



International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	BEYOND ACCURACY: EXPLAINABILITY, WORKFLOW, AND GOVERNANCE IN THE CLINICAL IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE
----------------------	---

DOI	https://doi.org/10.31435/ijitss.3(51).2026.6007
------------	---

RECEIVED	12 June 2026
-----------------	--------------

ACCEPTED	15 September 2026
-----------------	-------------------

PUBLISHED	21 September 2026
------------------	-------------------

LICENSE



The article is licensed under a **Creative Commons Attribution 4.0 International License**.

© The author(s) 2026.

This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

BEYOND ACCURACY: EXPLAINABILITY, WORKFLOW, AND GOVERNANCE IN THE CLINICAL IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE

Jagoda Pałubska (Corresponding Author, Email: palubskajagoda0@gmail.com)

Independent Public Health Care Institution of the Ministry of the Interior and Administration in Kielce named after St. John Paul II, Kielce, Poland

ORCID ID: 0009-0000-3833-7977

Oliwer Muller

Kielce Hospital of St. Aleksandra Sp. z o.o., Kielce, Poland

ORCID ID: 0009-0005-6197-8461

Zuzanna Rafałowska

University Clinical Hospital in Wrocław Borowska 213, 50-556 Wrocław, Poland

ORCID ID: 0009-0003-6699-3121

Stanisław Rutz

Wrocław Medical University, wybrzeże Ludwika Pasteura 1, 50-367 Wrocław, Poland

ORCID ID: 009-0007-7918-9625

Anna Wojtala

Szpital Uniwersytecki nr 2 im. dr. Jana Biziela, ul. Ujejskiego 75, 85-168 Bydgoszcz, Poland

ORCID ID: 0009-0008-4082-0835

Julia Kret

Szpital Specjalistyczny im. F. Ceynowy, ul. dr. A. Jagalskiego 10, 84-200 Wejherowo, Poland

ORCID ID: 0009-0003-2780-8448

Aleksandra Wyrostkiewicz

Plac Hirszfelda 12, 53-413 Wrocław, Poland

ORCID ID: 0009-0005-3679-7570

Aleksander Sadkowski

Collegium Medicum w Bydgoszcz Uniwersytet Mikołaja Kopernika w Toruniu, Jagiellońska 15, 85-067 Bydgoszcz, Poland

ORCID ID: 0009-0008-3300-2447

Paulina Bigda

Wojewódzki Szpital Zespolony im. Ludwika Rydygiera ul. Św. Józefa 53-59, 87-100 Toruń, Poland

ORCID ID: 0009-0005-5022-6184

Jan Gdula

Collegium Medicum w Bydgoszczy Uniwersytetu Mikołaja Kopernika w Toruniu, Jagiellońska 15, 85-067 Bydgoszcz, Poland

ORCID ID: 0009-0007-9987-3758

Lilianna Zielińska

Plac Hirszfelda 12, 53-413 Wrocław, Poland

ORCID ID: 0009-0004-2471-841X

Michał Szpunar

Collegium Medicum w Bydgoszcz Uniwersytet Mikołaja Kopernika w Toruniu, Poland

ORCID ID: 0009-0000-5981-3696

Adrianna Zabiegałowska

Wojewódzki Szpital Zespolony im. Ludwika Rydygiera ul. Św. Józefa 53-59, 87-100 Toruń, Poland

ORCID ID: 0009-0005-3048-0097

Magda Buśko

Collegium Medicum w Bydgoszczy Uniwersytetu Mikołaja Kopernika w Toruniu, Poland

ORCID ID: 0009-0005-6451-8612

ABSTRACT

Artificial intelligence (AI) is increasingly discussed as a transformative force in medicine, yet its practical value depends less on algorithmic promise alone than on whether tools can be integrated safely, intelligibly, and sustainably into clinical care. This review examines how free full-text PubMed literature describes the real-world implementation of AI in healthcare and identifies the technical, organizational, ethical, and social conditions that shape adoption. A structured narrative review was conducted using a targeted PubMed search with a free full-text filter, supplemented by focused searches on explainability, governance, workflow, bias, and implementation in clinical settings. The final core synthesis drew on 21 peer-reviewed articles published between 2019 and 2025, including reviews, qualitative studies, survey studies, consensus guidance, and quality-improvement evaluations. Across the literature, implementation emerged as a sociotechnical challenge rather than a purely technical one. Recurrent themes included the gap between proof-of-concept performance and routine use, the need for practical explainability, risks related to bias and patient safety, workflow and workforce adaptation, and the importance of lifecycle governance. Real-world studies suggest that AI can reduce documentation burden, improve structured data capture, and support triage or decision support, but adoption remains uneven and often limited by poor integration, uncertain accountability, insufficient training, and weak post-deployment monitoring. The review concludes that clinical AI should be evaluated not only by accuracy but also by its fit with human work, organizational routines, fairness, and governance.

KEYWORDS

Artificial Intelligence, Healthcare Implementation, Explainability, Governance, Workflow, Patient Safety

CITATION

Pałubska, J., Muller, O., Rafałowska, Z., Rutz, S., Wojtala, A., Kret, J., Wyrostkiewicz, A., Sadkowski, A., Bigda, P., Gdula, J., Zielińska, L., Szpunar, M., Zabiegałowska, A., & Buśko, M. (2026). Beyond accuracy: Explainability, workflow, and governance in the clinical implementation of artificial intelligence. *International Journal of Innovative Technologies in Social Science*, 3(51). [https://doi.org/10.31435/ijitss.3\(51\).2026.6007](https://doi.org/10.31435/ijitss.3(51).2026.6007)

COPYRIGHT

© The author(s) 2026. This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

Introduction

Artificial intelligence (AI) has become one of the defining technological narratives in contemporary medicine. Machine learning, natural language processing, computer vision, and, more recently, generative models are now associated with almost every segment of healthcare, from diagnostic imaging and clinical decision support to documentation, scheduling, population health, and virtual triage. In broad strategic accounts, AI is frequently described as part of an “AI-enabled” or “AI-augmented” health system in which automation and decision support affect patient administration, monitoring, clinical intervention, and the organizational management of care (Reddy et al., 2019). At the same time, there is a persistent gap between what AI appears able to do in controlled studies and what health systems are actually able to deploy safely and sustainably in everyday practice.

That implementation gap is now one of the most important questions in the field. Shaw et al. (2019) argued early that the future of AI in healthcare would depend not only on technical design but also on issues such as meaningful decision support, explainability, privacy, consent, bias, security, scalability, and the changing nature of health work. Subsequent evidence has supported that claim. A systematic review of AI applications implemented in real-life clinical practice found that, despite an enormous volume of research in model development and validation, comparatively few AI applications had actually been evaluated at the clinical frontline (Yin et al., 2021). This imbalance matters because an algorithm that performs well in retrospective development data is not the same thing as a tool that clinicians, managers, and patients can rely on in real-world settings.

The problem, therefore, is not simply whether AI works, but under what conditions it becomes usable, trustworthy, governable, and sustainable. In medical settings, implementation is shaped by a dense network of technical and social variables: data quality, user interface design, staffing patterns, institutional routines, legal responsibility, training, public expectations, and the emotional and moral norms of care. If AI is inserted into a workflow without attention to those factors, even a technically strong tool may be ignored, overruled, misused, or quietly abandoned. Conversely, a system that is less spectacular in technical terms may produce real value if it fits clinical work, communicates uncertainty clearly, and is embedded in a credible governance structure.

This perspective is especially relevant for journals such as *The International Journal of Innovative Technologies in Social Science*, where the key question is not only what a technology can do, but how it reshapes practice, institutions, and human relationships. Clinical AI is a quintessential sociotechnical object. It depends on data infrastructures, regulation, vendor ecosystems, and organizational readiness, but it also alters how responsibility is distributed, how expertise is interpreted, and how workers experience autonomy and burden. The literature reviewed here repeatedly suggests that the success of AI in healthcare depends on its social embedding as much as on its computational architecture.

Several recurring tensions define the current landscape. The first is the tension between performance and explainability. In public discourse, explainability is often treated as the cure for concern about medical AI. Yet a multidisciplinary analysis by Amann et al. (2020) shows that explainability in healthcare is not merely a technical design preference. It intersects with legal liability, informed consent, clinical communication, and ethical principles such as autonomy and justice. Later research has suggested that what stakeholders often want is not maximal technical transparency, but forms of explanation that are meaningful in context and helpful for action. At the same time, the literature warns against superficial accounts of explainability that create the appearance of trustworthiness without addressing deeper issues of validation, bias, or relevance to local practice.

A second tension concerns bias and safety. AI is frequently promoted as a way of improving consistency and reducing human error, but it can also inherit, reproduce, or amplify inequities present in clinical data and institutional practice. Challen et al. (2019) frame bias not as a peripheral ethical issue but as a core patient-safety concern. More recent literature extends this point by showing that bias does not arise only from data and algorithms. It can also emerge in the way users interpret model outputs, in how interfaces encourage deference or skepticism, and in how organizational contexts reward speed over critical reflection (Kostick-Quenet & Gerke, 2022; Militello et al., 2025). Safety, in other words, is not guaranteed simply because a tool is computational.

A third tension is the relationship between AI and clinical work. Many clinicians and health systems are attracted to AI because of workforce shortages, administrative overload, and rising demand. Documentation tools, triage systems, and decision support systems are often justified as ways of reducing burden and improving efficiency. Emerging evidence suggests that these benefits can be real. Quality-improvement and

review studies on AI-supported documentation report reduced perceived documentation burden, improved efficiency, and more time available for patient care in some settings (Lee et al., 2024; Albrecht et al., 2025). But the same literature also points to legal, technical, and organizational complications, including integration with electronic records, quality control, after-hours use, and uncertainty about liability. Clinical work is not merely a set of tasks to be optimized; it is also a domain of judgment, collaboration, and professional identity.

A fourth tension concerns governance. The pace of AI development has often outstripped the pace of institution-building around it. Reddy et al. (2020) proposed a governance model for AI in healthcare in response to exactly this problem, emphasizing that ethical and regulatory discussions need to be translated into concrete governance processes. The 2025 FUTURE-AI consensus guideline develops this further by organizing trustworthy healthcare AI around six principles—fairness, universality, traceability, usability, robustness, and explainability—and by proposing best practices across the full lifecycle of design, validation, deployment, and monitoring (Lekadir et al., 2025). These frameworks suggest that implementation should not begin at the moment a tool is procured or switched on. It begins much earlier in data curation, stakeholder engagement, and validation design, and it continues after deployment through monitoring, feedback, and adaptation.

The literature also makes clear that implementation is perceived differently by different stakeholders. Clinicians, nurses, managers, patients, and members of the public may all recognize potential benefits while worrying about different risks. A multi-stakeholder systematic review by Vo et al. (2023) found broad positivity toward healthcare AI alongside recurring concerns about privacy, transparency, bias, regulation, professional autonomy, and the human-AI relationship. Studies of specific professional groups tell a similar story. General practitioners in the United Kingdom stressed the limits of future technology for empathy, value-based care, and nuanced clinical reasoning (Blease et al., 2019). Nurses and other frontline professionals have emphasized both the promise of AI and the need for training, role clarity, and protection of compassionate care (Rony et al., 2024). Qualitative work from critical care and primary care likewise shows that adoption can stall when AI is perceived as optional, poorly explained, insufficiently useful, or inadequately supported by organizational policy (Mlodzinski et al., 2023; Larsson et al., 2025).

Against this background, the present article offers a structured narrative review of free full-text PubMed literature on the clinical implementation of AI. The review deliberately moves beyond broad claims that AI is “promising” and focuses instead on the conditions under which it becomes clinically meaningful, ethically defensible, and organizationally sustainable. Three research questions guide the analysis:

RQ1. How is the implementation of AI in clinical healthcare described in recent free full-text PubMed literature?

RQ2. Which technical, organizational, ethical, and human factors most consistently appear as barriers or enablers of implementation?

RQ3. What practical implications follow for responsible deployment in settings such as decision support, triage, documentation, and primary care?

By centering implementation rather than performance alone, the review aims to contribute to both medical and social-science discussions of AI. In medicine, it synthesizes what the literature says about how tools become usable in routine care. In social science, it highlights how AI adoption depends on governance, labor, accountability, and institutional legitimacy. The central argument developed throughout the article is that clinical AI succeeds not when it is merely accurate, but when it is explainable enough for its context, safe enough for its stakes, compatible enough with human work, and governed well enough to remain trustworthy over time.

Methodology

This article was designed as a structured narrative review. The decision not to present the paper as a systematic review or meta-analysis was deliberate. The literature relevant to clinical AI implementation is methodologically heterogeneous and includes conceptual essays, qualitative studies, survey-based studies, quality-improvement evaluations, scoping reviews, systematic reviews, and consensus guidelines. Because the present article seeks to synthesize how implementation is understood across these different forms of evidence, a structured narrative approach was considered more appropriate than statistical aggregation or formal effect-size comparison.

The review focused specifically on free full-text literature indexed in PubMed. This boundary was selected for two reasons. First, it ensured that all included publications were directly accessible in full text, which is particularly useful for a review intended to map themes across methods rather than to rely on abstracts

alone. Second, PubMed remains one of the most widely used entry points for biomedical and health-services literature and therefore provides a practical and transparent basis for an applied review of clinically relevant AI research. A targeted search strategy was conducted in March 2026 using combinations of terms related to artificial intelligence, machine learning, clinical implementation, healthcare workflow, triage, documentation, explainability, governance, bias, and patient safety. The PubMed “free full text” filter was applied throughout the search process.

The core search logic combined AI terms with implementation-oriented terms. Examples of search strings included: (“artificial intelligence” OR “machine learning”) AND (healthcare OR medicine OR clinical) AND (implementation OR deployment OR workflow); (“artificial intelligence” OR “machine learning”) AND (explainability OR governance OR bias OR “patient safety”); and focused follow-up searches such as (“artificial intelligence” AND documentation), (“artificial intelligence” AND triage), (“artificial intelligence” AND “primary care”), (“artificial intelligence” AND nursing), and (“artificial intelligence” AND qualitative study). The search process was iterative rather than purely linear. Initial searches were used to identify key domains of implementation concern, and subsequent focused searches were conducted to ensure that each domain was represented by free full-text studies with clear relevance to clinical practice.

Studies were considered eligible when they met four broad criteria. First, they had to be published in English between 2019 and 2025. This time frame was selected because it captures the period in which implementation-oriented discussion of clinical AI became much more visible, especially in relation to real-world deployment, explainability, governance, and human factors. Second, the publication had to be indexed in PubMed and accessible in full text through the free full-text filter. Third, the study had to address AI in a genuinely clinical or health-services context rather than in a purely technical or speculative context. Fourth, the study had to contribute directly to the review question by discussing one or more implementation-relevant issues, such as integration into workflow, frontline use, ethical governance, human oversight, stakeholder expectations, explainability, safety, or barriers to adoption.

Papers were excluded when they focused only on algorithm development or retrospective performance testing without discussion of practical clinical deployment. Studies that dealt with highly technical aspects of model architecture but did not address healthcare implementation, human use, or organizational adoption were also excluded. Editorials and brief commentaries were not automatically excluded, but they were retained only when they offered substantive conceptual value and were clearly situated in the implementation debate. To maintain the practical focus of the review, priority was given to studies that either examined actual real-world clinical use, synthesized implementation evidence, or reported stakeholder perspectives directly relevant to adoption.

After targeted searching and full-text screening, a final core corpus of 21 publications was selected for detailed analysis. These publications included systematic or scoping reviews, qualitative interview studies, descriptive survey studies, quality-improvement work, governance or consensus papers, and conceptual analyses. The included literature covered a range of clinical contexts, including primary care, critical care, nursing, clinical documentation, decision support, medical safety, and broader cross-cutting discussions of healthcare AI governance. While the corpus was not exhaustive of all clinical AI research, it was sufficient to support a robust thematic synthesis focused on implementation rather than on technical benchmarking.

For each included publication, the following information was extracted manually: publication year, study type, clinical or organizational setting, stakeholder group, AI use case, main implementation issue, and principal practical implication. These details were then used to compare studies across domains rather than discipline by discipline. During the reading process, recurring themes were coded inductively and then grouped into higher-order implementation categories. Five broad domains proved especially persistent across the literature: the gap between proof-of-concept and real-world deployment; explainability and meaningful use; bias, safety, and the need for calibrated human oversight; workflow, labor, and professional adaptation; and governance across the lifecycle of AI development and deployment.

No formal quantitative quality appraisal tool was applied. This was not because quality is unimportant, but because the material reviewed was too diverse in design and purpose for a single scoring approach to be meaningful. A checklist appropriate for randomized intervention studies, for example, would not be informative when applied to qualitative studies, consensus guidelines, or conceptual ethics papers. Instead, the review followed a principle of analytical sufficiency: studies were included when their methods and arguments were clear enough to support thematic comparison, and stronger weight was given in the synthesis to publications with transparent methods, explicit data sources, or broad integrative scope.

This review design has limitations. Because the search was restricted to PubMed and to free full-text publications, relevant literature available only through subscription databases or other indexing platforms may not have been captured. The use of a structured narrative approach also means that the review does not produce pooled estimates or formal causal claims. In addition, some included studies reflect perceptions, expectations, or early implementation experience rather than mature long-term deployment. These limits were accepted because the purpose of the article is interpretive and mapping-oriented: to understand what recent accessible literature suggests about the conditions under which AI becomes workable in clinical care. Within those limits, the approach provides a transparent and practically useful synthesis for a journal positioned at the intersection of technology and society.

Results

The final synthesis included 21 free full-text PubMed publications published between 2019 and 2025. The evidence base was heterogeneous by design, reflecting the mixed state of AI implementation research. It contained conceptual implementation papers, multidisciplinary ethics analyses, systematic and scoping reviews, qualitative interview studies, descriptive surveys, a quality-improvement evaluation, and a large international consensus guideline. In methodological terms, this diversity is itself informative: implementation is not treated in the literature as a narrow technical matter but as a problem that cuts across clinical evidence, organizational behavior, ethics, policy, and human factors.

The literature was also unevenly distributed across clinical settings. Diagnostic and decision-support discussions remained prominent, but a notable share of the more concrete implementation evidence came from domains where workflow burden is already highly visible, such as clinical documentation and triage. Several studies dealt with implementation at the systems level rather than within a single specialty, especially those focused on governance, stakeholder expectations, or the ethics of deployment. Only a limited portion of the reviewed literature reported patient outcomes directly. More often, the literature concentrated on enablers and barriers to adoption, professional perceptions, safety concerns, and the institutional conditions under which AI might or might not become sustainable in routine practice.

Table 1 summarizes representative studies included in the review. As shown in Table 1, the evidence base ranges from early implementation theory (Shaw et al., 2019) and foundational ethical analyses of explainability (Amann et al., 2020) to newer empirical work on ambient documentation tools (Albrecht et al., 2025), triage integration in primary care (Larsson et al., 2025), and lifecycle governance guidance (Lekadir et al., 2025). Table 2 synthesizes the five major implementation domains that organize the following analysis. Across these sources, one consistent pattern emerges: the central challenge is not simply building AI tools, but embedding them in clinical environments in ways that are useful, safe, and institutionally legitimate.

Table 1. Representative free full-text PubMed studies included in the review

Study	Design and setting	Main implementation insight
Shaw et al. (2019)	Conceptual implementation analysis in healthcare	AI adoption depends on decision support value, privacy, bias, security, scalability, and organizational fit.
Amann et al. (2020)	Multidisciplinary analysis of healthcare AI	Explainability has technical, legal, ethical, and patient-facing functions.
Yin et al. (2021)	Systematic review of real-life clinical AI studies	Real-world implementation studies remain relatively few and often report limited patient or economic outcomes.
van der Veer et al. (2021)	Citizens' juries on AI decision-making	In healthcare contexts, people may accept lower explainability if expected accuracy is substantially higher.
Mlodzinski et al. (2023)	Web-based survey in critical care	Barriers include trust, workflow fit, technical integration, and organizational readiness.
Yoo et al. (2023)	Qualitative study of healthcare professionals	Implementation requires training, realistic expectations, and role clarity.
Lee et al. (2024)	Scoping review of AI-assisted documentation	Efficiency gains are promising, but legal, integration, and quality concerns persist.

Study	Design and setting	Main implementation insight
Albrecht et al. (2025)	Quality-improvement survey of ambient AI documentation	Clinicians report less documentation burden and better workflow after implementation.
Larsson et al. (2025)	Qualitative primary care study of AI triage	Routine use remained limited because the tool was not fully normalized in local workflows.
Lekadir et al. (2025)	International consensus guideline	Deployable AI requires fairness, universality, traceability, usability, robustness, and explainability across the full lifecycle.

Note. The table presents representative studies; the full synthesis drew on 21 publications.

Table 2. Major implementation domains identified across the reviewed literature

Domain	Recurrent issues in the literature	Practical implications for health systems
Proof-of-concept vs. real-world use	Many AI tools are technically promising, but comparatively few are evaluated in mature frontline practice.	Prioritize local evaluation, realistic use cases, and post-pilot sustainability planning.
Explainability and meaningful use	Stakeholders need explanations that support action, challenge, communication, and review in context.	Match explanatory design to user role, task, and clinical stakes.
Bias, safety, and oversight	Bias can arise from data, interfaces, workflow pressure, and user interpretation.	Use human factors methods, clear uncertainty communication, and calibrated oversight.
Workflow, labor, and training	AI redistributes work, including verification, exception handling, and coordination.	Assess hidden labor, support training, and protect professional judgment.
Governance across the lifecycle	Deployment requires validation, traceability, monitoring, update control, and accountability.	Treat governance as continuous stewardship rather than one-time approval.

Note. These domains structure the synthesis presented in the following subsections.

The implementation gap: AI is visible in discourse, but selective in practice

One of the clearest findings across the reviewed literature is the persistence of a gap between AI's visibility as a medical innovation and its actual routinization in clinical practice. Reddy et al. (2019) describe a broad landscape in which AI may influence patient administration, clinical decision support, monitoring, and intervention, thereby creating the impression that healthcare is entering an AI-enabled era. Yet when one looks for evidence of mature frontline deployment, the picture becomes much narrower. Yin et al. (2021), in their systematic review of AI applications in real-life clinical practice, identified only 51 implementation studies despite the far larger volume of research on model development. Their review is important because it makes visible the difference between a field rich in technical experimentation and a field rich in real-world implementation evidence. Many applications had reached the stage of prototype, evaluation, or pilot use, but relatively few had generated strong evidence on patient outcomes, provider outcomes, or economic effects in routine practice.

The same pattern appears in later literature. Shaw et al. (2019) framed implementation as a major challenge precisely because healthcare AI is usually introduced into already complex systems characterized by legacy technologies, professional norms, regulatory constraints, and interdependent workflows. Their use of the NASSS perspective highlighted that adoption problems arise not only from the technical tool itself but also from the condition being managed, the organization, the external policy environment, and the long-term adaptation required for scale-up and sustainability. The key point is that implementation failure is rarely attributable to a single defect. It is more often the product of misalignment between a technology and the context into which it is introduced.

Real-world studies published after 2020 suggest that adoption is most likely where the use case is bounded, the operational pain point is already widely recognized, and the tool can be integrated into a task that clinicians readily understand. Documentation is an instructive example. Lee et al. (2024), in a scoping review of AI in clinical documentation, found recurring evidence that AI tools can improve documentation efficiency and accuracy, especially by automating repetitive aspects of note generation and structured data capture. Albrecht et al. (2025) add a more implementation-proximate perspective through a quality-improvement survey of ambient AI documentation, reporting improved perceptions of documentation workflow, reduced work burden, and less burnout-related strain in the evaluated setting. These studies suggest that adoption can gain traction when AI addresses a visible administrative burden and the benefit is readily legible to users.

However, even in apparently favorable domains, implementation remains conditional. Lee et al. (2024) also note recurring concerns about legal liability, integration with existing electronic health record systems, data privacy, and the risk of documentation errors or overreliance. In other words, a workflow benefit does not eliminate the need for validation, human review, and institutional rules. The same technology that saves time may introduce new responsibilities, including verification, error correction, and oversight of machine-generated text. Implementation therefore changes work rather than simply removing it.

Triage offers a contrasting case. In theory, triage is an appealing AI use case because it involves prioritization, structured symptom capture, and the possibility of routing patients more efficiently. Yet the qualitative study by Larsson et al. (2025) shows how difficult routine integration can be in primary care. Healthcare professionals described the AI-based triage system as insufficiently embedded in established routines, insufficiently supported by organizational policy, and often less useful than expected in relation to free-text patient information. The tool was not rejected in principle, but it had not become normalized in practice. This is a critical distinction. Adoption is not equivalent to procurement, training, or initial rollout. A tool is truly implemented only when it becomes part of normal work, and the literature suggests that many healthcare AI applications have not yet reached that stage.

The critical care literature points in the same direction. Mlodzinski et al. (2023), in a web-based survey study, examined barriers to implementation of machine learning and AI-based tools in critical care. Their findings indicate that implementation is slowed by a complex mix of concerns around trust, workflow disruption, technical integration, and the organizational capacity to use such tools effectively. Critical care is data-rich and often imagined as an ideal site for AI, but it is also high-pressure, high-stakes, and intolerant of tools that generate ambiguity or additional cognitive burden. The implication is that data abundance alone does not guarantee implementability.

Overall, the reviewed studies suggest that the implementation gap persists because health systems must do more than acquire AI. They must decide where it is useful, who is accountable, how the output is interpreted, how it enters existing documentation and decision pathways, how exceptions are handled, and how performance is monitored over time. The more an application touches consequential clinical judgment or institutional coordination, the more demanding implementation becomes. This is why the literature repeatedly distinguishes between proof-of-concept success and deployable value. In practice, AI becomes sustainable only when local workflow, organizational incentives, and governance structures are aligned with the technology's intended role.

Explainability as a contextual requirement rather than a universal solution

Explainability is one of the most frequently invoked solutions to the social and ethical problems of healthcare AI, but the reviewed literature presents a more nuanced picture. The central question is not whether explanation matters; it clearly does. The more important question is what kind of explanation is needed, for whom, and for what purpose. Amann et al. (2020) offer one of the clearest multidisciplinary accounts of this issue. They argue that explainability in healthcare spans multiple domains, including technical understanding, legal responsibility, clinical communication, ethical legitimacy, and patient autonomy. This is significant because it resists the narrow assumption that explainability is only a matter of inspecting model internals. In clinical settings, explanation is also about whether a tool can be used responsibly, audited appropriately, and discussed meaningfully in care.

Later literature supports this contextual understanding. The systematic literature review by Alkhanbouli et al. (2025) shows that explainable AI has become especially prominent in disease prediction research, with methods such as SHAP and LIME frequently used to provide post hoc explanations of model behavior. Yet the same review also points to unresolved problems, including limited dataset diversity, model complexity, and challenges of integrating explainability into practice rather than leaving it as an abstract modeling exercise.

This is a crucial finding for implementation-oriented research. A tool can be technically “explainable” in the machine-learning sense and still fail to provide the kind of actionable clarity that clinicians or patients need at the point of use.

The literature also challenges the idea that explainability is always the highest priority. van der Veer et al. (2021), through two citizens’ juries, found that in healthcare scenarios participants were willing to prioritize accuracy over explainability more than they were in non-healthcare scenarios. This does not mean that explainability was unimportant. Rather, it suggests that stakeholders may accept trade-offs when the perceived clinical benefit is strong, especially if other forms of oversight are in place. The implication is that explainability is relational and situational. In some settings, a simpler but slightly less accurate model may be preferable because it is easier to interpret, contest, or communicate. In other settings, users may accept a more complex model provided that it is well validated, its limitations are known, and its outputs are embedded within accountable clinical review.

This distinction helps explain why the implementation literature often speaks less about transparency in the abstract and more about meaningful use. Shaw et al. (2019) emphasize meaningful decision support rather than explanation for its own sake. The issue is whether an AI tool fits the informational needs of the clinical moment. A clinician using decision support in time-pressured care does not necessarily need a complete explanation of the model architecture, but they do need to know what the output means, how uncertain it is, where it is likely to fail, and what type of action it is supposed to guide. Similarly, an institution concerned with governance requires traceability, auditing capacity, and enough explanatory structure to review errors or adverse events after deployment.

The professional-perspective literature reinforces this pragmatic reading. In their qualitative study of healthcare professionals, Yoo et al. (2023) found that expectations about medical AI were tied not just to accuracy but to strategies for implementation, which included the need for systems that professionals could understand and work with. Blease et al. (2019) likewise showed that general practitioners were attentive to the limits of technology in the domain of nuanced reasoning and person-centered care. Their concerns were not reducible to a demand for more technical information. Rather, they reflected doubt about whether AI outputs could be integrated into the kind of contextual judgment that primary care requires. Explanation, in such settings, is meaningful only if it supports clinical reasoning rather than displacing it.

The recent FUTURE-AI consensus guideline is useful because it places explainability inside a broader implementation framework rather than treating it as an isolated virtue. Lekadir et al. (2025) identify explainability as one of six guiding principles, alongside fairness, universality, traceability, usability, and robustness. This ordering matters. It suggests that explainability is necessary, but insufficient on its own. A clinically deployable tool must also be usable, resilient, and traceable across the lifecycle of development and deployment. In practical terms, the literature suggests that health systems should not ask a binary question—whether a tool is explainable or not—but a contextual question: does the form of explanation provided support safe reliance, challenge, communication, and oversight in this particular setting?

Across the reviewed studies, the answer appears to depend on setting, stakeholder, and stakes. For documentation tools, the central explanatory need may concern verification, editability, and the provenance of generated text. For triage, it may concern why a patient was routed or prioritized in a certain way. For clinical prediction, it may concern the features or patterns contributing to a risk estimate and the conditions under which the model should not be used. This suggests that “explainability” is better understood not as a single property but as a family of implementation requirements linked to action, accountability, and trust. The literature therefore supports a shift from generic calls for transparent AI toward context-sensitive explanatory design.

Bias, patient safety, and the need for calibrated human oversight

A third major theme in the reviewed literature is that safety concerns in clinical AI are inseparable from bias, user behavior, and the design of human oversight. Challen et al. (2019) make an influential argument that bias should be understood as a clinical safety problem rather than as a merely abstract ethical issue. In healthcare, biased outputs can affect diagnosis, prioritization, resource allocation, and risk scoring in ways that translate directly into harm. This framing is especially important because it shifts bias from the margins of implementation discourse to the center of patient safety.

Systematic evidence on AI and safety also points to a complicated picture. Choudhury and Asan (2020) reviewed the role of AI in patient safety outcomes and found that AI has the potential to improve healthcare quality and safety, but that the evidence base remained heterogeneous and difficult to compare because of

inconsistent reporting, variable benchmarks, and differing study designs. Their review suggests that safety claims about AI should be made cautiously. Potential benefit exists, but benefit depends on how a system is evaluated, introduced, and monitored. Without standardization and robust implementation design, the language of “safety improvement” can easily outpace evidence.

More recent work extends this insight by emphasizing that implementation risk does not lie only in the model. Kostick-Quenet and Gerke (2022) draw attention to “imperfect users,” arguing that AI systems are always interpreted by humans whose own biases, heuristics, and working conditions shape how outputs are used. This argument is highly relevant for clinical practice. Even a technically sound model may be used poorly if its interface invites overtrust, if uncertainty is poorly communicated, or if time pressure encourages clinicians to accept recommendations uncritically. Conversely, a well-designed interface may encourage reflective use, appropriate skepticism, and safer calibration between human and machine judgment.

Militello et al. (2025) bring a human factors perspective to this issue by showing that interface and communication design can either exacerbate or mitigate bias in AI-based clinical decision support. Their analysis suggests that risk communication, display design, and user-centered evaluation are not secondary matters of convenience. They are part of the safety architecture of the system. This observation has major practical significance. Many implementation failures are explained after the fact in terms of “user resistance” or “lack of trust,” but the literature reviewed here indicates that some of what appears as user failure is actually design failure. If a system makes uncertainty invisible, obscures the basis of a recommendation, or fits poorly with cognitive workflow, it may encourage unsafe patterns of use.

The multi-stakeholder literature confirms that bias and safety remain live concerns for diverse groups. Vo et al. (2023) found that while many stakeholders were broadly optimistic about healthcare AI, recurring concerns included privacy, bias, inequity, regulation, professional autonomy, and transparency. What is notable is that these concerns were not isolated to one group. They appeared across patients, clinicians, and other stakeholders, suggesting that the legitimacy of AI in healthcare depends in part on whether systems can be shown to be fair and governable in practice. Ratti et al. (2025) similarly argue that ethical and social discussions need to move beyond generic principles toward more situated consideration of how AI actually operates in healthcare settings. This reinforces the idea that safety is contextual. It is not enough to say that a model has been validated somewhere; one must ask how it behaves in a specific workflow, with a specific user population, under specific institutional constraints.

The literature also implies that human oversight should not be treated as a simple safeguard added after technical development. Oversight needs to be calibrated. Too little human involvement can create blind reliance on algorithmic output. Too much poorly structured oversight can create automation bias in another form, where clinicians appear to be “in the loop” but lack the information, time, or authority needed to challenge the system meaningfully. The implementation question is therefore not just whether there is a human in the loop, but what that human is actually able to do. Can the clinician understand the output? Can they access the relevant context? Do they know the system’s failure modes? Are there procedures for escalation, override, and auditing? The reviewed literature repeatedly suggests that these questions matter more than nominal declarations of supervision.

In this sense, bias and safety are intertwined with organizational design. A hospital may deploy the same algorithm in two different ways and achieve very different safety profiles depending on interface, staffing, training, and monitoring. This is why the literature reviewed here supports the concept of calibrated oversight: oversight that is proportionate to the clinical stakes, informed by the tool’s limitations, and embedded within a system of post-deployment review. Without such calibration, AI risks being either underused because it is distrusted or overused because it is granted an authority that its real-world validation does not justify.

Workflow, labor, professional identity, and the lived experience of adoption

A fourth theme running through the literature is that AI implementation changes work. It may reduce some tasks, redistribute others, and create entirely new forms of oversight, coordination, and professional negotiation. This point is essential because AI is often promoted through an efficiency narrative that implies straightforward labor savings. The reviewed literature suggests a more complex reality. In clinical settings, AI rarely eliminates work; instead, it shifts the composition, timing, and accountability structure of work.

Documentation provides one of the clearest examples of this dynamic. Lee et al. (2024) report that AI-assisted documentation tools can improve efficiency and accuracy across clinical settings by reducing repetitive charting demands, supporting note generation, and improving consistency in structured data capture. Albrecht et al. (2025) go further by examining clinician perceptions after implementation of ambient AI

documentation. Their quality-improvement survey found reduced work burden and improved satisfaction in important domains related to documentation. These findings help explain why documentation has become one of the most visible entry points for AI in actual clinical environments: the administrative burden is already widely recognized, the pain point is immediate, and the potential value proposition is understandable to users and institutions alike.

Yet the same studies make clear that AI documentation is not a frictionless solution. Machine-generated notes require review. Errors may be subtle rather than obvious. Voice capture and ambient listening raise questions about data governance, recording practices, and patient comfort. Integration with electronic health record platforms can be technically uneven, and liability for inaccurate or incomplete notes remains a significant concern (Lee et al., 2024). Thus, one kind of burden may be reduced while another kind is created. Clinicians may spend less time typing but more time verifying, editing, and deciding how much trust to place in generated text. What changes, then, is not simply the quantity of work but the form of cognitive and legal responsibility attached to it.

The qualitative professional literature broadens this insight. Blease et al. (2019) found that general practitioners viewed AI through a mixed lens of promise and limitation. They recognized potential benefits in efficiency and standardization, yet they doubted that AI could replace the interpretive, empathetic, and value-sensitive aspects of primary care. Their concern was not merely sentimental. It reflected a professional understanding that much of medicine involves contextual judgment, negotiation of uncertainty, and relational care—domains in which automation may support work without supplanting the human clinician. In implementation terms, this means that AI adoption may be most acceptable when it augments practice while leaving professional reasoning and patient-facing interaction intact.

Nursing perspectives point in a similar direction. Rony et al. (2024) identified themes of readiness, expected benefits, workflow change, training needs, ethical concerns, and the importance of maintaining balance between technological and human elements of care. The study is valuable because nursing is sometimes underrepresented in AI implementation discussions despite the fact that nurses are central to monitoring, communication, documentation, and bedside coordination. The findings suggest that implementation is likely to be more sustainable when it acknowledges how AI intersects with caring labor, not only with data processing. If AI is perceived as imposing additional monitoring, undermining the patient-professional relationship, or redistributing responsibility without support, adoption is likely to be contested even if the technology is technically impressive.

The critical care literature also illustrates how implementation is shaped by workload, trust, and environmental pressure. Mlodzinski et al. (2023) describe barriers that reflect the complexity of introducing AI into settings where clinicians already operate under high cognitive load. In such environments, AI tools cannot be evaluated only in terms of theoretical usefulness. They must also be judged by how they affect alert burden, interpretive effort, and team coordination. A system that generates extra steps or ambiguous recommendations may be rejected not because staff are resistant to innovation, but because it does not fit the tempo and decision ecology of the unit.

The study by Yoo et al. (2023) provides a useful bridge between frontline expectations and implementation strategy. Healthcare professionals in their study did not discuss AI solely in terms of technical capability. They also emphasized conditions for successful implementation, including appropriate training, realistic expectations, institutional support, and role clarity. This is an important corrective to the common assumption that AI adoption depends mainly on convincing individual clinicians of technological value. The literature suggests instead that implementation is collective. It depends on whether organizations create conditions in which professionals can use the technology competently, challenge it when needed, and integrate it without undermining teamwork or professional identity.

Larsson et al. (2025) demonstrate what happens when these conditions are only partially met. In the Swedish primary care setting they studied, the AI-based triage tool had not become normalized. Professionals described it as optional, organizationally under-supported, and not fully aligned with existing routines. Some relied more on patients' original free-text submissions than on the AI summaries. This finding is especially important because it shows how implementation can remain superficial. A tool may be available, known, and even formally introduced, yet still fail to become part of routine care if users do not find it sufficiently helpful or if policy and workflow do not stabilize its role.

Taken together, these studies suggest that AI implementation should be understood as labor reorganization rather than mere automation. Adoption affects documentation practices, interpretive work, professional judgment, communication with patients, and the invisible coordination tasks that hold clinical

workflows together. For this reason, effective implementation requires more than training users on a new interface. It requires understanding what kinds of work are being displaced, what kinds of work are being added, and how those changes interact with professional values. The social-science significance of this point is substantial: AI in medicine is not just about better tools, but about changing the terms under which clinical labor is performed and evaluated.

Governance, accountability, and the lifecycle view of implementation

If workflow explains how AI is encountered in practice, governance explains how it is authorized, bounded, and monitored over time. The reviewed literature consistently shows that governance cannot be treated as a secondary matter to be added after a technically successful tool has been built. Instead, governance shapes implementation from the start, because decisions about data sources, validation populations, intended use, monitoring processes, and responsibility structures all determine whether deployment is clinically legitimate and sustainable.

Reddy et al. (2020) provide one of the clearest early attempts to move from abstract ethics to an applied governance model for healthcare AI. Their work is important because it recognizes that implementation requires more than broad endorsement of principles such as fairness and transparency. Institutions need practical processes through which AI is reviewed, approved, monitored, and revised. Without such processes, governance remains aspirational rather than operational. In real settings, questions quickly arise about who is allowed to procure or approve an AI tool, what evidence is sufficient for deployment, how local validation is handled, how model updates are governed, and who is accountable if the system produces harm.

Shaw et al. (2019) similarly argue that implementation is shaped by factors far beyond the algorithm itself, including external policy, regulation, market forces, infrastructure, and the changing role of both public and private actors. One of their key contributions is to show that AI implementation sits within a broader ecosystem, not just within a clinical department. This means governance must also be ecological. It must address vendor relations, interoperability, data stewardship, cybersecurity, consent, and the institutional capacity to monitor systems once they are live. A hospital may purchase an AI-enabled tool, but if it cannot audit performance drift, evaluate local bias, or govern updates, it has not truly governed the technology.

The FUTURE-AI guideline by Lekadir et al. (2025) offers the most comprehensive lifecycle framework among the reviewed sources. By organizing trustworthy and deployable healthcare AI around six principles—fairness, universality, traceability, usability, robustness, and explainability—the framework signals that governance is not reducible to ethics review or regulatory compliance. It extends across design, development, validation, deployment, and continuous monitoring. The significance of this lifecycle view is difficult to overstate. Many implementation problems occur not because no one considered ethics, but because systems were not designed for ongoing governance after deployment. Models may drift, patient populations may change, workflows may evolve, and users may interact with systems differently from what designers anticipated. Governance that ends at launch is therefore inadequate.

The ethical literature reviewed here reinforces this point. Ratti et al. (2025), in a two-pronged scoping review, argue that discussions of AI in healthcare often remain trapped at the level of generic principles and need to become more sensitive to the situated realities of use. This position aligns with the rest of the implementation literature. A model may satisfy abstract ethical criteria while still creating practical injustice if it is deployed in a setting with poor digital infrastructure, unclear escalation procedures, or hidden barriers to patient access. Situated ethics therefore complements governance by insisting that institutions ask how principles operate in context rather than merely whether they are named in policy documents.

Stakeholder-preference research suggests that governance also matters for public legitimacy. Vo et al. (2023) found widespread concern about privacy, regulation, accountability, and professional autonomy across stakeholder groups. These concerns imply that adoption is not just an internal technical matter for hospitals or vendors. It is also an issue of social license. Health systems need to show that AI is not a hidden infrastructure operating without scrutiny, but a governed set of tools whose evidence base, limits, and oversight are taken seriously. This becomes even more important when AI supports high-stakes tasks such as diagnosis, triage, or risk stratification.

The reviewed literature also suggests that governance and usability are inseparable. A system that cannot be audited easily, challenged meaningfully, or understood well enough for appropriate use poses both governance and workflow problems. This is why FUTURE-AI's pairing of traceability, usability, and explainability is so useful. It acknowledges that good governance is not only a matter of formal committees or

policies; it also concerns whether the tool is structured in ways that make monitoring and responsible use possible at the point of care.

Across the literature, one recurring problem is the absence of strong post-deployment monitoring. Many AI studies focus on development or pre-implementation evaluation, whereas fewer examine what happens after routine use begins. Yet implementation is an ongoing process, not a one-time event. A system that appears successful in early deployment may later show different performance across sites, user groups, or patient populations. Documentation tools may drift in quality as templates, specialties, or languages change. Triage systems may perform differently when case mix changes seasonally or when patient inputs become less structured than expected. Governance therefore requires feedback loops, local performance review, and mechanisms for modification or withdrawal.

This lifecycle perspective leads to a broader conclusion: implementation should be judged by governability as much as by innovation. A technically advanced system that cannot be monitored, explained in context, integrated into policy, or updated responsibly may be less valuable than a more modest system that fits existing governance capacity. The literature reviewed here strongly supports that conclusion. In healthcare, innovation without governance is not neutral experimentation. It is a potential source of organizational fragility and patient risk.

Discussion

The literature reviewed in this article supports a clear and, in some ways, sobering conclusion: the central challenge of clinical AI is not proof of technical possibility but the achievement of sociotechnical fit. Across free full-text PubMed studies published between 2019 and 2025, implementation repeatedly appears as a problem of alignment among models, users, institutions, and norms of care. This finding matters because AI in medicine is still frequently discussed as if adoption were the natural next step after sufficient accuracy has been demonstrated. The reviewed literature suggests otherwise. A model may perform impressively and still fail in practice if it is hard to interpret in context, difficult to integrate into workflow, weakly governed, or misaligned with professional judgment.

One of the most striking patterns in the literature is that adoption appears strongest in bounded use cases with highly visible pain points, especially documentation. Documentation tools can often be evaluated against relatively concrete outcomes such as time spent charting, perceived burden, note completeness, and user satisfaction. They also address a task that clinicians already experience as onerous and administrative. This may explain why documentation is emerging as one of the most plausible near-term sites of AI normalization. By contrast, implementation appears more difficult in domains such as primary care triage, where workflows are more relational, information is less structured, and the consequences of misclassification are harder to contain. The contrast suggests that the implementability of AI is partly shaped by the structure of the task itself. The more bounded and standardized the task, the more feasible early adoption may be. The more the task depends on contextual interpretation, negotiated judgment, or complex organizational coordination, the more demanding implementation becomes.

This pattern has broader implications for how health systems should evaluate AI opportunities. It suggests that “high impact” should not be defined only in terms of theoretical clinical importance. It should also include implementability: the degree to which a use case allows meaningful validation, controllable integration, understandable outputs, and clear lines of accountability. In that sense, a modest AI application that reduces documentation burden safely may be more valuable than a more spectacular diagnostic system that cannot be governed or normalized in routine care.

A second major implication concerns explainability. The literature does not support either of the two most common simplistic positions: that explainability is unnecessary if performance is high, or that explainability alone will solve stakeholder concerns. Instead, the reviewed studies suggest that explanation functions as a contextual resource. What matters is not maximal transparency in an abstract sense, but whether the explanation available is sufficient for appropriate use, challenge, communication, and review in a given setting. This is why the concept of practical or actionable explainability is more useful than a generic call for “transparent AI.” A clinician, a patient, a governance board, and a hospital IT team do not need exactly the same explanation. They need explanations suited to their roles in the implementation chain. Designing for that plurality is a difficult but necessary part of responsible deployment.

The same contextual logic applies to oversight. The reviewed literature repeatedly shows that human oversight should not be treated as a symbolic reassurance mechanism. A clinician is not meaningfully “in the loop” if the system’s recommendations are opaque, if the user lacks time to interrogate them, or if

organizational culture encourages passive acceptance of machine output. Effective oversight requires well-designed interfaces, appropriate uncertainty communication, adequate training, and institutional permission to contest or override the system. In this sense, the literature invites a shift from the slogan of human-in-the-loop to the more substantive idea of calibrated oversight. Oversight must be proportionate to the stakes, informed by the model's limitations, and supported by workflow and policy.

The review also suggests that implementation should be analyzed as labor transformation. AI changes who does what, when, and under what conditions of responsibility. It can reduce repetitive burden, but it can also create new forms of verification work, quality assurance, exception handling, and legal exposure. These effects are especially important in medicine, where professional identity and trust are closely tied to responsibility for judgment. The qualitative literature on primary care, nursing, and critical care suggests that frontline staff are not simply resisting innovation. Rather, they are interpreting AI through the lens of practical accountability and care quality. If implementation threatens empathy, professional discretion, or situational awareness, skepticism is understandable. This implies that successful implementation requires institutions to treat professional concerns not as obstacles to be overcome, but as diagnostic signals about where the technology is misaligned with clinical reality.

Governance emerges from the literature as the domain that links all the others. Explainability, safety, workflow fit, and user training all depend on governance decisions about evidence, procurement, validation, monitoring, and update control. The strongest frameworks in the review—especially the governance model proposed by Reddy et al. (2020) and the lifecycle principles articulated in FUTURE-AI (Lekadir et al., 2025)—treat implementation as ongoing stewardship rather than one-time rollout. This is a crucial advance over earlier, more static thinking. AI systems can drift, user populations can shift, and organizational pressures can change how a tool is used. Therefore, governance must be continuous. It needs to include local validation, feedback loops, auditability, and the capacity to suspend or revise deployment when performance or context changes.

From a social-science perspective, the literature also shows that AI implementation is a problem of legitimacy. A hospital or health system is not simply choosing a technical instrument; it is deciding how to distribute authority among clinicians, algorithms, vendors, patients, and administrators. That decision has implications for trust, labor, inequality, and institutional accountability. Studies of stakeholder views indicate that concerns about privacy, autonomy, professional authority, and fairness are not peripheral to adoption; they are part of the conditions under which adoption becomes socially acceptable. For a journal like *IJTSS*, this is a crucial insight. AI in healthcare is not an isolated clinical technology but a site where technical systems and social values are actively renegotiated.

Several practical recommendations follow from the reviewed literature. First, health systems should prioritize use cases where workflow problems are already clear and where benefits can be evaluated with concrete operational and clinical measures. Second, local validation should be treated as a precondition of deployment rather than a formality. Third, implementation plans should include human factors evaluation, not only model performance evaluation. Fourth, governance should be designed across the lifecycle, including update management, auditability, post-deployment monitoring, and criteria for withdrawal. Fifth, training should address not only how to use a tool but how to interpret it critically, communicate about it, and recognize when reliance is inappropriate.

This review has limitations. It is not a systematic review and does not claim exhaustive coverage of all healthcare AI implementation literature. The search was confined to PubMed and to free full-text publications, which means relevant work indexed elsewhere or available only behind paywalls may have been missed. The review also covers multiple forms of evidence, from conceptual papers to empirical studies, which is a strength for thematic synthesis but a limitation for precise comparison. In addition, much of the literature still reflects early or partial implementation rather than mature, long-term deployment. This means that some conclusions remain provisional. Even so, the convergence of findings across study designs is notable and supports meaningful synthesis.

Future research should move in at least four directions. First, it should examine post-deployment performance and governance more systematically rather than focusing mainly on pre-implementation promise. Second, it should generate stronger evidence on patient outcomes, safety outcomes, and economic consequences in routine care. Third, it should investigate the hidden labor of AI implementation, including verification work, exception management, and the burden of maintaining responsible use. Fourth, it should expand comparative analysis across clinical settings, especially beyond well-resourced institutions and beyond documentation or imaging. More research is especially needed on triage, nursing workflows, low-resource contexts, and the long-term effects of generative AI on clinical records and communication.

In sum, the literature suggests that the future of AI in medicine will be shaped less by abstract technical superiority than by implementation quality. The question is not whether AI can enter healthcare, because in limited ways it already has. The more important question is what forms of AI can be made safe, useful, accountable, and compatible with the moral and organizational structure of care. That is ultimately not only a technical question, but also an institutional and social one.

Conclusions

The reviewed free full-text PubMed literature indicates that the implementation of artificial intelligence in healthcare is best understood as a sociotechnical process rather than a simple matter of software deployment. Across studies on documentation, triage, governance, explainability, safety, and professional experience, the same pattern recurs: AI becomes clinically meaningful only when it is aligned with human work, integrated into organizational routines, and governed across its full lifecycle. Accuracy remains important, but it is not enough. Tools must also be explainable in context, safe in use, compatible with professional judgment, and accountable after deployment.

The literature further suggests that not all clinical AI applications face the same implementation conditions. Bounded tasks with visible workflow burdens, such as documentation, currently appear more implementable than high-context tasks such as triage or primary care reasoning. This difference is important because it shows that adoption is shaped by task structure, institutional readiness, and the degree to which responsibility can be clearly distributed. It also shows why governance, training, and human factors cannot be treated as secondary concerns.

For technology-oriented social-science research, the main lesson is that AI in medicine should not be evaluated only through performance claims. It should also be assessed through its effects on labor, accountability, fairness, and institutional legitimacy. For healthcare organizations, the practical implication is equally clear: successful implementation requires local validation, calibrated oversight, meaningful user training, and post-deployment monitoring rather than one-time rollout.

Future work should strengthen real-world evaluation, especially with respect to patient outcomes, safety, economic effects, and the long-term consequences of generative AI and other rapidly evolving tools. Until then, the strongest conclusion supported by the literature is a restrained one: clinical AI is most likely to succeed when it is introduced not as a replacement for medicine's human and organizational foundations, but as a carefully governed supplement to them.

Acknowledgments: No external funding or institutional research assistance was received for the preparation of this review article.

Ethical Considerations: This article is a structured narrative review of previously published literature and does not involve original research with human participants, animals, or identifiable personal data. Formal ethics committee approval was therefore not required.

Conflicts of Interest: No conflicts of interest to declare.

REFERENCES

1. Albrecht, M., Shanks, D., Shah, T., Hudson, T., Thompson, J., Filardi, T., Wright, K., Ator, G. A., & Smith, T. R. (2025). Enhancing clinical documentation with ambient artificial intelligence: A quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. *JAMIA Open*, 8(1), ooaf013. <https://doi.org/10.1093/jamiaopen/ooaf013>
2. Alkhanbouli, R., Almadhaani, H. M. A., Alhosani, F., & Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions. *BMC Medical Informatics and Decision Making*, 25(1), 110. <https://doi.org/10.1186/s12911-025-02944-6>
3. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1), 310. <https://doi.org/10.1186/s12911-020-01332-6>
4. Blease, C., Kaptchuk, T. J., Bernstein, M. H., Mandl, K. D., Halamka, J. D., & DesRoches, C. M. (2019). Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. *Journal of Medical Internet Research*, 21(3), e12802. <https://doi.org/10.2196/12802>
5. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3), 231–237. <https://doi.org/10.1136/bmjqs-2018-008370>

6. Choudhury, A., & Asan, O. (2020). Role of artificial intelligence in patient safety outcomes: Systematic literature review. *JMIR Medical Informatics*, 8(7), e18599. <https://doi.org/10.2196/18599>
7. Kostick-Quenet, K. M., & Gerke, S. (2022). AI in the hands of imperfect users. *npj Digital Medicine*, 5(1), 197. <https://doi.org/10.1038/s41746-022-00737-z>
8. Larsson, I., Siira, E., Nygren, J. M., Petersson, L., Svedberg, P., Nilsen, P., & Neher, M. (2025). Integrating AI-based triage in primary care: A qualitative study of Swedish healthcare professionals' experiences applying normalization process theory. *BMC Primary Care*, 26(1), 340. <https://doi.org/10.1186/s12875-025-03057-9>
9. Lee, C., Britto, S., & Diwan, K. (2024). Evaluating the impact of artificial intelligence (AI) on clinical documentation efficiency and accuracy across clinical settings: A scoping review. *Cureus*, 16(11), e73994. <https://doi.org/10.7759/cureus.73994>
10. Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., ... Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388, e081554. <https://doi.org/10.1136/bmj-2024-081554>
11. Militello, L. G., Diiulio, J., Wilson, D. L., Nguyen, K. A., Harle, C. A., Gellad, W., & Lo-Ciganic, W.-H. (2025). Using human factors methods to mitigate bias in artificial intelligence-based clinical decision support. *Journal of the American Medical Informatics Association*, 32(2), 398–403. <https://doi.org/10.1093/jamia/ocae291>
12. Mlodzinski, E., Wardi, G., Viglione, C., Nemati, S., Crotty Alexander, L., & Malhotra, A. (2023). Assessing barriers to implementation of machine learning and artificial intelligence-based tools in critical care: Web-based survey study. *JMIR Perioperative Medicine*, 6, e41056. <https://doi.org/10.2196/41056>
13. Ratti, E., Morrison, M., & Jakab, I. (2025). Ethical and social considerations of applying artificial intelligence in healthcare—a two-pronged scoping review. *BMC Medical Ethics*, 26(1), 68. <https://doi.org/10.1186/s12910-025-01198-1>
14. Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3), 491–497. <https://doi.org/10.1093/jamia/ocz192>
15. Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, 112(1), 22–28. <https://doi.org/10.1177/0141076818815510>
16. Rony, M. K. K., Kayesh, I., Bala, S. D., Akter, F., & Parvin, M. R. (2024). Artificial intelligence in future nursing care: Exploring perspectives of nursing professionals - A descriptive qualitative study. *Heliyon*, 10(4), e25718. <https://doi.org/10.1016/j.heliyon.2024.e25718>
17. Shaw, J., Rudzicz, F., Jamieson, T., & Goldfarb, A. (2019). Artificial intelligence and the implementation challenge. *Journal of Medical Internet Research*, 21(7), e13659. <https://doi.org/10.2196/13659>
18. Van der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., Oswald, M., & Peek, N. (2021). Trading off accuracy and explainability in AI decision-making: Findings from 2 citizens' juries. *Journal of the American Medical Informatics Association*, 28(10), 2128–2138. <https://doi.org/10.1093/jamia/ocab127>
19. Vo, V., Chen, G., Aquino, Y. S. J., Carter, S. M., Do, Q. N., & Woode, M. E. (2023). Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. *Social Science & Medicine*, 338, 116357. <https://doi.org/10.1016/j.socscimed.2023.116357>
20. Yin, J., Ngiam, K. Y., & Teo, H. H. (2021). Role of artificial intelligence applications in real-life clinical practice: Systematic review. *Journal of Medical Internet Research*, 23(4), e25759. <https://doi.org/10.2196/25759>
21. Yoo, J., Hur, S., Hwang, W., & Cha, W. C. (2023). Healthcare professionals' expectations of medical artificial intelligence and strategies for its clinical implementation: A qualitative study. *Healthcare Informatics Research*, 29(1), 64–74. <https://doi.org/10.4258/hir.2023.29.1.64>