" OR "deskilling" OR "expertise" OR "judgment")

Google Scholar

- allintitle: ("generative AI" OR "ChatGPT" OR "LLM") ("healthcare" OR "clinical") ("bias" OR "deskilling" OR "expertise")

All searches used quotation marks for exact phrase matching and were limited to peer-reviewed articles in English. The final search was executed on July 15, 2025. Supplementary hand-searching of reference lists from included studies and relevant reviews identified no additional records.

## Inclusion and Exclusion Criteria

### Inclusion Criteria

- Peer-reviewed journal articles published in English between January 2015 and June 2025, reflecting the emergence and maturity of GenAI in healthcare.
- Studies focusing on clinical applications of GenAI (eg, ChatGPT, Med-PaLM, BioGPT).
- Empirical, conceptual, or review articles addressing GenAI's impact on efficiency, decision support, clinician expertise, or judgment.
- Research from medical, health informatics, and AI-engineering domains.

### Exclusion Criteria

- Studies focusing exclusively on non-generative predictive or classification AI systems were excluded unless a generative component was explicitly evaluated.
- Non-peer-reviewed publications (editorials, short communications, abstracts, and commentaries without empirical or conceptual depth).
- Studies unrelated to clinical or eHealthcare settings (eg, AI in finance).
- Non-English papers and duplicates.
- Retracted papers or those lacking methodological transparency.

The years ranging from 2015 to 2025 were chosen because they are the same time as the Fourth Industrial Revolution and the rapid spread of GenAI in clinical settings. This gives a balanced view of both early and mature uses.

## Screening and Selection Process

The PRISMA flow model's four stages—identification, screening, eligibility, and inclusion—were used in the multi-stage screening process (Figure 1). After deduplication, the first search found 11269 records. Using the publication year filter got rid of 4179 records, leaving 5653. After filtering by publication type, this number went down to 3837 records. There were 3734 records left after removing 103 articles that were not in English. We used the Scimago Journal Rank (SJR), and also the citation metrics and impact factor to check the quality of the journals and kept only the Q1-Q4 studies.

This rigorous screening reduced the dataset down to 2217 articles. Duplicate and ineligible articles, were removed leaving with a final dataset of 1686 articles. The title and abstract screening was conducted independently by two reviewers using standardized inclusion/exclusion criteria. Each reviewer screened 50% of records, with 20% overlap for calibration. Inter-rater agreement was substantial (Cohen's $\kappa = 0.87$). Discrepancies (n=42) were resolved through discussion until 100% consensus was reached.

During the full-text screening stage, 84 articles were independently assessed by two reviewers, resulting in the exclusion of 55 studies for the following documented reasons: 22 studies focused exclusively on non-generative AI systems (predictive/classification models without text/image generation capabilities); 14 were non-peer-reviewed materials such as editorials, letters to the editor, or conference abstracts lacking empirical or conceptual depth; 9 examined AI applications outside clinical or eHealthcare settings (eg, AI in finance, agriculture, or non-medical domains); 5 represented duplicate publications already captured earlier in the screening process; 3 were either retracted papers or lacked sufficient methodological transparency for quality appraisal; and 2 were non-English language publications. Final

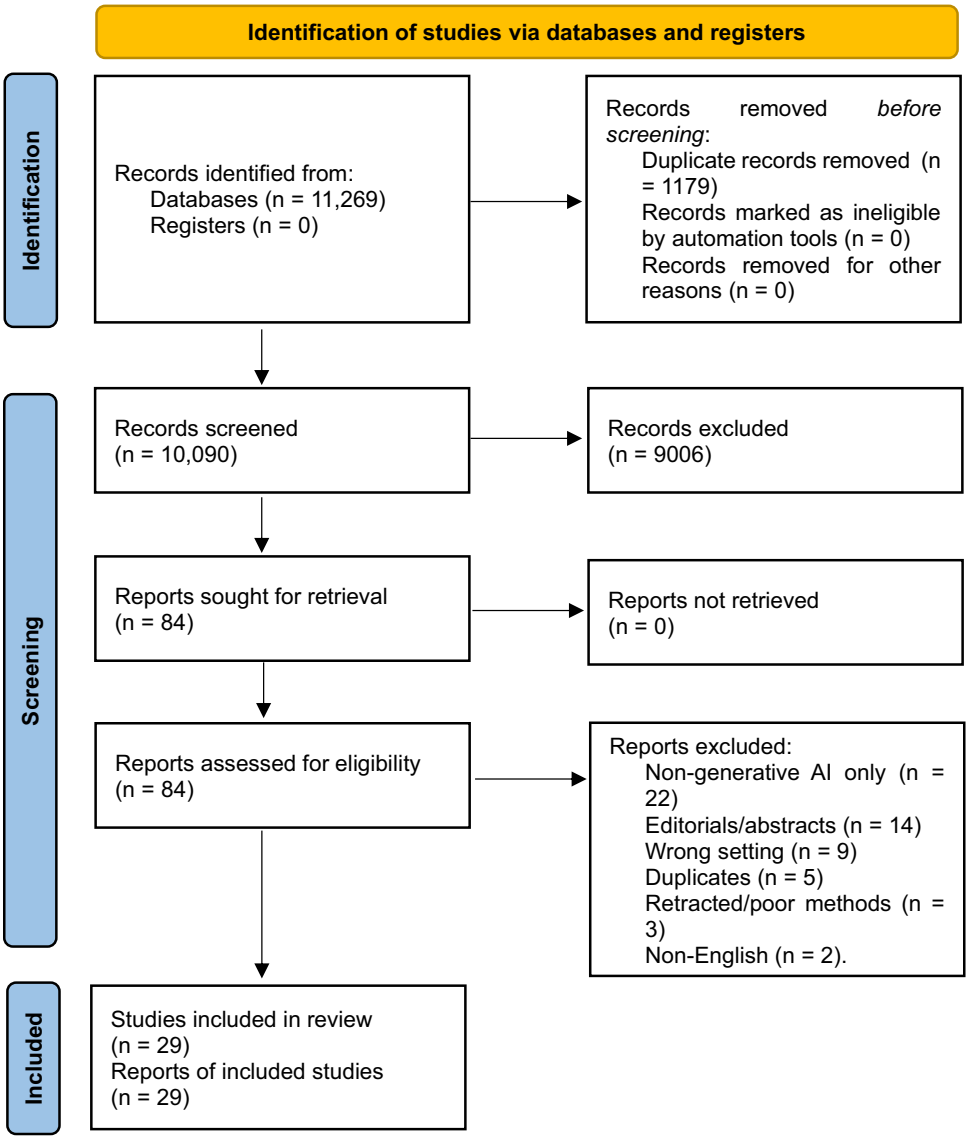


**Figure 1** Screening and selection process using PRISMA.

inclusion decisions were made by consensus, resulting in 29 studies for synthesis. The second reviewer provided written consent to be acknowledged for screening support without co-authorship.

## Data Extraction and Analysis

Data extraction was performed independently by the primary author using a standardized extraction form (Appendix A). The form captured: study identification, publication details, study characteristics, population/sample, GenAI tool details, comparator/context, outcomes relevant to each research question (RQ1-RQ3), methodological quality notes, and key findings. A second reviewer independently verified data extraction for 20% of studies (n=6) to ensure consistency. No discrepancies were identified. Extracted data were synthesized into Table B1 (Appendix B) for transparency and comparability across studies. The evidence from the studies are analyzed in relation to the three research questions. Emergent dimensions such as efficiency enhancement, automation bias, deskilling, trust and interpretability, and governance frameworks, were considered in the analysis of findings.

To strengthen analytical validity, findings were triangulated with insights from conceptual frameworks in cognitive psychology, human–AI interaction theory, and medical professionalism. Thematic patterns were then integrated into three overarching analytical categories corresponding to RQ1–RQ3, ensuring alignment between empirical evidence and theoretical interpretation.

The included studies demonstrated substantial heterogeneity in study design (experimental, observational, simulation-based, and conceptual analyses), clinical domains (eg, radiology, primary care, mental health), AI system maturity (prototype versus deployed systems), and outcome measures. Due to this variability, quantitative meta-analysis was not appropriate. Instead, a structured narrative synthesis approach was employed to identify recurring themes and patterns across studies.

## Quality Assessment

The methodological robustness of each included study was appraised using the Critical Appraisal Skills Programme (CASP) checklist, assessing credibility, transferability, dependability, and confirmability. Empirical studies were evaluated for sample adequacy and bias control, while conceptual and review studies were assessed for coherence and theoretical contribution. CASP appraisal (Appendix C and Table C1) indicated generally moderate-to-high methodological quality across included studies. Fourteen studies were rated high quality (≥8/10), and fifteen were rated moderate quality (6–7/10). No studies scored ≤5. Higher scores were most frequently observed among randomized and experimental diagnostic studies (eg,[27,28]), while conceptual, viewpoint, and narrative publications typically met moderate methodological criteria but lacked empirical validation or longitudinal data. Common methodological limitations included reliance on simulation-based designs, absence of long-term outcome assessment, and limited real-world implementation data. Only studies scoring above 70% quality threshold were retained in the final synthesis.

# Results

A total of twenty-nine peer-reviewed studies published between 2022 and 2025 were included in this systematic review, representing a diverse range of empirical, systematic, conceptual, and mixed-method investigations on the role of GenAI in eHealthcare. The dataset encompassed experimental and observational studies examining model performance and diagnostic decision support,[14,27,29–31] systematic and scoping reviews evaluating broader evidence bases,[11,32–35] and conceptual or theoretical contributions exploring cognitive, ethical, and professional dimensions of AI integration.[36–40] Collectively, these studies addressed clinical contexts including radiology, neurology, endoscopy, telehealth, surgery, and medical education, with sample sizes ranging from small experimental cohorts (eg, 27 radiologists in Dratsch et al[41]) to large multi-study syntheses (eg, 161 systematic reviews in Morone et al[11]) GenAI applications spanned diagnostic reasoning, documentation automation, educational assessment, and decision-support systems, enabling a comprehensive evaluation of both technological efficiency and its cognitive and ethical ramifications for clinicians. The findings in this review are discussed in relevance to the three research questions outlined in the introduction section.

## RQ1: Efficiency and Decision-Support Enhancements

Findings indicated that GenAI significantly enhances clinical efficiency, diagnostic support, and operational throughput. Systematic and comparative analyses demonstrate that GenAI, especiallyarchitectures such as GPT-4, has achieved measurable gains in accuracy, reasoning quality, and documentation productivity.

Sadeghi et al[32] and Morone et al[11] discovered that explainable and machine-learning-based systems enhanced diagnostic precision and predictive accuracy across various specialties, yielding efficiency improvements of 40–45% in diagnostic and prognostic tasks. Iqbal et al[35] synthesized 17 reviews and corroborated workflow acceleration in documentation, patient communication, and literature synthesis, whereas Bhuyan et al[3] highlighted time savings and cost reduction via automated charting and scheduling.

Experimental studies yielded more detailed insights. Ueda et al[30] demonstrated that GPT-4 exhibited superior consistency and information-retrieval accuracy in radiology compared to tested models, surpassing both GPT-3.5 and Llama 2. Yu et al[27] similarly showed that AI-assisted radiologists were better at finding problems on chest X-rays, but the results were different for each person. Shieh et al[31] found that ChatGPT-4 was 87% accurate in clinical education and reasoning, while 3.5 was only 48% accurate. Beheshti et al[34] found that it was 73% accurate across 128 evaluations, which shows that it is reliable for tasks like retrieving information and answering questions.

Performance advantages transcend diagnostics. Goodell et al[42] showed that tool-augmented models that combined retrieval-augmented generation (RAG) and external computation APIs cut down on medication-dose mistakes by as much as 88%. Budzyn et al[14] reported shorter endoscopic procedures and enhanced lesion detection, whereas Ranji[43] documented a 30% decrease in diagnostic latency for endocrine disorders. Gomez et al[28] documented enhanced diagnostic accuracy in telehealth when clinicians employed explainable AI (XAI) decision-support tools for strep throat screening, highlighting AI's significance in remote triage and the enhancement of primary care.

Narrative and conceptual contributions bolster these quantitative advancements. Najjar[44] underscored the significance of AI in image segmentation and radiomics-based predictive modeling Mittermaier et al[45] and Natali et al[13] stressed the importance of objective skill assessment and workflow standardization; and Bongurala et al[37] recognized documentation automation as a crucial factor in enhancing clinician productivity. These studies collectively indicate that GenAI enhances decision-support ecosystems by facilitating expedited data synthesis, alleviating administrative burdens, and broadening diagnostic capabilities.[29,38]

Nonetheless, efficiency seems to be contingent upon context. Arvai et al[33] and Banerji et al[36] observed that efficiency emerges solely when models are integrated into established clinical pathways and supported by clinician involvement. Abhari et al[46] and Mosqueira-Rey et al[47] similarly found that human-in-the-loop (HITL) designs produce models that are more adaptable and aware of their surroundings than fully automated ones. So, even though GenAI clearly speeds up diagnostic and administrative tasks, long-term efficiency gains need human oversight and co-design.

## RQ2. Negative Impacts on Diagnostic Expertise and Professional Judgment

Even though it looked beneficial, almost every study warned against cognitive and professional decline that could happen if people relied too much on AI. Tikhomirov et al[40] provided a foundational theoretical framework, differentiating between technological and cognitive efficiency, and cautioned against a "AI chasm" that separates clinicians from interpretive reasoning. Similar issues have been found in other studies: Dratsch et al[41] showed that radiologists of all experience levels were biased toward automation, and Goh et al[48] showed that clinicians often put too much faith in wrong LLM suggestions.

Numerous studies demonstrated the presence of deskilling, a prevalent theme in GenAI literature. The available literature reflects both empirical findings and theoretical interpretations. Where empirical evidence is presented, findings derive primarily from experimental simulations, observational studies, and short-term evaluations. Conceptual discussions regarding long-term deskilling and cognitive erosion remain largely theoretical and warrant further longitudinal investigation. Budzyn et al[14] recorded diminished interpretive engagement among endoscopists and highlighted real-world deskilling, showing reductions up to 20% in unassisted colonoscopy diagnosis accuracy following AI adoption.; Natali et al[13] characterized "upskilling inhibition"; and Khan[39] associated AI utilization in neurodiagnostics with deteriorating pattern-recognition abilities. This trend is also seen in schools: Arvai et al[33] found that doctors were worried about losing skills and feeling stressed out by technology, and Shieh et al[31] said that relying too much on ChatGPT in training could hurt independent clinical reasoning. While several studies report associations between AI assistance and changes in clinician decision-making behavior, direct longitudinal evidence demonstrating sustained professional deskilling remains limited.

Bias and excessive confidence pose additional risks to the integrity of judgment. Hasanzadeh et al[4] delineated automation, feedback-loop, and dataset biases that can perpetuate inequities and mislead decision-making. Hirosawa et al[29] noted that GPT-4 did not achieve the accurate final diagnosis in 16% of instances, highlighting the constraints of pattern-based reasoning lacking contextual interpretation. Beheshti et al[34] measured inconsistencies in medical fields, showing that accuracy could range from 60% to 85% depending on how hard the task was. Gomez et al[28] introduced a behavioral aspect, demonstrating that even precise AI outputs led clinicians to excessive testing due to distrust, thereby highlighting the paradox of confidence calibration.

Cognitive over-delegation is also evident in theoretical contributions. Choudhury & Chaudhry[38] elucidated a "self-referential learning loop" wherein persistent human reliance on AI not only diminishes clinicians' skills but also taints model retraining data, thereby compromising overall epistemic quality. Tikhomirov et al[40] and Banerji et al[36] identified the risk of ecological debounding, which refers to the disconnection of clinical reasoning from real-world variability when practitioners depend on abstracted AI representations.

Lastly, ethical and organizational issues make individual cognitive risks worse. Arab et al[49] discovered that AI models validated in trials frequently exhibit suboptimal performance in real-world conditions due to data heterogeneity, while Bhuyan et al[3] cautioned against privacy violations and inadequate governance mechanisms. Bongurala et al[37] emphasized that excessive automation of documentation could undermine the narrative and empathetic aspects of care. These findings collectively indicate that GenAI, despite its technological robustness, may unintentionally undermine essential diagnostic reasoning, situational awareness, and moral responsibility if not properly regulated or contextualized.

## RQ3. Recommendations and Mitigation Strategies

The reviewed studies articulate a consistent set of strategies to preserve clinician expertise while leveraging GenAI's advantages. Three main areas of mitigation are: human-centered design, education and governance, and evaluation and transparency.

The most common suggestion for a safeguard is human-centered design. Sadeghi et al,[32] Abhari et al,[46] and Mosqueira-Rey et al[47] contended that explainability and human-in-the-loop (HITL) mechanisms enhance trust and interpretability, thereby mitigating automation bias. Banerji et al,[36] and Budzyn et al[14] supported clinician participation throughout AI lifecycles, from data curation to validation to guarantee contextual relevance. Similarly, Goodell et al[42] showed that adding human verification loops to tool-augmented LLMs made them much more reliable when it came to numbers.

Education and governance are very important. Arvai et al[33] suggested the implementation of structured AI literacy curricula and psychological support to mitigate technostress, whereas Beheshti et al[34] and Shieh et al[31] advocated for the integration of prompt-engineering and critical-evaluation training into medical education. Hasanzadeh et al[4] proposed regulatory imperatives via a model-life-cycle bias-mitigation framework, while Iqbal et al[35] advocated for ethical oversight and transparency policies prior to complete clinical integration. Arab et al[49] and Natali et al[13] subsequently introduced comprehensive governance models (AI-HIF and reskilling frameworks) that align with CFIR and TAM principles to reconcile innovation with professional integrity.

Evaluation and transparency concentrate on the standardization of technical and methodological practices. Morone et al[11] presented the CLASMOD-AI and PRISMA-AI frameworks to enhance methodological rigor in AI reporting. Beheshti et al[34] proposed an Adjusted Accuracy Metric to standardize LLM performance evaluations, whereas Ueda et al[30] and Yu et al[27] advocated for adaptive calibration systems that customize AI support to clinician proficiency. All of these strategies point to the need for ongoing audits, learning that includes feedback, and open benchmarking to keep clinicians accountable and involved.

Ethical and epistemic considerations enhance these technical directives. Tikhomirov et al[40] advocated for AI design informed by cognitive science that maintains reflective reasoning, while Choudhury & Chaudhry[38] emphasized the necessity of algorithmic accountability laws to mitigate epistemic dependency. Khan[39] and Bongurala et al[37] underscored the necessity of curriculum reforms that maintain core interpretive and communication skills as fundamental components of training. The overall message is clear: improvements in efficiency should not come at the expense of moral agency, clinical cognition, or patient-centered care.

# Discussion

The studies collectively presented a dual narrative. GenAI shows measurable benefits in terms of numbers, such as speeding up diagnoses, making paperwork easier, and improving decision support. Qualitatively, its unregulated utilization poses a threat to professional de-skilling and epistemic complacency. The literature indicates the development of a new equilibrium model: AI functioning as a cognitive enhancer rather than a cognitive substitute. Empirical evidence[27,29,42] indicated that clinician performance is optimized when AI outputs are analyzed rather than uncritically accepted. This corroborates the socio-technical framework posited by Tikhomirov et al[40] and Banerji et al,[36] which asserts that optimal decision-making results from distributed cognition between humans and machines. On the other hand, proof of automation bias and cognitive off-loading shows that cultures and institutions need to change so that professional judgment stays important even as AI technology gets better. Given evidence of automation bias reported across multiple studies included in this review, structured human-in-the-loop oversight mechanisms may mitigate overreliance on AI-generated outputs.

Methodological reviews[11,35] identified an additional challenge: inconsistent evaluation standards obstruct generalisability and regulatory oversight. The distinction between augmentation and substitution remains ambiguous in the absence of standardized metrics. In the same way, ethical and psychological aspects[4,33] show that clinician trust and identity need to be actively built through open governance and clear AI design. Recent evidence[50] on provable AI ethics frameworks proposes technical architectures for embedding explainability, constraint validation, and ethical guardrails directly into AI systems rather than relying solely on policy-level oversight. Such "ethical firewall" approaches may offer a more operational pathway toward trustworthy clinical AI deployment.

In summary, GenAI's incorporation into eHealthcare ought to adhere to a "co-evolutionary" framework—enhancing efficiency while fortifying, rather than undermining, the clinician's cognitive ecosystem. Future implementation must actualize human-centered frameworks, longitudinal skill retention studies, and transparent, verifiable model architectures. Healthcare can only realize GenAI's transformative potential by aligning technological acceleration with professional stewardship, without undermining its epistemic foundations. Considering observed variability in AI performance across clinical contexts, governance frameworks and monitoring systems are recommended to ensure safe implementation.

## Synthesis and Implications

The synthesis of findings across twenty-nine studies revealed a complex duality in the role of GenAI within eHealthcare: it acts both as a catalyst for efficiency and precision and as a potential disruptor of clinical expertise and judgment. This ambivalence carries significant theoretical and practical implications for how technology, cognition, and professional autonomy interact in modern medical practice.

### Theoretical Implications

The findings enhanced the theoretical understanding of how GenAI transforms the cognitive and epistemic foundations of clinical decision-making. Research by Tikhomirov et al,[40] Banerji et al,[36] and Choudhury & Chaudhry[38] emphasizes the necessity to redefine the clinician–AI relationship as a dynamic cognitive partnership instead of a tool–user hierarchy. This change means that traditional ideas about medical expertise and diagnostic reasoning, which are based on experience and implicit judgment, need to change to include augmented cognition and distributed intelligence. Tikhomirov et al[40] suggest a difference between technological efficiency (AI performance) and cognitive efficiency (human interpretive capacity). This is an important theoretical lens for judging the depth, not just the speed, of diagnostic reasoning. Moreover, frameworks developed in studies[4,11,35] enhance methodological theory by incorporating AI ethics, transparency, and bias mitigation into evidence appraisal models like PRISMA-AI and CLASMOD-AI. These studies collectively emphasize that GenAI should not be regarded solely as a technology that enhances efficiency but as a component of a cognitive ecosystem that interacts with human judgment, institutional norms, and socio-technical infrastructures. This theoretical reframing necessitates a multidisciplinary research paradigm that integrates cognitive psychology, medical epistemology, and human–AI interaction to elucidate how clinicians develop, validate, and sustain expertise in AI-mediated contexts.

### Practical Implications

The review delineates several pragmatic strategies for healthcare organizations, educators, and policymakers to responsibly incorporate GenAI into clinical practice. Studies[27,29,30,42] provide empirical evidence that GenAI tools can significantly improve clinical accuracy, workflow efficiency, and decision-support capacity when implemented under structured supervision. Nonetheless, research conducted in[13,14,33] warns that prolonged exposure without intentional re-skilling could result in "AI-induced deskilling". So, institutions need to make sure that AI literacy, continuous competency development, and human-in-the-loop protocols are all part of their daily clinical work.

In practice, the review supports hybrid decision-support systems that find a balance between automation and human oversight. In these systems, AI-generated outputs are used as advice, not as final decisions. Ethical governance frameworks, like the ones suggested in,[4,49] stress the importance of clear rules, accountability systems, and algorithmic transparency in order to protect clinician independence and patient trust. Education systems ought to integrate AI fluency into medical curricula, as proposed by Arvai et al[33] and Shieh et al,[31] to guarantee that future clinicians can critically evaluate AI recommendations instead of merely accepting them.

On a larger scale, practical effects include policy and system design. Healthcare organizations ought to adopt explainable AI interfaces,[32] uphold data-diversity standards,[4] and institute continuous audit mechanisms[34] to ensure that AI deployment aligns with patient safety and professional development goals. In short, the main point of this review is to help healthcare leaders choose a GenAI adoption method that is based on evidence, ethical, and human-centered, while also making things more efficient and protecting professional judgment.

## Limitations and Future Directions for Research

The exclusion of non-English publications (n = 103) represents a potential source of language bias. Given the global nature of generative AI development, relevant contributions from non-English-speaking regions may not have been captured. Future reviews incorporating multilingual search strategies could provide a more globally representative synthesis.

Future studies should transcend experimental accuracy evaluations to examine the longitudinal effects of ongoing engagement with Generative AI on clinical cognition, autonomy, and professional judgment. Longitudinal and real-world studies are necessary to examine the behavioral, cognitive, and organizational effects of prolonged reliance on AI in various clinical settings. Empirical research must prioritize the creation of validated and multidimensional metrics for AI-induced deskilling, encompassing both cognitive and procedural dimensions of expertise attrition. Experimental and simulation-based methodologies may elucidate the progression of automation bias and overreliance across varying degrees of clinical experience and task complexity.

Future research should employ longitudinal randomized designs comparing clinicians using AI-assisted workflows versus independent diagnostic reasoning over extended periods. Experimental protocols could measure retention of differential diagnosis accuracy, time-to-decision without AI access, and susceptibility to automation bias under varying AI reliability conditions. Such designs would allow quantification of epistemic quality, defined as the clinician's ability to independently generate accurate and contextually grounded diagnostic hypotheses. This approach would provide clearer thresholds for safe AI integration and clarify whether observed performance shifts represent temporary adaptation effects or sustained cognitive change.

Another important area is trust calibration and interpretability. Future research should investigate the impact of explainability features, uncertainty visualization, and user-interface design on clinicians' confidence, critical engagement, and corrective judgment, drawing on findings from Gomez et al[28] and Yu et al[27] Interdisciplinary research that combines ethics, human factors, and policy will also be important for creating governance models that strike a balance between professional responsibility and technological efficiency. This entails assessing regulatory instruments like the Algorithmic Accountability Act and conducting empirical evaluations of hybrid human–AI decision frameworks within actual healthcare environments.

# Conclusion

This systematic review demonstrates that GenAI represents a transformative force in eHealthcare, simultaneously enhancing efficiency, diagnostic accuracy, and decision support while challenging the foundations of clinical expertise and professional judgment. Across the 29 reviewed studies, evidence consistently shows that GenAI can streamline workflows, reduce administrative burden, and augment complex reasoning tasks, yet excessive reliance may erode clinicians' cognitive engagement, autonomy, and interpretive depth. The findings underscore that the future of medical practice depends not merely on technological capability but on the design of human-centred, explainable, and ethically governed AI ecosystems that sustain professional competence and accountability.

To ensure GenAI serves as an intelligent collaborator rather than a substitute for human reasoning, healthcare systems must invest in continuous education, transparent AI integration, and regulatory safeguards. Evidence suggests that clinician performance may degrade when incorrect AI outputs are accepted without independent verification, particularly in simulated diagnostic contexts. Structured safeguards—such as mandatory independent reasoning prior to AI exposure, calibrated trust training, and performance monitoring during AI-assisted tasks—may mitigate overreliance. Experimental research should quantify thresholds of AI reliance (eg, percentage of cases using AI suggestions) and measure longitudinal skill retention to determine whether cognitive degradation is transient or persistent.

Ultimately, the responsible adoption of GenAI requires operationally defined balance—leveraging analytical augmentation while preserving clinician-led interpretive authority. While early evidence suggests associations between generative AI use and automation bias, altered diagnostic reasoning patterns, and concerns regarding deskilling, the current literature remains predominantly short-term, heterogeneous, and observational. Robust longitudinal, real-world, and threshold-based experimental studies are necessary to clarify long-term cognitive and professional impacts and to establish evidence-based parameters for safe and sustainable AI integration in clinical practice.

## Highlight

This review highlights that generative AI can meaningfully improve diagnostic accuracy, workflow efficiency, and decision support across diverse clinical settings, but requires human-centred design, governance, and continuous clinician reskilling to prevent deskilling and protect safe, accountable patient care.

## Ethics Statement

No human subjects were involved in the study.

## Funding

No funding has been received for this study.

## Disclosure

The author reports no conflicts of interest in this work.

## References

1. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Bus Inf Syst Eng*. 2023;66(1):111–126. doi:10.1007/s12599-023-00834-7
2. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary evidence of the use of generative AI in health care clinical services: systematic narrative review. *JMIR Med Inform*. 2024;12e52073. doi:10.2196/52073
3. Bhuyan SS, Sateesh V, Mukul N, et al. Generative artificial intelligence use in healthcare: opportunities for clinical excellence and administrative efficiency. *J Med Syst*. 2025;49(1):10. doi:10.1007/s10916-024-02136-1
4. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digit Med*. 2025;8(1):154. doi:10.1038/s41746-025-01503-7
5. Chatterjee S, Fruhling A, Kotiadis K, Gartner D. Towards new frontiers of healthcare systems research using artificial intelligence and generative AI. *Health Syst*. 2024;13(4):263–273. doi:10.1080/20476965.2024.2402128
6. Chen Y, Esmaeilzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges. *J Med Internet Res*. 2024;26 (e53008):e53008. doi:10.2196/53008
7. Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJC. Generative AI for transformative healthcare: a comprehensive study of emerging models, applications, case studies, and limitations. *IEEE Access*. 2024;12:31078–31106. doi:10.1109/ACCESS.2024.3367715
8. Rashidieranjbar F, Farhadi A, Zamanifar A. Revolutionizing healthcare with generative artificial intelligence technologies. In: *Generative Artificial Intelligence (AI) Approaches for Industrial Applications*. Springer; 2025:189–221. doi:10.1007/978-3-031-76710-4_10
9. He R, Sarwal V, Qiu X, et al. Generative AI models in time varying biomedical data: a systematic review. *J Med Internet Res*. 2024;26. doi:10.2196/59792
10. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat Med*. 2025;31(8):2546–2549. doi:10.1038/s41591-025-03727-2
11. Morone G, De Angelis L, Cinnera AM, 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
12. Xu R, Wang Z. Generative artificial intelligence in healthcare from the perspective of digital media: applications, opportunities and challenges. *Heliyon*. 2024;10(12):e32364. doi:10.1016/j.heliyon.2024.e32364
13. Natali C, Marconi L, Duran LDD, Cabitza F. AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. *Artif Intell Rev*. 2025;58(11). doi:10.1007/s10462-025-11352-1
14. Budzyń K, Romańczyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol Hepatol*. 2025. doi:10.1016/S2468-1253(25)00133-5
15. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. *J Saf Sci Resil*. 2024;5(4):460–469. doi:10.1016/j.jnlssr.2024.06.001
16. Thurzo A, Varga I. Revisiting the role of review articles in the age of AI-Agents: integrating AI-Reasoning and AI-Synthesis reshaping the future of scientific publishing. *Bratislavské LekáRske Listy/Bratislava Med J* [Internet]. 2025;126(4):381–393. doi:10.1007/s44411-025-00106-8
17. Levin C, Naimi E, Saban M. Evaluating GenAI systems to combat mental health issues in healthcare workers: an integrative literature review. *Int J Med Inform*. 2024;191:105566. doi:10.1016/j.ijmedinf.2024.105566
18. Miles G, Giles L, Kerr B, Norman B, Sibbring G. MSR201 optimising performance of generative artificial intelligence (GENAI) in systematic literature review (SLR) screening using PICOS criteria. *Value Health*. 2024;27(12):S478. doi:10.1016/j.jval.2024.10.2435

19. Neo NWS, Gunawan J, Levett-Jones T, Khoo ET, Chua WL, Liaw SY. Generative artificial intelligence in healthcare simulation-based education: a scoping review. *Clin Simul Nurs*. 2025;108:101819. doi:10.1016/j.ecns.2025.101819
20. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implement Sci*. 2024;19(1). doi:10.1186/s13012-024-01357-9
21. Sallam M, Khalil R, Sallam M. Benchmarking generative AI: a call for establishing a comprehensive framework and a generative AIQ test. *Mesopotam J Artif Intell Healthc*. 2024;69–75. doi:10.58496/mjaih/2024/010
22. Wang L, Bhanushali T, Huang Z, et al. Evaluating generative AI in mental health: a systematic review of capabilities and limitations. *JMIR Ment Health*. 2025;12:e70014. doi:10.2196/70014
23. Li Y, Datta S, Rastegar-Mojarad M, et al. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. *J Am Med Inform Assoc*. 2025;32:616–625. doi:10.1093/jamia/ocaf030
24. Tranfield D, Denyer D, Smart P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *Br J Manag*. 2003;14(3):207–222. doi:10.1111/1467-8551.00375
25. Denicol J. Why clear managerial recommendations matter in major projects research: searching for relevance in practice. *Int J Proj Manag*. 2022;40(2):98–100. doi:10.1016/j.ijproman.2022.01.004
26. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097
27. Yu F, Moehring A, Banerjee O, et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat Med*. 2024;30(3):837–849. doi:10.1038/s41591-024-02850-w
28. Gomez C, Smith B, Zayas A, Unberath M, Canares T. Explainable AI decision support improves accuracy during telehealth strep throat screening. *Commun Med*. 2024;4(1). doi:10.1038/s43856-024-00568-x
29. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Evaluating ChatGPT-4's accuracy in identifying final diagnoses within differential diagnoses compared with those of physicians: experimental study for diagnostic cases. *JMIR Form Res*. 2024;8:e59267. doi:10.2196/59267
30. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol*. 2023;42(1):3–15. doi:10.1007/s11604-023-01474-3
31. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE Step 2 CK and clinical case reports. *Sci Rep*. 2024;14(1). doi:10.1038/s41598-024-58760-x
32. Sadeghi Z, Alizadehsani R, Cifci MA, et al. A review of explainable artificial intelligence in healthcare. *Comput Electr Eng*. 2024;118:109370. doi:10.1016/j.compeleceng.2024.109370
33. Arvai N, Katonai G, Mesko B. Health care professionals' concerns about medical AI and psychological barriers and strategies for successful implementation: scoping review. *J Med Internet Res*. 2025;27:e66986. doi:10.2196/66986
34. Beheshti M, Toubal IE, Alaboud K, et al. Evaluating the reliability of ChatGPT for health-related questions: a systematic review. *Informatics*. 2025;12(1):9. doi:10.3390/informatics12010009
35. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci*. 2025;32(1). doi:10.1186/s12929-025-01131-z
36. Banerji CRS, Shah AB, Dabson B, et al. Clinicians must participate in the development of multimodal AI. *EClinicalMedicine*. 2025;84:103252. doi:10.1016/j.eclinm.2025.103252
37. Bongurala AR, Save D, Virmani A, Kashyap R. Transforming health care with artificial intelligence: redefining medical documentation. *Mayo Clin Proc Digit Health*. 2024;2(3):342–347. doi:10.1016/j.mcpdig.2024.05.006
38. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. *J Med Internet Res*. 2024;26e56764. doi:10.2196/56764
39. Khan FA. AI in clinical diagnostics: is overreliance eroding clinical expertise? *PLOS Digit Health*. 2025;4(8):e0000959. doi:10.1371/journal.pdig.0000959
40. Tikhomirov L, Semmler C, McCradden M, Searston R, Ghassemi M, Oakden-Rayne L. Medical artificial intelligence for clinicians: the lost cognitive perspective. *Lancet Digit Health*. 2024;6:e589–e594. doi:10.1016/S2589-7500(24)00095-5
41. Dratsch T, Chen X, Mehrizi MR, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*. 2023;307(4). doi:10.1148/radiol.222176
42. Goodell AJ, Chu SN, Rouholiman D, Chu LF. Large language model agents can use tools to perform clinical calculations. *NPJ Digit Med*. 2025;8(1). doi:10.1038/s41746-025-01475-8
43. Ranji SR. Large language models—misdiagnosing diagnostic excellence? *JAMA Network Open*. 2024;7(10):e2440901. doi:10.1001/jamanetworkopen.2024.40901
44. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*. 2023;13(17):2760. doi:10.3390/diagnostics13172760
45. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digit Med*. 2023;6(1):113. doi:10.1038/s41746-023-00858-z
46. Abhari S, Afshari Y, Fatehi F, et al. Exploring ChatGPT in clinical inquiry: a scoping review of characteristics, applications, challenges, and evaluation. *Ann Med Surg*. 2024;86(12):7094–7104. doi:10.1097/MS9.0000000000002716
47. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev*. 2022;56(4):3005–3054. doi:10.1007/s10462-022-10246-w
48. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning. *JAMA Network Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969
49. Arab RAE, Abu-Mahfouz MS, Abuadas FH, et al. Bridging the gap: from AI success in clinical trials to real-world healthcare implementation—a narrative review. *Healthcare*. 2025;13(7):701. doi:10.3390/healthcare13070701
50. McDermid JA, Jia Y, Porter Z, Habli I. Artificial intelligence explainability: the technical and ethical dimensions. *Phil Trans the R Soc Math Phys Eng Sci* [Internet]. 2021;379(2207). doi:10.1098/rsta.2020.0363