Technema: Journal of Intelligent Engineering and Computing
Vol 1 No 2 June 2026, Hal 139-151
ISSN: 3163-6020 (Print) ISSN: 3163-5997 (Electronic)
Open Access: https://sovereignresearch.org/technema

The Rise of Agentic Artificial Intelligence in Medicine: A Systematic Review of Autonomous Systems, Multi-Agent Collaboration, and Ethical Governance

Marzuki Pilliang[1*], Meilita Dwi Paundrianagari[2], Chindy Amir[3]
[1-2] Universitas Salakanagara, Indonesia
email: marzuki.pilliang@ieee.org[1]

Article Info :
Received:
01-06-2026
Revised:
12-06-2026
Accepted:
19-06-2026

Abstract

Artificial Intelligence (AI) in medicine has traditionally functioned as a passive decision-support tool that provides recommendations without autonomous action. The emergence of Agentic AI introduces a new paradigm in which AI systems can plan, execute, and self-correct tasks, creating a shift toward more collaborative human–machine interaction. However, existing literature remains fragmented, with technical development, multi-agent coordination, and ethical governance often discussed separately. This systematic review aims to synthesize current evidence on Agentic AI in medicine by examining autonomous architectures, multi-agent collaboration, clinical validation, ethical challenges, and future implementation barriers. The study followed the PRISMA 2020 framework and employed thematic narrative synthesis of 34 primary studies identified from major academic databases. The findings indicate that orchestrator-based multi-agent architectures integrated with large language models and retrieval-augmented generation represent the dominant approach in medical Agentic AI development. However, real-world clinical validation remains limited, with most studies still focusing on simulations and proof-of-concept evaluations. Furthermore, ethical and governance studies represented the largest research category, highlighting concerns regarding accountability, transparency, and human oversight. This review concludes that the future adoption of Agentic AI requires simultaneous advancement of technical capabilities and governance frameworks, supported by evaluation approaches that consider accuracy, computational efficiency, clinical safety, and human control.

Keywords: *Accountability, Agentic Artificial Intelligence, Ethical Governance, Multi-agent Collaboration, System Autonomy, Systematic Review.*

INTRODUCTION

Artificial intelligence (AI) has become an integral component of contemporary healthcare, supporting a broad spectrum of clinical activities ranging from disease diagnosis and medical imaging to treatment planning and healthcare management. Existing AI applications have substantially improved the efficiency and accuracy of clinical decision-making by functioning as decision-support systems that assist healthcare professionals in interpreting complex medical information (Acharya et al., 2025). Despite these advances, most healthcare AI systems remain fundamentally reactive because they rely on explicit human instructions before initiating any subsequent action or recommendation. Recent developments in large language models, autonomous planning, adaptive reasoning, and memory mechanisms have introduced Agentic AI, allowing intelligent systems to perceive clinical contexts, formulate objectives, execute sequential tasks, and continuously refine their behaviour with limited human intervention (Abou Ali et al., 2025). This technological transition represents a fundamental change in the relationship between humans and AI, shifting intelligent systems from passive analytical tools toward active collaborators within clinical workflows. As a consequence, Agentic AI is increasingly regarded as a new generation of intelligent teammates capable of participating alongside healthcare professionals in complex medical decision-making processes (James Zou et al., 2025).

The growing autonomy of intelligent agents has stimulated rapid innovation across numerous medical specialties, where AI is beginning to perform tasks that extend well beyond conventional prediction and classification. Current developments indicate that autonomous agents are being designed to support patient monitoring, clinical documentation, diagnostic reasoning, multidisciplinary

collaboration, and personalized healthcare delivery through increasingly sophisticated decision-making capabilities (AlZu'Bi et al., 2025). Advances in system architecture have further enabled these agents to integrate retrieval-augmented generation, persistent memory, and multi-agent collaboration, thereby improving contextual reasoning and interoperability across healthcare information systems (Tanasă Simona-Vasilica Oprea, 2026). These innovations demonstrate that future healthcare AI will depend not only on predictive performance but also on the ability of autonomous systems to coordinate complex workflows while interacting with clinicians and digital infrastructures. Similar transformations have been reported in radiology, where intelligent agents are evolving from image interpretation tools into workflow coordinators that actively assist reporting and clinical decision processes (Dietrich, 2025). Such progress suggests that Agentic AI is gradually becoming embedded within routine healthcare operations rather than remaining an isolated analytical technology.

Although the technological potential of Agentic AI continues to expand, its increasing autonomy has also generated important concerns regarding accountability, transparency, and governance in healthcare environments. Autonomous agents are capable of making independent decisions, coordinating multiple tasks, and interacting with other intelligent systems, creating decision pathways that are substantially more complex than those of traditional clinical decision-support tools (Chow et al., 2026). Previous research has shown that responsibility may become increasingly difficult to determine when multiple autonomous agents collectively contribute to a clinical outcome, giving rise to the phenomenon of diffused responsibility in AI-assisted healthcare (Bleher et al., 2022). The challenge extends beyond technical performance because healthcare institutions must also ensure that autonomous systems remain explainable, trustworthy, and aligned with existing ethical and regulatory expectations. Scholars therefore argue that addressing responsibility gaps requires governance models based on shared accountability among developers, healthcare providers, and regulatory institutions rather than assigning responsibility to a single actor (Lang et al., 2023). These concerns indicate that the successful implementation of Agentic AI depends not only on technological advancement but also on the development of governance frameworks capable of supporting safe and responsible clinical adoption.

Despite the remarkable progress achieved in recent years, the current body of literature on Agentic AI in healthcare remains fragmented in both conceptual scope and practical implementation. Existing reviews consistently indicate that research has advanced more rapidly in developing autonomous architectures than in establishing common definitions, evaluation frameworks, and governance principles capable of guiding their safe clinical deployment (Njei et al., 2026). The conceptual distinction between *AI agents* and *Agentic AI* also remains inconsistent across the literature, resulting in overlapping terminology that complicates comparisons among studies and obscures the actual level of system autonomy being investigated (Collaco et al., 2026). This lack of conceptual consistency has made it increasingly difficult to evaluate technological maturity across different healthcare domains and to identify which approaches are sufficiently robust for real-world clinical adoption. Furthermore, many published studies continue to emphasize prototype development and architectural innovation while providing only limited discussion of implementation challenges, organizational readiness, and long-term governance. Consequently, the expanding volume of publications has not yet translated into a coherent body of knowledge capable of supporting evidence-based adoption of Agentic AI in healthcare.

Another important limitation concerns the imbalance between technological performance and real-world clinical validation. Most reported applications remain confined to simulation environments or laboratory evaluations, whereas prospective clinical validation, interoperability testing, and large-scale implementation studies remain relatively limited (Liu et al., 2026). Although several studies have demonstrated promising improvements in reasoning capability and clinical accuracy through autonomous agent systems, these benefits are frequently accompanied by increased computational costs, longer response times, and greater system complexity that may reduce their practicality in routine healthcare settings (Liu et al., 2026). Similar concerns have been raised regarding the ethical implications of multi-agent collaboration, where increasing autonomy introduces new challenges related to transparency, human oversight, and accountability that cannot be addressed solely through technical optimization (Xie et al., 2026). These findings suggest that technological advancement alone is insufficient to ensure successful implementation because clinical adoption also depends on organizational, ethical, and regulatory preparedness. Such conditions reveal a critical need for a

comprehensive synthesis that simultaneously examines technological capabilities, implementation readiness, and governance challenges within a unified analytical framework.

This study addresses these limitations by providing a comprehensive literature review that integrates discussions on the architecture, clinical applications, implementation challenges, and governance implications of Agentic AI within healthcare. Rather than examining autonomous intelligent systems from a single technological perspective, this review positions technical innovation, clinical implementation, and ethical governance as interconnected dimensions that collectively determine the readiness of Agentic AI for healthcare practice. The study is designed to provide a structured understanding of how autonomous intelligent agents are evolving across different clinical domains while identifying the principal barriers that continue to limit their widespread adoption. It also offers a critical synthesis of current evidence to clarify conceptual inconsistencies and highlight directions for future investigation. Through this integrated perspective, the review is expected to contribute to the development of a more coherent foundation for future research, informed policy formulation, and responsible implementation of Agentic AI in modern healthcare.

RESEARCH METHODS

This study employed a non-empirical systematic literature review following the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA 2020) guidelines to ensure a transparent, reproducible, and traceable review process (Page et al., 2021). A narrative thematic synthesis was adopted because the reviewed studies differed substantially in research objectives, methodological approaches, application domains, and reported outcomes, making quantitative meta-analysis inappropriate. Literature searches were conducted in IEEE Xplore, SpringerLink, ScienceDirect, MDPI, and Google Scholar, covering publications published between 2022 and 2026. Search strategies were adapted to the syntax of each database using Boolean combinations of keywords, including *"agentic AI"*, *"AI agent"*, *"multi-agent system"*, *"autonomous AI"*, *"healthcare"*, *"medicine"*, *"clinical"*, and *"hospital"*, with additional terms such as *"LLM-based agent"*, *"retrieval augmented generation (RAG)"*, *"clinical decision support"*, and *"digital health"* where appropriate. Study eligibility was established using the PICOC (Population, Intervention, Comparison, Outcomes, and Context) framework, considering only peer-reviewed English-language publications that investigated autonomous AI systems, multi-agent collaboration, clinical implementation, or ethical and governance issues in healthcare. The eligibility criteria applied throughout the screening process are summarized in Table 1.

Table 1. Eligibility Criteria

Aspect	Inclusion Criteria	Exclusion Criteria
Population/Context	Studies investigating Agentic AI, AI agents, or multi-agent systems in healthcare, medicine, hospitals, clinical practice, or public health.	Studies conducted exclusively in non-medical domains without clear healthcare relevance.
Intervention/Concept	Studies discussing autonomous AI architectures, multi-agent collaboration, LLM-based agents, retrieval-augmented generation (RAG), or ethical and governance aspects of Agentic AI.	Studies addressing only conventional machine learning or deep learning without autonomous agent capabilities.
Outcomes	Studies reporting system architectures, technical performance, clinical validation, implementation findings, ethical analysis, governance frameworks, or implementation challenges.	Studies lacking substantive technical, conceptual, or empirical findings.
Study Design	Experimental studies, simulation studies, case studies, proof-of-concept research, conceptual	Editorials, commentaries, opinion papers, letters,

	frameworks, qualitative studies, and systematic or scoping reviews used as supporting evidence.	duplicate publications, and non-peer-reviewed preprints.
Language & Accessibility	English-language articles with full-text availability.	Non-English publications or articles without accessible full-text versions.

The study selection followed the PRISMA 2020 workflow, including identification, duplicate removal using Mendeley, title and abstract screening, full-text eligibility assessment, and final inclusion of relevant studies. Two independent reviewers performed the screening process using the Rayyan platform, while disagreements were resolved through discussion or consultation with a third reviewer. Inter-reviewer agreement was assessed using Cohen's Kappa, with a minimum acceptable value of 0.80 to ensure screening reliability (Ismael et al., 2021). Data extraction was conducted using a standardized extraction form that recorded bibliographic information, study characteristics, application domains, autonomous architectures, levels of autonomy, large language model integration, multi-agent coordination mechanisms, clinical validation approaches, ethical considerations, governance issues, and reported implementation barriers. Because the included literature encompassed technical, clinical, conceptual, and ethical studies, methodological quality was assessed according to study type by considering reproducibility, validation strategy, methodological transparency, and analytical rigor rather than applying a single appraisal instrument to all studies. Finally, the extracted evidence was synthesized through thematic narrative analysis organized around five predefined research questions addressing autonomous system architectures, multi-agent collaboration, clinical applications and validation, ethical governance, and future research directions, thereby ensuring that the review process remains transparent and reproducible.

RESULTS AND DISCUSSION

Literature Selection Process and Characteristics of Included Studies

The systematic search process identified 518 records from five electronic databases, consisting of 122 articles from IEEE Xplore, 91 articles from SpringerLink, 37 articles from ScienceDirect, 28 articles from MDPI, and 240 articles from Google Scholar. The application of the PRISMA 2020 framework enabled a transparent and reproducible selection procedure by documenting each stage of identification, screening, eligibility assessment, and inclusion. After duplicate removal using Mendeley reference management software, 460 unique records remained for title and abstract screening. This initial filtering stage was necessary because the emerging field of Agentic AI contains substantial conceptual overlap with conventional artificial intelligence research that does not involve autonomous reasoning, planning, or action execution capabilities. The increasing number of retrieved publications reflects the rapid expansion of research interest in autonomous AI systems for healthcare environments, particularly following advances in large language models and agent-based architectures (Page et al., 2021; Acharya et al., 2025).

The title and abstract screening stage resulted in the exclusion of 340 records that did not satisfy the predefined eligibility criteria. A substantial proportion of excluded studies focused on conventional machine learning or deep learning approaches without incorporating autonomous agent characteristics, indicating that the distinction between predictive AI and agentic intelligence remains an important conceptual boundary in healthcare research. Other excluded publications addressed AI applications outside medical contexts, lacked English accessibility, or represented non-substantive academic formats such as editorials and opinion articles. This pattern demonstrates that although artificial intelligence adoption in healthcare has expanded significantly, research specifically addressing autonomous decision-making systems remains comparatively specialized. The identification of these limitations during screening strengthened the thematic relevance of the final evidence base used in this review (Durante et al., 2024).

Following full-text assessment, 86 additional records were excluded because they failed to address the research questions related to autonomous architectures, multi-agent collaboration, clinical applications, ethical governance, or implementation barriers. The most frequent exclusion reasons involved literature reviews without original analytical contributions and studies that discussed AI systems without sufficient evidence of agency, autonomy, or healthcare applicability. Manual backward

and forward reference tracking from key publications did not identify additional studies meeting the inclusion criteria, indicating that the database search strategy provided adequate coverage of the targeted research domain. The final synthesis consisted of 34 peer-reviewed studies that represented technical investigations, conceptual frameworks, and ethical or governance analyses of Agentic AI in medicine. This composition reflects the multidisciplinary nature of autonomous healthcare AI, where engineering innovation, clinical translation, and normative considerations develop simultaneously (Basile Njei et al., 2026).

The characteristics of the included studies reveal a substantial acceleration of Agentic AI research during the period 2024–2026. Among the 34 selected articles, 27 studies (79.4%) were published between 2024 and 2026, while seven studies (20.6%) originated from 2022–2023. The distribution indicates that Agentic AI has entered an intensive development phase following the emergence of advanced generative models capable of reasoning, retrieval augmentation, and autonomous task execution. The classification of study characteristics is presented in Table 1, which summarizes publication periods, methodological orientations, and clinical application domains identified through the thematic synthesis process. This distribution provides an overview of how recent literature has shifted from theoretical exploration toward practical experimentation and governance-oriented analysis.

Table 1. Characteristics of Included Studies

Category	Classification	Number of Studies	Percentage
Publication Period	2022–2023	7	20.6%
	2024–2026	27	79.4%
Study Design	Ethical, qualitative, and policy analysis	14	41.2%
	Technical or empirical studies	11	32.4%
	Conceptual frameworks and architecture proposals	9	26.5%
Clinical Domain	Radiology and medical imaging	11	32.4%
	Oncology	6	17.6%
	Cardiology	4	11.8%
	Hospital workflow and health information systems	4	11.8%
	Drug discovery and bioinformatics	3	8.8%

The methodological composition of the reviewed literature indicates that ethical, qualitative, and policy-oriented studies represented the largest category, accounting for 14 studies (41.2%), followed by technical or empirical investigations with 11 studies (32.4%) and conceptual architecture studies with nine studies (26.5%). This distribution differs from earlier AI healthcare research patterns that were predominantly focused on algorithmic performance evaluation and predictive accuracy. The strong representation of governance-oriented studies suggests that autonomous AI introduces additional concerns related to accountability, transparency, and human oversight beyond conventional clinical decision support systems. Studies examining responsibility allocation and ethical implementation have emphasized that increasing autonomy requires corresponding mechanisms for maintaining trust and regulatory compliance (Bleher et al., 2022; Lang et al., 2023). The findings indicate that the scientific discussion surrounding Agentic AI in medicine has expanded from technological capability assessment toward broader socio-technical evaluation.

The final evidence base demonstrates that Agentic AI in medicine is developing as an interdisciplinary research field involving artificial intelligence engineering, clinical science, and healthcare governance. The coexistence of technical prototypes, conceptual models, and ethical analyses indicates that the maturity of this technology cannot be evaluated solely through computational performance metrics. Instead, the reviewed studies highlight the importance of examining autonomy levels, clinical reliability, explainability, and institutional readiness as interconnected dimensions of implementation. This perspective aligns with emerging frameworks that consider autonomous medical AI as a socio-technical system requiring coordination between algorithms, healthcare professionals, and

regulatory structures (Meszaros et al., 2022; Eisenberg et al., 2026). The synthesis of included studies provides the foundation for subsequent analysis of Agentic AI architectures, multi-agent collaboration mechanisms, clinical validation challenges, and governance requirements.

Autonomous Architectures and Levels of Agentic Intelligence in Medical Systems

The synthesis of the included studies identified three dominant architectural patterns underlying Agentic AI implementation in medical environments, namely single-agent systems, orchestrator-based multi-agent systems, and decentralized multi-agent systems. These architectural categories were derived from the analysis of 11 technical studies and nine conceptual framework studies included in the review. Unlike conventional artificial intelligence systems that primarily generate predictions from predefined models, Agentic AI architectures incorporate autonomous reasoning, planning, tool utilization, and adaptive task execution mechanisms. The emergence of these architectures reflects a shift from passive decision-support systems toward interactive computational entities capable of coordinating complex medical workflows. This transformation corresponds with the broader conceptualization of agentic intelligence as an AI paradigm designed to achieve complex objectives through autonomous perception, reasoning, and action cycles (Acharya et al., 2025).

The single-agent architecture represents the earliest and simplest implementation pattern identified in the reviewed literature, where one large language model (LLM)-based agent performs reasoning, information retrieval, and task execution through integrated tools. These systems generally demonstrate limited to moderate autonomy because decision processes remain concentrated within a single computational entity. Although this architecture offers advantages in simplicity, deployment efficiency, and easier monitoring, it also introduces limitations when handling complex clinical scenarios requiring multiple specialized reasoning processes. The absence of internal role separation may increase the risk of incomplete analysis, particularly when medical problems require simultaneous interpretation of heterogeneous information sources such as imaging data, clinical records, and biomedical literature. Recent surveys have emphasized that autonomous medical systems require increasingly sophisticated coordination mechanisms because healthcare decision-making involves multiple interacting domains rather than isolated computational tasks (Durante et al., 2024).

The orchestrator-based multi-agent architecture emerged as the most frequently identified approach among the reviewed studies because it enables task decomposition through a central coordinating agent. In this configuration, an orchestrator agent assigns specific subtasks to specialized agents, such as diagnostic reasoning, evidence retrieval, patient monitoring, or clinical recommendation generation. Studies by AlZu'Bi et al. (2025), Tanasă et al. (2026), and Kalathas et al. (2026) demonstrated that this architecture can improve workflow organization by combining specialized capabilities while maintaining centralized coordination. However, dependence on the orchestrator introduces potential vulnerabilities because errors at the coordination layer may propagate across downstream agents. This condition highlights the importance of designing reliable supervisory mechanisms capable of monitoring agent interactions and preventing cascading failures in clinical contexts.

The architectural classification obtained from the literature synthesis is summarized in Table 2, which presents the defining characteristics, representative studies, and estimated autonomy levels of each Agentic AI architecture. The classification demonstrates that increasing autonomy is generally associated with greater coordination complexity, as systems progress from individual reasoning agents toward collaborative computational ecosystems. The reviewed evidence indicates that multi-agent architectures dominate current medical Agentic AI research because they provide greater flexibility for managing complex clinical objectives. Nevertheless, higher autonomy does not automatically guarantee improved clinical reliability because additional agent interactions may introduce communication errors, computational burdens, and difficulties in tracing decision pathways. These findings reinforce the perspective that architecture selection should consider both functional capability and governance requirements.

Table 2. Classification of Agentic AI Architectures in Healthcare

Architecture Type	Main Characteristics	Representative Studies	Autonomy Level
Single-agent architecture	One LLM-based agent performing reasoning, retrieval, and tool utilization	Acharya et al. (2025); Woo et al. (2025)	Limited–moderate autonomy
Orchestrator-based multi-agent architecture	Central coordinator delegates tasks to specialized agents using structured workflows	AlZu'Bi et al. (2025); Tanasă et al. (2026); Kalathas et al. (2026)	Moderate–high autonomy
Decentralized multi-agent architecture	Peer-to-peer collaboration, consensus mechanisms, and distributed reasoning	Zhou et al. (2025); Pylarinou et al. (2026)	High autonomy

The decentralized multi-agent architecture represents a more advanced direction in Agentic AI development by enabling multiple autonomous agents to collaborate without complete dependence on a single controller. Zhou et al. (2025) demonstrated this approach through agentic bioinformatics systems that coordinate literature mining, hypothesis generation, and computational validation processes. Similarly, Pylarinou et al. (2026) implemented a six-agent framework for surgical monitoring by combining ReAct reasoning, Reflexion-based self-correction, and confidence-guided human oversight. These approaches demonstrate that decentralized collaboration can improve adaptability when addressing complex medical tasks involving diverse data modalities. However, distributed autonomy also increases the complexity of verification because each agent may contribute independent reasoning pathways that require systematic auditing.

Large language model integration was identified as a central component across the reviewed architectures, appearing in 85.3% of included studies. LLMs provide the reasoning foundation that enables agents to interpret clinical information, generate responses, and coordinate interactions with external tools, while retrieval-augmented generation (RAG) mechanisms are frequently adopted to improve factual grounding and reduce hallucination risks. Kalathas et al. (2025) and Woo et al. (2025) reported that RAG-enhanced agent systems improve evidence retrieval and clinical answer accuracy by connecting generative models with validated knowledge sources. Despite these advantages, dependence on LLM-based reasoning introduces challenges related to reliability, computational requirements, and transparency of generated decisions. The integration of LLMs therefore requires complementary validation mechanisms rather than relying solely on model capability improvements.

The overall findings indicate that Agentic AI architectures in medicine are progressing toward higher levels of autonomy through increasingly sophisticated coordination mechanisms. The dominance of multi-agent approaches suggests that future medical AI systems will likely operate as collaborative networks rather than isolated models. However, autonomy expansion must be accompanied by mechanisms that maintain interpretability, human supervision, and responsibility allocation throughout the decision-making process. Chow (2026) emphasized that ethical medical agents require integrated modeling of autonomy, persistence, and human-AI cooperation to prevent inappropriate delegation of clinical authority. These findings position architectural design as not only a technical challenge but also a critical determinant of safety, accountability, and sustainable adoption in healthcare environments.

Multi-Agent Collaboration, Clinical Applications, and Real-World Validation Challenges

The analysis of the included studies revealed that multi-agent collaboration in medical Agentic AI systems is primarily organized through three operational mechanisms: domain-based task delegation, iterative communication protocols, and collective decision aggregation. These mechanisms allow individual agents with specialized functions to contribute to a shared clinical objective rather than operating as independent computational units. Task specialization enables systems to distribute complex medical reasoning processes into smaller and more manageable components, while communication protocols facilitate continuous information exchange between agents during problem-solving cycles. The emergence of collaborative architectures indicates that medical Agentic AI is

moving beyond single-model performance optimization toward coordinated intelligence capable of addressing multidimensional healthcare problems. This development is consistent with research describing agentic systems as collaborative frameworks where autonomous entities interact dynamically to achieve complex goals (Durante et al., 2024).

The first collaboration mechanism identified in the literature involves domain-specific agent specialization, where each agent is assigned a particular clinical or analytical responsibility. For example, an agent may focus on evidence retrieval, another on diagnostic interpretation, and another on treatment recommendation assessment. Pylarinou et al. (2026) demonstrated this approach through a surgical monitoring system consisting of six specialized agents that performed continuous observation, reasoning, and correction processes using ReAct and Reflexion mechanisms. Similarly, Kalathas et al. (2026) developed a multi-agent healthcare framework incorporating Corrective RAG, allowing agents to identify retrieval failures and refine information acquisition processes autonomously. These findings suggest that specialization improves system adaptability, although coordination quality becomes a decisive factor determining overall reliability.

Iterative communication represents the second major mechanism supporting multi-agent collaboration, particularly through reasoning frameworks such as ReAct and Reflexion. These approaches allow agents to generate intermediate reasoning steps, evaluate previous outputs, and modify responses when inconsistencies are detected. The incorporation of self-correction mechanisms provides an important improvement compared with static AI models because medical environments require continuous adjustment based on changing evidence and contextual information. However, iterative interaction also increases computational demands because each additional reasoning cycle requires additional processing time and resource consumption. Research examining autonomous AI teams has highlighted that increased collaboration capability may create a trade-off between reasoning performance, efficiency, and operational scalability (James Zou et al., 2025).

The comparative characteristics of multi-agent collaboration mechanisms and their associated challenges are summarized in Table 3. The synthesis demonstrates that collaborative approaches provide several advantages, including improved robustness, multimodal information processing, and enhanced specialization of medical reasoning tasks. Nevertheless, these advantages are accompanied by emerging risks related to communication complexity, bias propagation, and limited transparency of collective decision processes. The table highlights that increasing the number of interacting agents does not necessarily produce proportional improvements because system performance depends on coordination strategies, evaluation mechanisms, and governance structures. This finding supports the argument that multi-agent medical systems should be evaluated as integrated socio-technical frameworks rather than collections of independent AI components.

Table 3. Multi-Agent Collaboration Mechanisms and Associated Challenges in Medical Agentic AI

Collaboration Mechanism	Main Function	Potential Benefits	Key Challenges
Domain-based task delegation	Assigning specialized roles to individual agents	Improved task specialization and workflow efficiency	Coordination dependency and role conflicts
Iterative reasoning communication (ReAct/Reflexion)	Continuous reasoning, evaluation, and self-correction	Increased adaptability and error reduction	Higher computational cost and latency
Consensus-based collaboration	Combining outputs from multiple agents	Improved decision robustness and uncertainty management	Difficult auditability and responsibility attribution

Clinical application analysis showed that Agentic AI research has expanded across multiple medical domains, with radiology and medical imaging representing the most frequently studied area. Eleven of the 34 included studies (32.4%) focused on imaging-related applications, followed by oncology with six studies (17.6%), cardiology and hospital workflow systems with four studies each (11.8%), and drug discovery or bioinformatics with three studies (8.8%). The strong representation of

radiology reflects the suitability of multimodal AI systems for processing complex visual and textual medical information. Dietrich (2025) and Gruson et al. (2026) emphasized that radiology and laboratory medicine are among the earliest fields experiencing a transition from automated analysis toward autonomous AI-assisted workflows. Similar developments have been reported in orthopedics, ophthalmology, and neurological care, where agent-based systems are being explored for decision support and evidence synthesis (Billi et al., 2026; Grammenos et al., 2026).

Despite expanding clinical applications, the synthesis identified a substantial gap between technological development and real-world clinical validation. Only six of the 34 reviewed studies (17.6%) evaluated Agentic AI systems using real clinical datasets, while merely two studies (5.9%) reported prospective testing or implementation within actual healthcare environments. This finding indicates that most current Agentic AI research remains positioned at the proof-of-concept or prototype validation stage rather than routine clinical deployment. Ta et al. (2022) demonstrated that successful healthcare AI adoption depends not only on technical performance but also on workflow integration, usability, and institutional readiness. Similarly, Gill et al. (2023) emphasized the importance of rigorous clinical evaluation frameworks, including randomized controlled trial designs, for assessing AI-assisted healthcare interventions.

A notable example of the performance–efficiency trade-off was reported by Liu et al. (2026), who compared an agent-based system with a conventional LLM baseline for medical tasks. The agentic system achieved improved accuracy, reaching 60.3% performance in specific tasks, but required more than ten times higher token consumption and generated latency exceeding twice that of the baseline model. These findings demonstrate that additional autonomy introduces operational costs that must be considered before clinical implementation. Several unresolved barriers identified across studies include limited interoperability with electronic health record systems, variations in clinical data distribution, insufficient longitudinal datasets, and the absence of standardized evaluation benchmarks. Eisenberg et al. (2026) and Jain et al. (2025) argued that clinical deployment requires evaluation frameworks capable of assessing not only accuracy but also reliability, safety, and practical integration within healthcare ecosystems.

The synthesis of collaboration mechanisms and clinical applications indicates that Agentic AI has demonstrated promising capabilities for supporting complex medical activities, but evidence for widespread clinical adoption remains limited. The current research trajectory suggests that future progress will depend on balancing autonomous reasoning capabilities with operational feasibility and clinical safety requirements. Multi-agent collaboration provides opportunities for improving medical AI performance, yet uncontrolled autonomy may introduce additional uncertainty if communication pathways and decision processes cannot be adequately monitored. The transition from experimental prototypes to clinical systems requires stronger validation methodologies, standardized benchmarks, and integration strategies that accommodate existing healthcare infrastructures. These findings reinforce the perspective that the success of Agentic AI in medicine will depend not only on technological advancement but also on its ability to satisfy clinical reliability and implementation requirements.

Ethical Governance, Implementation Barriers, and Future Research Directions

The thematic synthesis of ethical, qualitative, and policy-oriented studies identified governance as one of the most prominent concerns in the development of Agentic AI for medical applications. Among the 34 included studies, ethical and governance-related research represented the largest methodological category, accounting for 14 studies (41.2%), which exceeded the proportion of technical investigations and conceptual architecture studies. This distribution indicates that the advancement of autonomous medical AI has generated substantial normative concerns regarding responsibility, transparency, fairness, and human oversight. Unlike conventional clinical decision support systems that primarily assist healthcare professionals, Agentic AI introduces the possibility of autonomous planning, recommendation generation, and action execution, creating new challenges for accountability frameworks. Previous research has emphasized that increasing AI autonomy requires parallel development of governance mechanisms capable of maintaining trust and ensuring responsible deployment (Meszaros et al., 2022).

One of the central ethical challenges identified in the reviewed literature is the emergence of responsibility gaps in situations where autonomous systems contribute to clinical decisions.

Responsibility gaps occur when it becomes difficult to determine whether accountability should be assigned to developers, healthcare institutions, clinicians, or the AI system itself after an adverse outcome. Bleher et al. (2022) demonstrated that AI-driven clinical decision support systems can create uncertainty regarding responsibility attribution because decision processes are distributed across multiple actors and technological components. Lang et al. (2023) further argued that black-box characteristics of healthcare AI make complete responsibility assignment unrealistic, requiring a transition toward shared responsibilization models. This perspective suggests that future governance frameworks should focus not only on identifying individual responsibility but also on establishing coordinated accountability structures across the entire AI lifecycle.

Another significant issue concerns the increasing opacity produced by multi-agent architectures. Although individual agents may incorporate explainability mechanisms, interactions among multiple autonomous components can create a more complex phenomenon referred to as compound opacity. In such systems, the final output may represent the accumulation of several reasoning pathways, retrieval processes, and agent interactions that are difficult to reconstruct. Sara Salehi (2025) highlighted that layered opacity can reduce clinician confidence because understanding individual agent behavior does not necessarily explain the collective decision-making process. Xie et al. (2026) similarly emphasized that multi-agent systems introduce additional difficulties in auditing, error tracing, and accountability because mistakes may propagate between interconnected agents. These findings indicate that explainability approaches for Agentic AI must evolve beyond single-model interpretation toward mechanisms capable of representing interactions among autonomous entities.

Algorithmic bias and fairness represent additional governance challenges because autonomous agents may reproduce or amplify biases contained within training data, retrieval sources, or decision policies. The risk becomes more complex in multi-agent environments because biased outputs from one agent may influence the reasoning process of other agents during collaborative decision-making. Amadi et al. (2026) emphasized that trustworthy healthcare AI requires continuous monitoring of fairness, transparency, and reliability throughout system operation rather than relying solely on pre-deployment evaluation. Concerns regarding patient autonomy and informed consent have also emerged because patients may not fully understand how autonomous systems contribute to medical recommendations. Rueda et al. (2024) argued that procedural fairness in AI-based medical resource allocation requires explainability mechanisms to ensure that affected individuals can understand and evaluate decisions influencing their care.

The reviewed studies also highlighted several mitigation strategies designed to support safer Agentic AI implementation in healthcare. Human-in-the-loop mechanisms remain one of the most consistently proposed approaches, particularly through confidence thresholds that determine when autonomous recommendations require human verification. Thurzo et al. (2025) proposed risk-averse AI frameworks that prioritize safety constraints, while Pylarinou et al. (2026) incorporated confidence-guided human intervention mechanisms within surgical monitoring systems. Beyond technical safeguards, governance-oriented approaches emphasize the importance of multidisciplinary oversight involving clinicians, engineers, ethicists, and policymakers. Chow (2026) suggested that ethical generative AI systems should integrate autonomy modeling, transparency mechanisms, and human-AI cooperation principles from the earliest stages of system design.

Implementation barriers identified across the literature extend beyond ethical concerns and include technical, organizational, and educational limitations. The absence of standardized benchmarks remains a major obstacle because current Agentic AI systems are evaluated using heterogeneous metrics, datasets, and validation procedures. Jain et al. (2025) and Liu et al. (2026) highlighted that reliable assessment frameworks must consider multiple dimensions, including accuracy, computational efficiency, latency, safety, and clinical usefulness. Interoperability with existing healthcare infrastructure, particularly electronic health record systems, also represents a persistent challenge because autonomous agents require access to diverse and structured clinical information. Furthermore, limited AI literacy among healthcare professionals may hinder adoption because clinicians must understand both the capabilities and limitations of autonomous systems before integrating them into practice (Hanna et al., 2025).

Future research directions identified through the synthesis indicate that Agentic AI development should move beyond improving computational capability toward establishing comprehensive evaluation and governance ecosystems. Research priorities should include standardized

clinical benchmarks, longitudinal validation studies, interoperable architectures, and methods for quantifying the trade-off between autonomy and safety. The reviewed evidence suggests that successful implementation will require balancing technological advancement with institutional preparedness and ethical responsibility. Burgess et al. (2023) emphasized that clinician acceptance cannot be achieved solely through improvements in AI accuracy because trust depends on transparency, usability, and alignment with professional practices. The future trajectory of Agentic AI in medicine will therefore depend on integrating engineering innovation with governance frameworks that preserve accountability, patient autonomy, and clinical reliability.

CONCLUSION

This systematic review demonstrates that Agentic AI in medicine represents a fundamental transition from conventional AI systems that function primarily as passive analytical tools toward autonomous systems capable of planning, executing, and self-correcting complex tasks. The synthesis of 34 included studies indicates that multi-agent architectures supported by large language models and retrieval-augmented generation have become the dominant approach for developing autonomous healthcare systems. Although these architectures provide opportunities to improve clinical reasoning, workflow efficiency, and multimodal information processing, they also introduce new challenges related to coordination complexity, error propagation, transparency, and accountability. The evidence further shows that most current Agentic AI implementations remain at the proof-of-concept stage, with substantial gaps between simulated evaluations and real-world clinical validation. The limited availability of prospective clinical studies, standardized benchmarks, and interoperability strategies indicates that technological maturity has not yet been fully matched by implementation readiness. Therefore, the advancement of Agentic AI in medicine requires balanced development between computational innovation, clinical validation, and responsible governance frameworks.

The findings also reveal that ethical and governance considerations have become a central research priority, exceeding the proportion of studies focusing solely on technical development or conceptual architecture. This trend highlights that the critical question surrounding Agentic AI is no longer only whether autonomous systems can perform medical tasks, but how responsibility, supervision, and human control should be maintained when these systems influence clinical decisions. Healthcare institutions need evaluation frameworks that consider not only accuracy but also workflow compatibility, safety monitoring, and long-term reliability after implementation. Policymakers should develop adaptive regulatory approaches that correspond with different levels of AI autonomy, while researchers should prioritize external validation, evaluation methods balancing accuracy–cost–risk trade-offs, and AI literacy development among healthcare professionals. Despite the limitations related to heterogeneous evidence, limited primary studies, and the predominance of conceptual or simulation-based evaluations, this review provides a comprehensive foundation for understanding the technological and ethical trajectory of Agentic AI in medicine. Future success will depend less on maximizing autonomous capability alone and more on designing systems where autonomy is aligned with accountability, transparency, and meaningful human oversight.

REFERENCES

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access*, *13*, 18912–18936. https://doi.org/10.1109/ACCESS.2025.3532853

AlZu'Bi, S., Almiani, M., & Jararweh, Y. (2025). Agentic AI for Healthcare: Solutions to Intelligent Patient Care. *2025 12th International Conference on Information Technology (ICIT)*, 33–38. https://doi.org/10.1109/ICIT64950.2025.11049267

Andreea-Maria Tanasă Simona-Vasilica Oprea, A. B. (2026). Designing an Architecture of a Multi-Agentic AI-Powered Virtual Assistant Using LLMs and RAG for a Medical Clinic. *Electronics*, *15*, 334. https://doi.org/10.3390/electronics15020334

Andrej Thurzo, V. T. (2025). Embedding Fear in Medical AI: A Risk-Averse Framework for Safety and Ethics. *AI*, *6*, 101. https://doi.org/10.3390/ai6050101

Basile Njei Yazan A. Al-Ajlouni, U. (2026). Artificial intelligence agents in healthcare research: A scoping review. *PLOS One*, *21*, e0342182. https://doi.org/10.1371/journal.pone.0342182

Bernardo G. Collaco Syed Ali Haider, S. (2026). *The role of agentic artificial intelligence in healthcare: a scoping review*. *9*. https://doi.org/10.1038/s41746-026-02517-5

Chimaobi Amadi, A. O. (2026). *Building Trustworthy AI in Healthcare*. *14*. https://doi.org/10.1109/ACCESS.2025.3648410

Damien Gruson Bernard Gouget, W. (2026). From automation to agentic artificial intelligence in laboratory medicine: an opinion of the IFCC Division on Emerging Technologies. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *64*, 985–988. https://doi.org/10.1515/cclm-2025-1314

Dietrich, N. (2025). Agentic AI in radiology: emerging potential and unresolved challenges. *British Journal of Radiology*, *98*, 1582–1584. https://doi.org/10.1093/bjr/tqaf173

Dimitrios Kalathas Andreas Menychtas, I. (2025). *Incorporating Medical Assistants in eHealth Environments Using an Agentic RAG Approach*. *758*. https://doi.org/10.1007/978-3-031-96235-6_12

Dimitrios Kalathas Andreas Menychtas, P. (2026). Leveraging MCP and Corrective RAG for Scalable and Interoperable Multi-Agent Healthcare Systems. *Electronics*, *15*, 888. https://doi.org/10.3390/electronics15040888

Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., Ikeuchi, K., Vo, H., Fei-Fei, L., & Gao, J. (2024). *Agent AI: Surveying the Horizons of Multimodal Interaction*. https://doi.org/10.48550/arXiv.2401.03568

Fabrizio Billi, S. A. B. (2026). The Application of Agentic Artificial Intelligence in Orthopaedics. *Journal of Bone and Joint Surgery*, *108*, 278–282. https://doi.org/10.2106/JBJS.25.01497

Felix C. Oettl James A. Pruneski, B. (2026). Small language models: The big play for agentic artificial intelligence in orthopaedics. *Knee Surgery, Sports Traumatology, Arthroscopy*, *34*, 398–401. https://doi.org/10.1002/ksa.70126

Gerasimos Grammenos Aristidis G. Vrahatis, K. (2026). AI agents in Alzheimer's disease management: challenges and future directions. *Frontiers in Aging Neuroscience*, *17*. https://doi.org/10.3389/fnagi.2025.1735892

Hannah Bleher, M. B. (2022). Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, *2*, 747–761. https://doi.org/10.1007/s43681-022-00135-x

James C. L. Chow, K. L. (2026). From Dialogue Systems to Autonomous Agents: A Modeling Framework for Ethical Generative AI in Healthcare. *Information*, *17*, 361. https://doi.org/10.3390/info17040361

James Zou, E. J. T. (2025). The rise of agentic AI teammates in medicine. *The Lancet*, *405*, 457. https://doi.org/10.1016/S0140-6736(25)00202-8

Janos Meszaros Jusaku Minari, I. H. (2022). The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.927721

John J Hanna, R. J. M. (2025). The Infectious Diseases Orchestrator: Embracing AI Literacy in the Agentic Era. *Open Forum Infectious Diseases*, *13*. https://doi.org/10.1093/ofid/ofaf794

Jon Rueda Janet Delgado Rodríguez, I. (2024). "Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. *AI & SOCIETY*, *39*, 1411–1422. https://doi.org/10.1007/s00146-022-01614-9

Joshua J. Woo Andrew J. Yang, R. (2025). Custom Large Language Models Improve Accuracy: Comparing Retrieval Augmented Generation and Artificial Intelligence Agents to Noncustom Models for Evidence-Based Medicine. *Arthroscopy*, *41*, 565. https://doi.org/10.1016/j.arthro.2024.10.042

Juexiao Zhou Jindong Jiang, Z. (2025). Streamline automated biomedical discoveries with agentic bioinformatics. *Briefings in Bioinformatics*, *26*. https://doi.org/10.1093/bib/bbaf505

Katherine W Eisenberg, K. K. P. (2026). Governing autonomous AI in clinical care: the window is narrowing. *Health Affairs Scholar*, *4*. https://doi.org/10.1093/haschl/qxag141

Lang, B. H., Nyholm, S., & Blumenthal-Barby, J. (2023). Responsibility gaps and black box healthcare AI: Shared responsibilization as a solution. *Digital Society*, *2*(3), 52. https://doi.org/10.1007/s44206-023-00073-z

Md Asadul Islam Subbulakshmi Somu, F. M. F. A. (2026). The Rise of Agentic AI: Synthesis of Current Knowledge and Future Research Agenda. *Global Business and Organizational Excellence*, *45*, 402–416. https://doi.org/10.1002/joe.70019

Mohamad Abou Ali Fadi Dornaika, J. C. (2025). Agentic AI: a comprehensive survey of architectures, applications, and future directions. *Artificial Intelligence Review*, *59*. https://doi.org/10.1007/s10462-025-11422-4

Sara Salehi Yashbir Singh, K. (2025). Agentic AI and Large Language Models in Radiology: Opportunities and Hallucination Challenges. *Bioengineering*, *12*, 1303. https://doi.org/10.3390/bioengineering12121303

Sara Salehi Yashbir Singh, P. (2025). Beyond Single Systems: How Multi-Agent AI Is Reshaping Ethics in Radiology. *Bioengineering*, *12*, 1100. https://doi.org/10.3390/bioengineering12101100

Sarah Lebovitz Hila Lifshitz-Assaf, N. L. (2022). To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science*, *33*, 126–148. https://doi.org/10.1287/orsc.2021.1549

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. In *Frontiers of Computer Science* (Vol. 18, Number 6). Higher Education Press Limited Company. https://doi.org/10.1007/s11704-024-40231-1

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X. (2025). The rise and potential of large language model based agents: A survey. *SCIENCE CHINA Information Sciences*, *68*(2), 121101. https://doi.org/10.1007/s11432-024-4222-0

Yunsong Liu Zunamys I. Carrero, X. (2026). Benchmarking large language model-based agent systems for clinical decision tasks. *Npj Digital Medicine*, *9*. https://doi.org/10.1038/s41746-026-02443-6

Yushi Feng Junye Du, Y. (2026). PASS: Probabilistic Agentic Supernet Sampling for Interpretable and Adaptive Chest X-Ray Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*, 30717–30725. https://doi.org/10.1609/aaai.v40i36.40328

Zhibin Xie Hongyu Wang, L. (2026). Ethical issues in multi-agent AI systems for healthcare: a narrative review. *Frontiers in Public Health*, *14*. https://doi.org/10.3389/fpubh.2026.1792627