

Agentic AI Solutions: The Future for Intelligent and Secure Healthcare: A Review

Sonawane Shubhangi Chintaman¹, Bangar Abhishek Papat², Bhosale Sachin Bajirao³

¹ Department of Computer Engineering / Jaihind College of Engineering / SPPU, Pune, India

² Department of Computer Engineering / Jaihind College of Engineering / SPPU, Pune, India

³ Department of Computer Engineering / Jaihind College of Engineering / SPPU, Pune, India

Abstract — Healthcare organisations are exploring AI systems that do more than answer questions: they plan multi-step tasks, retrieve evidence, call software tools and coordinate with other agents. This capability promises support for clinical decision-making and administration, yet it widens the security and safety exposure of systems that handle highly sensitive patient data. This paper presents a structured narrative review of five verified IEEE journal articles published in 2025–2026. They cover the foundations of agentic AI, a taxonomy-based survey of healthcare agents, a multi-agent clinical decision support system, a multi-agent framework for Alzheimer's disease analysis and a privacy-preserving language-model framework for medical consultation. The papers are compared on objectives, methods, contributions, limitations and relevance to secure healthcare. The evidence indicates that retrieval-grounded, multi-agent designs are the dominant pattern, whereas human-in-the-loop control, adversarial testing, drift monitoring and regulatory compliance are rarely implemented. In the parts of the sources that could be read, no prospective clinical outcome trial is reported. The paper contributes a thematic synthesis, a five-study comparison, a conceptual layered architecture assembled from safeguards found in the literature, and a research agenda. The review is deliberately narrow and makes no claim of systematic coverage.

Keywords— Agentic AI; Healthcare; Large Language Models; Multi-Agent Systems; Clinical Decision Support; Privacy; Cybersecurity

I. INTRODUCTION

Artificial intelligence has long supported healthcare through single-purpose models that classify an image, score a risk or flag an abnormal value. A newer class of systems is built around large language models (LLMs) and behaves differently: it receives a goal, decides which steps are needed, consults sources, calls software tools and adjusts its plan as results arrive. This style of system is usually called agentic AI. A general survey describes it as autonomous AI that pursues complex goals with minimal human intervention and adapts to changing environments, in contrast with conventional systems that depend on structured instructions and close oversight [1]. In healthcare, a recent taxonomy-based survey reports LLM-based agents applied to tasks ranging from electronic health record (EHR) analysis and differential diagnosis to treatment planning and research workflows [2].

The attraction is clear. Clinicians face growing volumes of heterogeneous data, and administrative work consumes time that could be spent with patients. Agents promise to retrieve evidence, summarise records and coordinate routine steps. The difficulty is that the properties that make an agent useful, namely access to records, the ability to call tools and the freedom to act without approval at every step, also enlarge the

attack surface and the consequences of error. A chatbot that answers incorrectly can be ignored; an agent that queries a record system or drafts a chart entry on its own initiative cannot be treated so casually. Recent studies illustrate both sides: multi-agent clinical decision support [3], an agentic framework for multimodal Alzheimer's disease analysis [4] and a privacy-preserving framework for medical consultation [5].

This review aims to (i) explain the concept and architecture of healthcare agents, (ii) compare five verified IEEE journal papers critically, (iii) examine security and privacy risks together with candidate safeguards, and (iv) identify gaps that should be closed before wider deployment. Three kinds of statement are kept apart throughout: findings reported by the sources, proposals made by their authors, and our own synthesis, which is labelled as such and must not be read as a result of the cited work.

II REVIEW METHODOLOGY

The review is a narrow, structured narrative review, not a systematic one. Candidate papers were located with a general web search tool and then checked against publisher pages (IEEE Xplore and DOI records), indexing services (DBLP, DOAJ) and institutional or author repositories. Search phrases

combined agentic AI, LLM agents, multi-agent, clinical decision support, retrieval-augmented generation, healthcare, privacy-preserving language models, prompt injection, access control and electronic health records. IEEE Xplore, Scopus and Web of Science were not queried through their own search interfaces and no search log with record counts was kept, so no screening numbers are reported. Publications from 2022 to 2026 were preferred; the five selected were published in 2025–2026.

A paper was included if it was an IEEE journal article, addressed agentic or multi-agent AI in healthcare or a closely related foundation (concept, privacy), and had a DOI confirmed by at least one independent page. Conference papers were excluded even when relevant (an IEEE student-conference survey of healthcare AI agents was found and set aside for this reason), as were arXiv-only preprints, theses and non-IEEE journals. No base paper was supplied, so none is designated. Access to the papers differed, and claims are limited accordingly. The full text of the author-accepted manuscript of [2] was read. For [4], the abstract and portions of the full text were read. For [1], [3] and [5], only abstracts, metadata and short text fragments were accessible, so statements about them are limited to what those parts support, and the numerical results of [3] are not reported. Papers were compared on objective, method, contribution, limitation and relevance to secure healthcare, and then synthesised by theme. Only one of the five papers, [5], centres on privacy, which limits the security evidence base. The volume and page numbers of [2] and [5] could not be confirmed.

III LITERATURE REVIEW

1. Meaning and Characteristics of Agentic AI

Acharya et al. survey agentic AI as a general paradigm rather than a healthcare-specific one [1]. Their abstract attributes to agentic systems adaptability, advanced decision-making and self-sufficiency in evolving environments, and it treats ethical challenges by proposing approaches to goal alignment, resource constraints and environmental adaptability, while calling for further ethics research [1]. Vatsal et al. narrow the notion for medicine. They study systems in which an LLM is the main controller of planning and decisions and may use tools, keep memory and take actions, and they exclude purely symbolic or rule-based agents [2].

Synthesis: the two definitions agree on goal-directed, multi-step behaviour with limited step-by-step supervision, but they differ in precision. The first gives a conceptual picture; the second gives an operational criterion that can be checked

against a real system. For clinical use we interpret autonomy as a bounded spectrum. Retrieval and drafting may be automated, while orders, prescriptions and record changes should require human sign-off. This interpretation is ours and is not a finding of [1] or [2].

2. Architecture and Working of Healthcare AI Agents

Vatsal et al. organise healthcare agents into seven dimensions: cognitive capabilities, knowledge management, interaction patterns, adaptation and learning, safety and ethics, framework typology, and core tasks [2]. Cognitive capabilities include planning, perception of heterogeneous inputs, action through tools or application programming interfaces (APIs), meta-capabilities such as self-critique and deferral to clinicians, and the reconciliation of conflicting evidence. Knowledge management covers retrieval of external knowledge, memory modules that combine short-term buffers with longer-term stores, and updating or forgetting outdated knowledge. Interaction patterns include conversation, event-triggered activation, human-in-the-loop control and error recovery [2]. Read together, these dimensions suggest a working loop (synthesis). The agent receives a clinical request or event, converts the input into a usable representation, drafts a plan, retrieves evidence or calls a tool, checks the result, and then either proposes an output for review or performs a permitted action, recording each step. On collaboration, [2] reports that role-based multi-agent design is the prevailing pattern, with separate agents for functions such as retrieval, verification and synthesis, usually coordinated by an orchestrator that can enforce checks and route uncertain cases to clinicians. Multi-agent design was fully implemented in about 82% of the 49 studies, whereas central orchestration was mostly only partial, at about 57% [2].

3. Applications in Clinical and Administrative Workflows

The 49 studies analysed in [2] show that information-centred tasks are the most mature. Medical question answering and decision support were fully implemented in 28 studies (about 57%), documentation and EHR analysis in 22 (about 45%), and benchmarking or simulation in 39 (about 80%). Action-oriented tasks lag: treatment planning and prescription were fully implemented in only about 20% of studies, with about 59% not implementing them, and patient interaction and monitoring were fully implemented in 11 studies (about 22%) [2]. The survey also discusses workflow and administrative uses, such as prior-authorisation review, record coding and triage support [2].

Synthesis: clinical decision support is therefore the best-evidenced application, while hospital workflow automation,

patient monitoring and personalised treatment assistance are described mostly as individual task demonstrations or simulations. The surveyed work should not be read as proof of operational benefit in routine care.

4. Multi-Agent Decision Support and Disease Analysis

Öğdücü et al. address the difficulty clinicians face when growing data complexity slows accurate decisions. They propose a multi-layered, adaptive clinical decision support system (CDSS) composed of interacting LLM agents [3]. Fragments of the accepted manuscript indicate that retrieval from biomedical literature (a BioBERT embedding model with a Qdrant vector store) is handled separately from agents that gather evidence from the web [3]. Quantitative results were not read from the primary text and are not repeated here. Critical note (synthesis): web-sourced content is untrusted input, and any design that passes web results to an agent that can act should treat it as a route for prompt injection. We could not determine from the available text whether [3] addresses this.

Huang et al. present ADAM-1, a multi-agent reasoning framework that integrates microbiome profiles, clinical data and external knowledge bases through retrieval-augmented generation (RAG) and chain-of-thought reasoning in modular agents, to support Alzheimer's disease classification [4]. The data came from five nursing-home sites in central Massachusetts. Compared with XGBoost, ADAM-1 is reported to achieve a significantly higher mean F1 score and considerably lower variance. The framework is limited to a binary task with two data modalities, and neuroimaging and progression prediction are left for future work [4]. Lower variance on a small dataset is encouraging but is not evidence of clinical utility. The framework does contextualise its findings with literature-based evidence, which supports traceability [4].

5. Privacy-Preserving Language Models for Sensitive Data

Javed et al. target the privacy barrier to LLM-based medical consultation, where users must disclose private health information. Their MediGuard framework combines adaptive information obfuscation, secure access protocols and auditing so that only non-sensitive information is processed while the semantic content needed for medical inference is preserved [5]. According to the abstract, tests on several medical question-answering datasets show better privacy protection and clinical accuracy than existing methods, even under strict privacy constraints [5].

Critical note (synthesis): only the abstract was available, and the framework as described is a consultation system rather than

an agent that calls tools. Applying it to agents that touch record systems is therefore an assumption, and question-answering benchmarks do not reproduce live EHR environments. Its three components, obfuscation, access protocols and auditing, nevertheless correspond closely to safeguards agents need. They also address a weakness reported in [2]: privacy-preserving mechanisms were fully implemented in about 18% of the surveyed agents, partly in about 28% and not at all in about 53%.

6. Security and Privacy Risks and Candidate Safeguards

Five risks are central (synthesis informed by the safety dimension of [2]). (a) Prompt injection: instructions hidden in user text, documents or retrieved content redirect the agent. The survey lists prompt injection among the attacks that agent defences must address and describes resistance to manipulation of EHR-linked tools as part of safety guardrails [2]. (b) Unauthorised tool use: an agent holding broad credentials may read or write beyond its task. (c) Sensitive data leakage through outputs, logs, prompts sent to external services or retrieval; [2] describes privacy controls as preventing exfiltration of patient data from EHR-linked tools, and [5] is motivated by the same concern. (d) Weak access control, for example an agent acting under a shared service account instead of the permissions of the requesting clinician (synthesis). (e) Unreliable outputs: hallucination and contradictory evidence. Conflict resolution was absent in about 61% of studies and self-monitoring or deferral in about 53% [2].

Safeguards found in or implied by the sources include the following. Least-privilege access corresponds to the role-scoped access discussed under regulatory constraints in [2]. Human oversight appears as confirmation gates, uncertainty-triggered pauses and rollback controls, yet human-in-the-loop control was fully implemented in only about 12% of studies and absent in about 86% [2]. Audit trails appear as provenance logging and rationale tracing in [2] and as auditing in [5]. Secure data handling appears as obfuscation in [5] and as de-identification and federated learning in [2]. Explainability appears as literature-linked evidence in [4]. Continuous evaluation includes red-teaming and drift monitoring, but drift detection was absent in about 98% of studies [2]. Overall, only 3 of 49 studies (about 6%) fully implemented safety guardrails, 14 did so partly and 32 not at all [2].

7. Limitations, Ethical Considerations and Implementation Challenges

Technical limits recur across the sources. Most agents stay at the prototype stage and are evaluated on benchmarks or simulations [2]; ADAM-1 is restricted to one binary task [4];

and knowledge updating is rarely handled, so agents may rely on outdated guidance [2]. On the ethical side, [1] stresses goal alignment and calls for more work on ethics, and [2] reports that bias and fairness were fully addressed in about 10% of studies and regulatory or compliance constraints in about 10%, with about 86% not addressed. Implementation challenges, in our synthesis, include integration with record standards, allocation of accountability when an agent contributes to a decision, and the cost and latency of multi-agent pipelines. The last point is flagged in [2] as outside its scope.

8. Synthesis

Three patterns emerge. First, retrieval-grounded multi-agent design is the common technical answer to reliability [2], [3], [4]. Second, safety, privacy, oversight and compliance mechanisms trail the capability engineering [2], and the one

privacy-focused paper [5] tests a consultation setting rather than a tool-using agent. Third, reported evidence comes from benchmarks, simulations or small retrospective data, not from trials of clinical outcomes [2], [4]. These patterns support a cautious reading: agentic AI is promising as supervised decision support but unproven as an autonomous clinical actor.

IV. COMPARATIVE ANALYSIS

Table 1 compares the five studies. Reported metrics are not directly comparable because the tasks, data and evaluation methods differ, and the sources differ in kind (survey, system and framework).

Table 1. Comparison of the reviewed IEEE journal papers.

Ref., year	Objective	Method / technology	Major reported findings	Limitations	Relevance to secure healthcare
[1], 2025	Survey foundations, characteristics and ethics of agentic AI	Literature survey; framework for safe integration	Autonomy, adaptability and goal pursuit; goal-alignment issues	General scope; abstract-level access; no clinical evaluation	Conceptual and ethical basis
[2], 2026	Map LLM-based healthcare agents with one taxonomy	Review of 49 studies; 7 dimensions, 29 sub-dimensions	Multi-agent design about 82%; guardrails 3/49; human-in-the-loop about 12%; drift detection about 98% absent	Interpretive rubric; prototypes only; coverage ends June 2025	Strongest evidence on safety and oversight maturity
[3], 2025	Support decisions under growing data complexity	Multi-layer adaptive CDSS; interacting LLM agents; biomedical RAG; web agents	Multi-agent LLM design reported to improve decision support (abstract level)	Results not read; web-evidence risk; no deployment evidence seen	Decision-support architecture with external evidence
[4], 2025	Classify Alzheimer's disease from microbiome and clinical data	Multi-agent LLM; RAG; chain-of-thought	Higher mean F1 and lower variance than XGBoost	Binary task; two modalities; five-site dataset	Traceable, evidence-linked output with limited data
[5], 2025	Protect private data in LLM consultation	Adaptive obfuscation; secure access; auditing	Reported gains in privacy and accuracy on medical QA datasets	Abstract only; not a tool-using agent	Privacy and audit controls transferable to agents

Agreements. The sources agree that agents should be grounded in external knowledge and that specialised roles improve on a single model [2], [3], [4]. They also agree, at least implicitly, that sensitive data needs protection beyond what the base model provides [2], [5].

Differences. [1] is conceptual and general, [2] is a meta-level survey, [3] and [4] are implementations, and [5] is a privacy

framework. Only [2] measures how often safeguards are present, which makes it the most informative for the security question but also the most dependent on its own rubric.

Strengths and weaknesses. [2] provides transparent selection criteria and reproducible labels; [4] offers a defined comparison against a baseline; [5] links privacy to auditing. The weaknesses are the thin evidence about real deployments, the

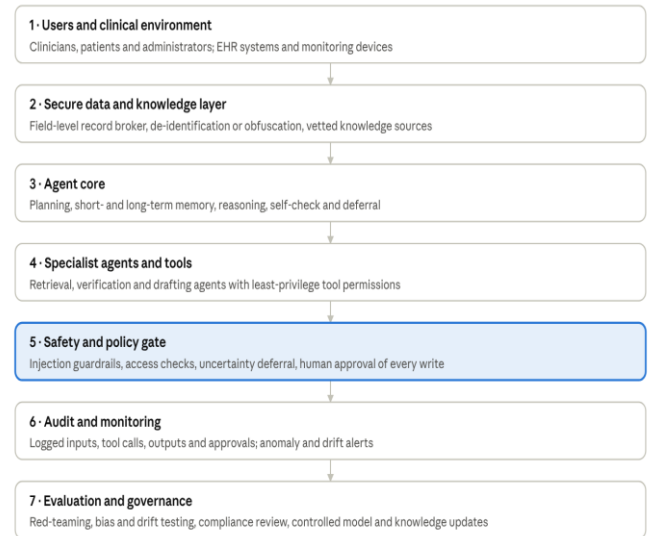
narrow tasks and the shortage of adversarial evaluation. Several of these judgements rest on abstracts only and should be re-checked against the full texts.

V. PROPOSED CONCEPTUAL ARCHITECTURE

Fig. 1 presents a conceptual architecture assembled from safeguards found in the reviewed literature. It has not been implemented or validated. Layer 1 contains the users and clinical environment. Layer 2 is a secure data and knowledge layer: a broker that exposes only the record fields a task needs, applies de-identification or obfuscation of the kind described in [5], and serves vetted knowledge sources for retrieval [2], [3]. Layer 3 is the agent core with planning, short- and long-term memory and reasoning [2]. Layer 4 holds specialist agents and tools for retrieval, verification and drafting, each running with least-privilege permissions [2], [3].

Layer 5 is a safety and policy gate between agent decisions and the outside world. It applies guardrails against injected instructions, checks access rights of the requesting user, defers to a clinician when uncertainty is high and requires human approval for any write action [2]. Layer 6 records an audit trail of inputs, tool calls and outputs [2], [5]. Layer 7 evaluates and governs the system through red-teaming, drift monitoring, bias testing and controlled updates [2]. The ordering reflects a design principle of ours: no agent output reaches a patient record or order without passing Layer 5.

No agent output reaches a patient record without passing the safety gate



conceptual layered architecture · 7 layers, 1 safety gate
Fig. 1. Proposed conceptual architecture for a secure healthcare agentic AI system (conceptual, not experimentally validated).

VI. RESEARCH GAPS AND FUTURE DIRECTIONS

Table 2 summarises the gaps supported by the reviewed evidence and the corresponding research directions.

Table 2. Research gaps and future directions.

Gap supported by evidence	Supporting studies	Future direction
Human-in-the-loop and error-recovery control rarely implemented	[2]	Standard approval gates, uncertainty thresholds and rollback, evaluated with clinicians
Little adversarial testing against prompt injection and unsafe tool use	[2]	Red-team suites for EHR-linked agents; treat web and retrieved content as untrusted
Privacy techniques tested on question answering, not agent workflows	[5], [2]	Test obfuscation and access control inside tool-using agents with leakage audits
Drift and knowledge-update mechanisms almost absent	[2]	Monitoring and controlled updating tied to guideline changes
Evidence from small or retrospective data; no outcome trials seen	[2], [4]	Prospective, multi-site studies with clinical and safety endpoints
Compliance and bias treated as policy statements	[2], [1]	Executable compliance checks; subgroup-stratified evaluation
Inconsistent definitions of agentic AI and autonomy	[1], [2]	Shared autonomy levels and reporting standards for healthcare agents

Further directions requiring investigation include clinician trust and workload studies, accountability rules for agent-assisted

decisions, and cost-benefit analysis of multi-agent pipelines. These items are our suggestions and are not claims made by the sources.

Health Inform., early access, Aug. 27, 2025, doi: 10.1109/JBHI.2025.3602983.

VII. CONCLUSION

Agentic AI offers a credible route to more intelligent healthcare support, mainly through retrieval-grounded, multi-agent decision support and analysis of complex data [2], [3], [4]. The reviewed evidence does not yet show that such systems are secure or clinically safe for autonomous use. Safety guardrails, human oversight, drift monitoring and compliance are rarely implemented [2], privacy techniques are tested mostly outside agent workflows [5], and the reported results come from benchmarks and small datasets. Deployment should therefore begin with supervised, least-privilege settings in which a clinician approves actions. Reliable progress will need adversarial testing, auditable operation, prospective clinical evaluation and clearer standards, as reflected in the proposed architecture. The conclusions are limited by the small, non-systematic sample of five papers and by abstract-level access to several of them.

REFERENCES

1. D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey," *IEEE Access*, vol. 13, pp. 18912–18936, 2025, doi: 10.1109/ACCESS.2025.3532853.
2. S. Vatsal, H. Dubey, and A. Singh, "Agentic AI in healthcare and medicine: A seven-dimensional taxonomy for empirical evaluation of LLM-based agents," *IEEE Access*, published Jan. 5, 2026, doi: 10.1109/ACCESS.2026.3651218. (Volume and page numbers not verified.)
3. Ç. U. Ögdü, K. Arslanoğlu, and M. Karaköse, "An adaptive multi-agent LLM-based clinical decision support system integrating biomedical RAG and web intelligence," *IEEE Access*, vol. 13, pp. 167390–167404, 2025, doi: 10.1109/ACCESS.2025.3613340.
4. Z. Huang, V. K. Sekhon, R. Sadeghian, M. L. Vaida, C. Jo, B. A. McCormick, D. V. Ward, V. Bucci, and J. P. Haran, "ADAM-1: An AI reasoning and bioinformatics model for Alzheimer's disease detection and microbiome-clinical data integration," *IEEE Access*, vol. 13, pp. 145953–145967, 2025, doi: 10.1109/ACCESS.2025.3599857.
5. H. Javed, F. Ali, B. Shah, N. Dilshad, and D. Kwak, "MediGuard: Protecting sensitive healthcare data with privacy-preserving language models," *IEEE J. Biomed.*