Journal of Computing Theories and Applications ISSN:3024-9104

Review Article

A Systematic Review of Agentic AI in Healthcare: An Evidence-Informed Seven-Principle Framework

Chandra Prakash *, Avneesh Sisodia, and Mary Lind

School of Computer Information Sciences, University of the Cumberlands, Williamsburg 40769, Kentucky, United States; e-mail cprakash@outlook.com; a.sisodia.k@gmail.com; marylind@gmail.com

* Corresponding Author : Chandra Prakash

Abstract: Agentic artificial intelligence (AI) systems capable of autonomous goal-directed behavior, multi-step planning, tool use, multi-agent coordination, and iterative self-correction represent a transition from passive clinical AI tools toward systems that can participate in complex healthcare workflows. However, empirical evidence remains fragmented across clinical decision support, patient monitoring, and administrative applications, and no systematic synthesis has evaluated which agentic principles have been technically demonstrated and which have accumulated sufficient evidence to support responsible clinical deployment. We conducted a PRISMA-informed systematic review of peer-reviewed empirical studies published between January 2025 and April 2026. Searches across five bibliographic databases and Google Scholar, supplemented by citation tracking, identified 443 unique records for screening, of which 25 met the predefined PICOS and quality appraisal criteria. Evidence was synthesized using an evidence-informed seven-principle framework derived from the integration of agentic AI, clinical AI, and healthcare governance literature. This framework provides a structured lens for examining how agentic principles are evaluated individually and in combination, enabling a deployment-readiness perspective that extends beyond capability-focused assessments alone. The evidence base was concentrated on technical capability principles, whereas human oversight, safety, compliance, and equity-related evaluation received comparatively limited attention. Most studies remained at the laboratory, benchmark, or proof-of-concept stage, and none reported demographic-stratified performance outcomes. Overall, the findings suggest a structural asymmetry in agentic healthcare AI: empirical research is advancing agentic capabilities more rapidly than it is generating evidence for the oversight, safety, equity, and governance mechanisms required for responsible clinical translation.

Keywords: Agentic AI; Autonomous agents; Clinical decision support; Digital health; Explainability; Human-in-the-loop; Multi-agent systems; Patient monitoring.

Received: May, 10th 2026
Revised: June, 7th 2026
Accepted: June, 11th 2026
Published: June, 21st 2026

1. Introduction

The healthcare sector is undergoing a profound transformation driven by artificial intelligence (AI) [1], [2]. Healthcare AI has evolved through several overlapping phases. Early systems were primarily rule-based, relying on expert-defined logic to support diagnosis, alert generation, and clinical decision support. While these systems remain important components of clinical workflows, they are inherently constrained by their rigidity and dependence on manually encoded knowledge [3]. Subsequently, machine learning and deep learning approaches expanded the capabilities of healthcare AI by learning patterns from medical imaging, structured electronic health record (EHR) data, claims data, and clinical text, enabling improved prediction and classification across a wide range of specialized tasks [4], [5]. More recently, generative AI and large language models (LLMs) introduced flexible natural-language interaction, information synthesis, and reasoning over unstructured clinical content [6]. Agentic AI represents a further evolution in this progression. Rather than merely generating predictions or responses, agentic systems can pursue goals, plan across multiple steps, use external tools, coordinate specialized agents, and adapt their actions based on feedback [7].

Early healthcare AI applications were predominantly narrow in scope, including rule-based clinical decision support systems (CDSS), image classification models, and natural language processing tools for clinical documentation [8]. Although effective within well-defined tasks, these systems shared a common limitation: they were fundamentally passive. They responded to queries or inputs but lacked the ability to initiate, plan, and execute actions autonomously across multi-step clinical workflows [9]. Agentic AI seeks to address this limitation. Unlike conventional AI models that typically generate single-turn outputs, agentic systems can decompose complex clinical objectives into intermediate tasks, invoke external tools such as EHR APIs and laboratory interfaces, coordinate specialized sub-agents, and iteratively refine their strategies based on intermediate feedback while operating under configurable levels of human supervision [9]–[11]. These capabilities position agentic systems as a distinct advancement beyond earlier generations of clinical AI. Recent empirical studies have begun to demonstrate these capabilities in practice. Ferber et al. [12] and Zhao et al. [7] reported the evaluation of autonomous agentic systems for oncology-related clinical decision support. Zhao et al. [7] introduced DeepRare, an agentic framework that achieved measurable performance on the MIMIC-IV-Rare benchmark for rare disease diagnosis. Liu et al. [13] proposed EvoMDT, a self-evolving multi-agent system designed to support cancer-related clinical decision-making. Collectively, these studies, published in leading venues, suggest an ongoing transition from conceptual discussions toward more rigorous empirical evaluation of agentic healthcare systems [14].

Despite this progress, the field still lacks a unified synthesis of the core agentic principles being evaluated across healthcare domains and, more importantly, of what the current empirical evidence supports regarding responsible clinical deployment. Several recent reviews have examined aspects of agentic AI in healthcare, each offering valuable yet relatively focused perspectives [15]. Collaco et al. [15] conducted a scoping review that identified seven eligible studies exhibiting agentic characteristics in clinical settings. However, their analysis primarily focused on application domains and conceptual distinctions between AI agents and agentic AI, without developing a principle-oriented synthesis or examining deployment-related governance considerations. Abou Ali et al. [16] conducted a systematic review of 90 peer-reviewed studies and organized the literature according to application areas and architectural paradigms, including reinforcement learning, multi-agent systems, and LLM-based agents. While this work provides a valuable taxonomic overview of existing implementations, it does not evaluate the empirical support for the principles that may underpin responsible clinical deployment. Similarly, Njei et al. [17] categorized 43 included studies into functional archetypes such as conversational agents, workflow assistants, and decision-support agents, while identifying mechanisms including tool use and self-correction. Although this functional classification advances understanding of agentic system design, it does not address which principles have accumulated sufficient empirical support for safe and effective clinical operation. Srinivasu et al. [18] reviewed the technical foundations of agentic AI, including federated learning and privacy-preserving architectures, providing detailed insights into implementation mechanisms but offering limited cross-domain synthesis of deployment-relevant principles. Finally, Hosseini and Seilani [19] examined agentic AI across multiple sectors, including healthcare; however, the breadth of their cross-sector perspective necessarily limited the depth of healthcare-specific analysis required to inform clinical translation and deployment decisions.

2. Literature Review and Research Gap

2.1. Prior Reviews of Agentic AI in Healthcare

The present review addresses a gap that remains only partially explored in prior reviews of agentic AI in healthcare (Table 1). Existing reviews have characterized agentic AI from the perspectives of applications, architectures, implementation mechanisms, or functional archetypes; however, they have not systematically evaluated empirical evidence through a healthcare-oriented principle framework designed to assess deployment readiness. By organizing evidence around seven empirically grounded principles rather than application domains, architectural paradigms, or functional categories, this review provides an evidence-informed analytical framework for examining the readiness of agentic AI systems for responsible clinical deployment.

This review synthesizes 25 empirical studies that met predefined PICOS eligibility criteria and a custom quality appraisal rubric informed by MMAT/CASP appraisal logic. Beyond identifying the presence of agentic capabilities, the review examines how principles co-occur, where evidence remains fragmented, and which deployment-critical dimensions receive limited empirical attention. The principle co-occurrence analysis (Figure 3) and domain-by-principle evidence matrix (Table 13) extend prior reviews by revealing not only which principles are evaluated, but also how they interact across healthcare domains and where important evidence gaps remain.

Table 1. Summary of prior reviews on agentic AI in healthcare

Ref.	Studies	Approach	Principle Framework	Co-occurrence Analysis	Primary Contribution / Gap Addressed
[15]	7	Scoping Review	No	No	Conceptual distinction between AI agents and agentic AI
[16]	90	Systematic Review	No	No	Architectural taxonomy of agentic AI paradigms
[17]	43	Scoping Review	No	No	Functional archetypes and agentic mechanisms
[18]	N/A	Narrative Review	No	No	Technical and mechanistic foundations of agentic AI
[19]	Multi-sector	Systematic Review	No	No	Cross-sector overview of agentic AI applications
Ours	25	PRISMA-Guided Systematic Review	Yes – Seven Principles	Yes	Evidence-informed framework for deployment-readiness assessment and empirical evidence synthesis

2.2. Contribution

Existing reviews are generally application-oriented, architecture-focused, or conceptually descriptive rather than explicitly centered on empirical evidence for deployment readiness. This review addresses that gap through a PRISMA 2020-guided systematic review of 443 screened records, resulting in a curated corpus of 25 empirical studies spanning Clinical Decision Support (CDS), Patient Monitoring and ICU environments, and Administrative Workflows.

This review contributes to the agentic AI and healthcare informatics literature in four ways. First, it synthesizes recent peer-reviewed empirical evidence on agentic AI systems across key healthcare domains. Second, it integrates established concepts from agentic AI, transparency, and healthcare governance literature into an evidence-informed seven-principle framework for deployment-readiness assessment. Third, it maps the empirical distribution and co-occurrence of these principles across included studies. Fourth, it identifies evidence gaps that should be addressed before agentic AI systems can be responsibly translated into routine clinical practice.

The framework's primary contribution lies in shifting the analytical focus from application mapping toward deployment-readiness assessment. Rather than asking only where agentic AI is being applied or which architectural paradigms are being adopted, the framework examines whether empirical studies address the broader set of technical, transparency, and governance principles required for responsible clinical use. In this way, the framework provides a structured lens for evaluating not only what agentic systems can do, but also whether sufficient evidence exists to support their safe and accountable deployment.

This review also differs from prior taxonomies and scoping reviews in its treatment of evidence. Studies were included only when they reported empirical evaluation, structured performance assessment, benchmark testing, clinical or workflow validation, or systematic coding of agentic system characteristics with analyzable outcomes. Purely conceptual frameworks, narrative taxonomies, opinion papers, and architectural proposals without empirical evaluation were excluded from the evidence synthesis, although they were used where appropriate to contextualize the field. Framework and taxonomy studies were retained only when

they incorporated a reproducible empirical component, such as benchmark comparisons, implementation evaluations, coded assessments, or reported outcome data.

This distinction enables the review to move beyond cataloging agentic AI concepts toward evaluating the current evidence base for deployment readiness. The contribution of this work does not lie in proposing entirely new principles; rather, it lies in integrating established principles into a structured, evidence-informed synthesis framework that makes deployment-readiness gaps visible across the emerging empirical literature.

3. Methodology

3.1. Search Strategy and Information Sources

The primary input to this review is a corpus of 443 records identified through systematic searches across multiple bibliographic sources. The search covered five major databases—IEEE Xplore, PubMed/MEDLINE, Scopus, Web of Science, and the ACM Digital Library—with an additional supplementary search conducted through Google Scholar. The review protocol was established prior to screening and specified the research questions, eligibility criteria, data extraction procedures, quality appraisal approach, and evidence synthesis strategy in accordance with PRISMA 2020 recommendations [20], [21]. The primary research question guiding the review is: RQ: What empirical evidence supports the deployment readiness of agentic AI in healthcare when assessed through capability-, transparency-, and governance-oriented principles?

The review scope was intentionally defined to align with its deployment-readiness objective. The January 2025–April 2026 publication window was selected because it represents the period during which peer-reviewed empirical evaluations of healthcare-focused agentic AI systems began to emerge in sufficient numbers for systematic analysis. The three selected domains—(a) Clinical Decision Support (CDS), (b) Patient Monitoring and ICU environments, and (c) Administrative Workflows—were chosen because they represent the principal healthcare settings in which agentic systems may influence clinical reasoning, patient surveillance, operational efficiency, and access to care. The review further adopted an empirical-only inclusion strategy because its objective is to evaluate evidence supporting deployment readiness rather than conceptual possibilities or theoretical system designs. Consequently, the findings should be interpreted as a synthesis of the current empirical evidence base rather than as a historical review of healthcare AI or a comprehensive taxonomy of agentic AI concepts.

This review was not prospectively registered in PROSPERO or another systematic-review registry. Nevertheless, the review protocol was developed before screening commenced and specified the research question, search strategy, eligibility criteria, screening process, quality appraisal rubric, data extraction fields, and narrative synthesis procedures. No major protocol deviations occurred after screening began, although the reporting of the search strategy, quality appraisal process, and exclusion rationale was expanded during revision to improve transparency and alignment with PRISMA reporting recommendations.

The search strategy employed a two-cluster Boolean structure applied consistently across all databases. Cluster 1 captured agentic AI concepts, while Cluster 2 captured healthcare-related terms. The two clusters were combined using the Boolean operator AND to restrict results to healthcare applications while maintaining broad coverage across variations in agentic AI terminology. All searches applied a publication-date filter spanning January 2025 to April 2026. Table 2 presents the complete search strategy, adapted where necessary to accommodate database-specific syntax, indexing structures, and field-tag conventions.

Table 3 summarizes retrieval counts by database, search execution dates, duplicate removal procedures, and cumulative record totals. Deduplication was performed in Rayyan using DOI-based matching followed by manual verification of title-author combinations where necessary. An additional 40 records were identified through backward and forward citation searching of included studies and relevant review articles, resulting in a final corpus of 443 records entering the PRISMA screening process. All searches were conducted between 15 and 16 April 2026 using a common Cluster 1 AND Cluster 2 search structure, with database-specific adaptations applied only where required by indexing and search-interface constraints. Cross-database deduplication was performed through DOI exact matching followed by manual review when ambiguity remained.

Table 2. Database-specific boolean search strings with field tags

Database	Cluster 1 — Agentic AI Constructs (joined internally by OR)	Cluster 2 — Healthcare Context Terms (joined internally by OR)	Field Tags Applied
PubMed / MEDLINE	"agentic AI"[Title/Abstract] OR "AI agents"[Title/Abstract] OR "autonomous agents"[Title/Abstract] OR "multi-agent system"[Title/Abstract] OR "LLM agent"[Title/Abstract] OR "AI agent orchestration"[Title/Abstract] OR "intelligent agents"[Title/Abstract]	"clinical decision support"[Title/Abstract] OR "intensive care unit"[Title/Abstract] OR "patient monitoring"[Title/Abstract] OR "healthcare administration"[Title/Abstract] OR "electronic health records"[Title/Abstract] OR "hospital workflows"[Title/Abstract] OR "clinical AI"[Title/Abstract]	*Title/ Abstract [Title/Abstract]*
IEEE Xplore	"agentic AI" OR "AI agents" OR "autonomous agents" OR "multi-agent system" OR "LLM agent" OR "AI agent orchestration" OR "intelligent agents"	"clinical decision support" OR "intensive care unit" OR "patient monitoring" OR "healthcare administration" OR "electronic health records" OR "hospital workflows" OR "clinical AI"	*Full Text & Metadata*
Scopus	TITLE-ABS-KEY("agentic AI" OR "AI agents" OR "autonomous agents" OR "multi-agent system" OR "LLM agent" OR "AI agent orchestration" OR "intelligent agents")	TITLE-ABS-KEY("clinical decision support" OR "intensive care unit" OR "patient monitoring" OR "healthcare administration" OR "electronic health records" OR "hospital workflows" OR "clinical AI")	*TITLE-ABS-KEY*
Web of Science	TS=("agentic AI" OR "AI agents" OR "autonomous agents" OR "multi-agent system" OR "LLM agent" OR "AI agent orchestration" OR "intelligent agents")	TS=("clinical decision support" OR "intensive care unit" OR "patient monitoring" OR "healthcare administration" OR "electronic health records" OR "hospital workflows" OR "clinical AI")	*TS (Topic: title, abstract, keywords, keyword plus)*
ACM Digital Library	Title:("agentic AI" OR "AI agents" OR "autonomous agents" OR "multi-agent system" OR "LLM agent") OR Abstract:("agentic AI" OR "AI agents" OR "autonomous agents")	Title:("clinical decision support" OR "patient monitoring" OR "electronic health records" OR "hospital workflows" OR "clinical AI") OR Abstract:("clinical AI" OR "healthcare" OR "clinical decision")	*Title + Abstract*
Google Scholar (Supplementary)	"agentic AI" OR "AI agents" OR "autonomous agents" OR "multi-agent system" OR "LLM agent"	"clinical decision support" OR "patient monitoring" OR "electronic health records" OR "healthcare" OR "clinical AI"	*All fields (title-weighted); first 20 pages harvested*

Note: Google Scholar does not support field-limited search via API; results were harvested by manual review of the first 20 result pages per query. The primary cluster terms were adapted to each database's controlled vocabulary where applicable (e.g., MeSH mapping in PubMed was disabled to preserve term specificity given the novelty of agentic AI terminology).

Table 3. Per-Database Retrieval Counts, Duplicate Removal, Citation-Search Additions, and Final Records Screened

Database	Search Execution Date	Raw Records Retrieved	Duplicates Removed (cross-database)	Records After Deduplication	Notes
PubMed / MEDLINE	15 April 2026	58	6	52	Date filter: Jan 2025 – April 2026; MeSH auto-mapping disabled to preserve term specificity; language filter: English.
IEEE Xplore	15 April 2026	74	3	71	Conference proceedings and journal articles included; early-access articles included if DOI-registered.
Scopus	15 April 2026	102	7	95	Subject area filter: Medicine + Computer Science; document type: Article, Conference Paper, Review.
Web of Science	15 April 2026	30	4	26	Web of Science Core Collection; document types: Article, Proceedings Paper; English only
ACM Digital Library	15 April 2026	20	2	18	ACM Full-Text Collection; conference and journal articles; English only.
Google Scholar (Suppl.)	16 April 2026	172	31	141	First 20 result pages per query variant; harvested manually; high duplicate overlap with above databases; included to identify potentially relevant records not indexed in the selected bibliographic databases; records were included in the synthesis only if peer-review status could be verified.
Citation searching	-	40		40	An additional 40 records identified through citation searching of included studies and reference lists of relevant systematic reviews.
TOTAL		496	53	443	Grand total after all deduplication: 443 records entered into PRISMA screening.

Note: Searches executed 15–16 April 2026. All databases searched with identical Cluster 1 AND Cluster 2 logic; field tags adapted per database. Cross-database deduplication performed using DOI exact-match.

3.2. Eligibility Criteria (PICOS)

The eligibility criteria were operationalized using the PICOS framework as follows:

- Population: Healthcare systems, healthcare institutions, clinical personnel, or patient populations operating within Clinical Decision Support (CDS), Patient Monitoring/ICU, or Administrative Workflow settings.
- Intervention: Included systems were required to demonstrate autonomous or semi-autonomous goal-directed behavior, defined as agent actions performed without continuous manual human intervention, together with at least one additional agentic capability such as multi-step planning, tool use, memory or state maintenance, self-correction, or multi-agent coordination.
- Comparison: Comparator evidence was extracted when available, including comparisons with rule-based systems, single-agent systems, generative AI systems without agentic capabilities, or human-only baselines.
- Outcomes: Studies were required to report empirical outcomes, either quantitative or qualitative, such as diagnostic accuracy, operational efficiency metrics, safety-related outcomes, clinician acceptance measures, regulatory compliance assessments, or validated benchmark performance with reported results.
- Study Design: Only peer-reviewed empirical studies were included, encompassing experimental evaluations, pilot studies, cohort analyses, benchmark evaluations, proof-of-concept systems with quantitative assessment, and mixed-methods studies containing empirical evidence. Narrative reviews, editorials, opinion papers, perspective articles, and theoretical frameworks without evaluation data were excluded.

Additional exclusion criteria included publications outside the January 2025–April 2026 review window, studies outside the three target healthcare domains, publications lacking assessable empirical outcomes, non-English-language publications, grey literature without verifiable peer-review status, and healthcare-tangential applications (e.g., agricultural AI, 6G networking studies without a clinical deployment context, or educational MCQ-generation systems). Table 4 summarizes the inclusion and exclusion criteria applied throughout the screening and review process.

Table 4. Inclusion and Exclusion Criteria

Inclusion	Exclusion
Peer-reviewed empirical studies published between January 1, 2025, and April 15, 2026	Editorials, viewpoints, commentaries, frameworks, and purely conceptual papers
Healthcare-relevant tasks, settings, users, or workflows	Reviews, protocols, and opinion pieces without primary empirical data
Studies evaluating an agentic mechanism such as autonomous planning, tool use, memory, self-correction, or multi-agent coordination. Taxonomy-oriented studies were included only when they contained a reproducible empirical coding or assessment component.	Studies on general-purpose LLMs with no agentic mechanism
Studies reporting explicit outcome data or structured evaluation.	Non-healthcare studies
Full-text English-language articles	Conference abstracts without sufficient methodological detail

3.3. Screening and PRISMA Flow

Records were screened in two sequential stages using the predefined PICOS criteria and the operational definition of agentic goal-directed behavior. In Stage 1, titles and abstracts were evaluated against the healthcare-domain and empirical-study requirements. In Stage 2, full-text articles were assessed for eligibility, outcome validity, and methodological quality using the custom appraisal rubric described in Section 3.4. Cases involving uncertainty or borderline eligibility were reviewed against the predefined inclusion and exclusion criteria, with final inclusion decisions resolved through author consensus. Figure 1 summarizes the study

identification, screening, eligibility assessment, and inclusion process following the PRISMA 2020 framework.

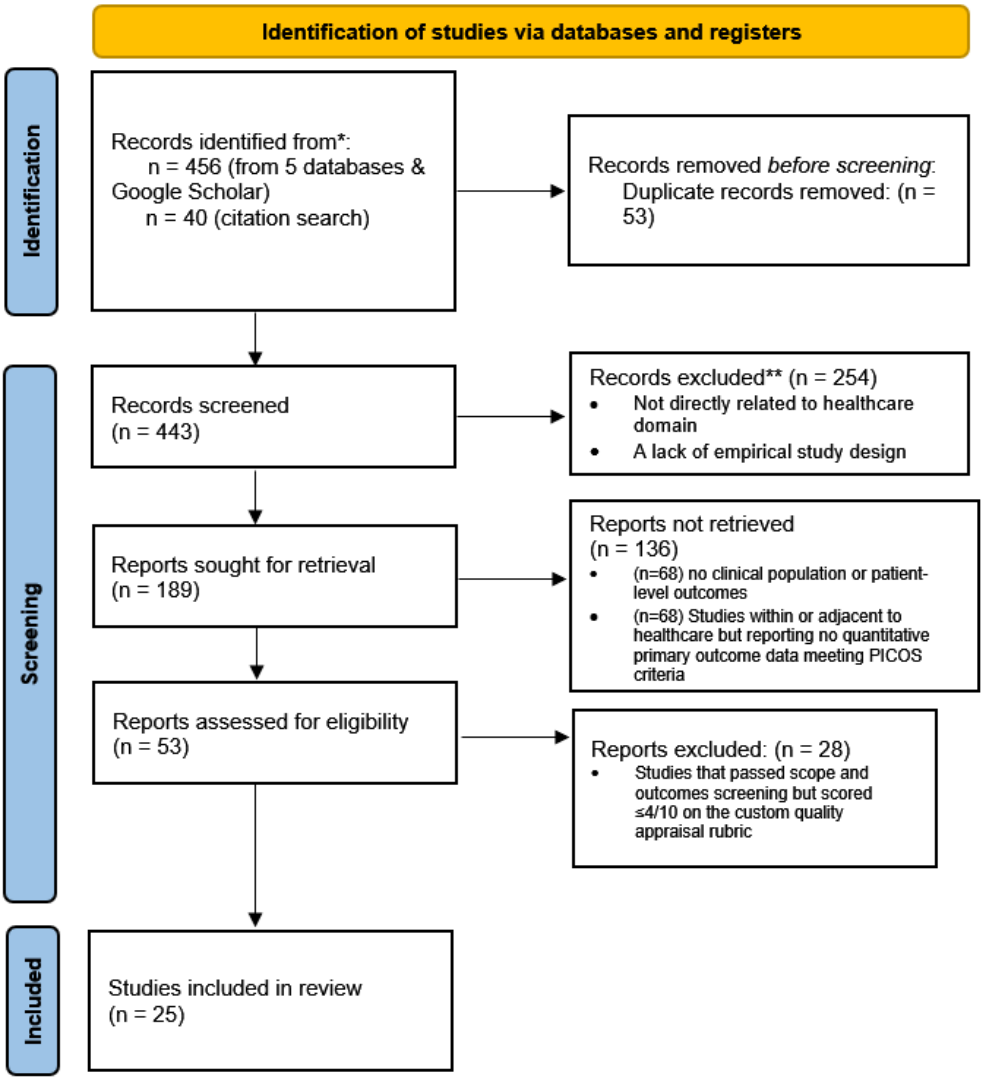

Figure 1. PRISMA 2020 flow diagram of study identification, screening, eligibility assessment, and final inclusion.

3.4. Quality Assessment

Methodological quality was assessed using a custom five-criterion appraisal rubric developed specifically for this review. A tailored rubric was considered necessary because the included studies were methodologically heterogeneous, encompassing benchmark evaluations, proof-of-concept systems, cohort analyses, pilot deployments, workflow evaluations, and mixed-methods investigations. No single established appraisal instrument, including MMAT or CASP, fully accommodates the diversity of study designs represented in the agentic AI literature, as many studies do not conform to conventional clinical-trial, qualitative, or mixed-methods research paradigms. The rubric was therefore informed by the appraisal logic of MMAT and CASP but was not implemented as a formal MMAT or CASP instrument. Quality categories were defined as follows:

- 8–10 points (High Quality): The study satisfies most appraisal criteria with no major methodological weaknesses.
- 5–7 points (Moderate Quality): The study provides sufficient empirical evidence to inform synthesis but contains limitations such as small sample sizes, lack of comparator groups, synthetic data, or restricted generalizability.

- 0–4 points (Low Quality): Methodological limitations are sufficiently substantial to reduce confidence in the evidence and preclude inclusion in the synthesis.

Each study received a score ranging from 0 to 2 across five criteria:

1. Clarity of the research objective,
2. Appropriateness of the study design,
3. Adequacy of evaluation data,
4. Outcome validity and clinical relevance,
5. Transparency regarding limitations.

The maximum possible score was 10. Studies scoring 8–10 were classified as high quality and retained for direct evidence synthesis. Studies scoring 5–7 were classified as moderate quality and retained with their limitations explicitly considered during interpretation. Studies scoring 0–4 were classified as low quality and excluded from synthesis. Quality assessment was conducted using the predefined rubric, and borderline cases were reviewed against the operational definitions of the scoring criteria before final classification. Article-level appraisal scores are reported in Supplementary Appendix B.

3.5. Data Extraction

Data were extracted using a predefined extraction form, with the complete extraction fields provided in Supplementary Appendix A. The extraction form captured bibliographic information, source database, screening decisions, exclusion reasons, healthcare domain, target population or user group, agentic mechanisms, study design, comparator type (where applicable), primary outcomes, quality-assessment information, and synthesis classifications. Data extraction was conducted by the lead author and subsequently verified against the full text during synthesis to ensure consistency with the eligibility criteria and the evidence-informed seven-principle framework. Following pilot use, the extraction form was refined to better capture both conventional empirical outcomes and agentic AI-specific characteristics, including tool use, planning, multi-agent coordination, memory or state maintenance, self-correction, human-in-the-loop (HITL) oversight, explainability, and safety/compliance evidence. Missing numerical outcomes were not imputed. When outcome information was reported only narratively or lacked sufficient quantitative detail, the evidence was coded as descriptive and interpreted cautiously during synthesis. Studies without assessable empirical outcomes were excluded from the final evidence synthesis.

3.6 Synthesis Approach

Given the heterogeneity of the included studies, evidence was synthesized narratively rather than through meta-analysis. The studies varied substantially in clinical settings, agent architectures, evaluation designs, datasets, and reported outcomes, making statistical pooling inappropriate. Evidence was organized using the evidence-informed seven-principle framework and compared across healthcare domains, evaluation settings, and outcome categories.

Principle-level coding was based on explicit evidence reported within each study's methods, system architecture, evaluation design, and results sections. A principle was coded as present only when the study described and evaluated a corresponding capability, design feature, or governance mechanism. For example, tool use required evidence of interaction with external resources or systems; multi-agent collaboration required coordination among multiple agents; human-in-the-loop (HITL) oversight required a defined role for human review, approval, escalation, or intervention; and safety, alignment, and compliance required explicit evaluation or structured discussion of safety controls, privacy protection, auditability, regulatory considerations, or alignment mechanisms. References to a principle appearing only in background, discussion, or future-work sections were not considered sufficient evidence for principle-level coding.

When multiple publications appeared to describe the same underlying system or study, the most complete peer-reviewed version was retained as the primary record and overlapping reports were not double-counted. Where a subsequent publication extended an earlier system through a distinct evaluation, it was treated as a separate study only when it reported new empirical outcomes or a clearly different evaluation context.

Principle-level observations were aggregated to calculate frequency and co-occurrence patterns across the 25 included studies. Coding was conducted using the predefined

framework and cross-checked against extracted study characteristics during synthesis. Formal independent duplicate coding and inter-coder reliability statistics were not calculated and are acknowledged as a limitation of the review. Consequently, principle-level coding should be interpreted as a structured evidence-synthesis exercise rather than a formal content-analysis procedure. To enhance transparency, the coding rules, extracted variables, study-level quality scores, and principle-frequency results are reported throughout the manuscript and supplementary materials.

4. Evidence-Informed Seven-Principle Synthesis Framework

The evidence-informed seven-principle framework used in this review was developed deductively during protocol design, prior to study screening, to provide a consistent analytical basis for evaluating empirical evidence on agentic AI in healthcare. The framework was derived through triangulation across three complementary sources: foundational agent theory, emerging LLM-agent literature, and healthcare-specific regulatory and ethical guidance. This approach enables the evaluation of agentic AI not only as a technical architecture but also as a clinically governed system requiring transparency, oversight, accountability, and safety safeguards.

Foundational research on intelligent agents emphasizes autonomy, goal-directed behavior, environmental interaction, and adaptive decision-making as defining characteristics of agentic systems [22], [23]. More recent LLM-agent research, including ReAct [22], Toolformer [24], and multi-agent debate frameworks [25], extends these concepts by demonstrating how language models can reason across multiple steps, use external tools, collaborate with other agents, and revise outputs based on feedback. Healthcare-specific guidance introduces an additional requirement: agentic systems operating in clinical environments must also be explainable, governable, safe, and accountable [26]–[28]. The resulting framework organizes these requirements into three complementary dimensions: technical capability, transparency, and governance. Together, these dimensions provide a structured lens for evaluating not only what agentic systems can do, but also whether sufficient evidence exists to support their responsible deployment in high-stakes healthcare settings.

4.1. Technical Capability Principles

The first dimension of the framework captures the technical capabilities that distinguish agentic AI from conventional AI systems. Foundational work on rational agents and autonomous systems identifies autonomy, goal-directed behavior, environmental interaction, and adaptive decision-making as core characteristics of agentic behavior [22], [23]. More recent LLM-agent research, including ReAct [22], Toolformer [24], and multi-agent debate frameworks [25], further demonstrates how language models can reason across multiple steps, invoke external tools, coordinate specialized agents, and refine outputs based on intermediate feedback.

Drawing from this literature, four capability-oriented principles were retained. Autonomy and goal-directed behavior refer to the ability of a system to pursue defined objectives without requiring human initiation at every stage. Planning and multi-step reasoning describe the ability to decompose complex clinical objectives into sequenced actions. Tool use and orchestration refer to interaction with external resources such as EHR systems, laboratory information systems, imaging tools, guideline repositories, and clinical calculators. Multi-agent collaboration captures the use of multiple specialized agents that divide, coordinate, validate, or debate tasks to support decision-making [22]. Iterative self-correction was recognized as an important agentic capability but was not retained as a standalone principle. Instead, it was coded within planning, reasoning, and multi-agent validation because it typically functions as an implementation mechanism embedded within these broader capabilities rather than as an independent principle.

4.2. Transparency Principle

Explainability and transparency were elevated to a standalone principle because their role in healthcare differs fundamentally from their role in many non-clinical AI applications. Within clinical environments, explanation is not merely a desirable interface feature; it is closely linked to clinician accountability, informed decision-making, patient safety, and medicolegal defensibility. Clinical AI literature consistently treats interpretability, auditability, and

transparent reasoning as foundational requirements for responsible deployment rather than optional enhancements [29], [30].

Accordingly, explainability and transparency are defined as the ability of an agentic system to provide clinically meaningful, inspectable, and contestable reasoning for its recommendations or actions. This may include traceable reasoning pathways, structured arguments, evidence-linked outputs, confidence indicators, ablation-based explanations of agent roles, and clinician-facing rationales. Treating transparency as a distinct principle enables the review to assess whether agentic systems merely perform tasks effectively or also make their reasoning accessible and interpretable to clinical users.

4.3. Governance Principles

The third dimension of the framework captures the governance conditions required for the responsible deployment of autonomous and semi-autonomous systems in healthcare. These principles were informed by regulatory and ethical guidance, including the European Union AI Act, the World Health Organization (WHO) guidance on ethics and governance of AI for health, the U.S. Food and Drug Administration (FDA) guidance on AI/ML-based software as a medical device, and HIPAA-related requirements for data protection and auditability [26], [27], [28]. Collectively, these sources emphasize that healthcare AI systems should be evaluated not only in terms of technical performance, but also with respect to oversight, accountability, safety, privacy protection, and post-deployment governance. Consequently, the framework incorporates two governance-oriented principles that are particularly relevant to deployment readiness.

The first principle, human-in-the-loop (HITL) oversight, refers to the presence of defined human review, escalation, approval, or intervention points within an agentic workflow. This principle is especially important when agentic systems influence diagnosis, treatment decisions, triage processes, or patient communication. The second principle, safety, alignment, and compliance, encompasses mechanisms designed to prevent harmful outputs, preserve privacy and security, support auditability, align system behavior with clinical and ethical expectations, and satisfy applicable regulatory requirements [29], [30]. Together, these governance principles represent essential conditions for responsible deployment. High technical performance alone is insufficient to justify clinical adoption without corresponding evidence of oversight, safety, accountability, and regulatory preparedness.

4.4. Rationale for Merged or Excluded Principles

Several candidate principles were considered during framework development but were not retained as independent dimensions. The objective was not to maximize the number of principles, but to establish a framework that was conceptually distinct, operationally observable within the empirical literature, and relevant to deployment-readiness assessment.

Privacy was initially considered as a standalone principle because of its well-established importance in healthcare AI. However, it was ultimately incorporated within the Safety, Alignment, and Compliance principle because its practical implementation is largely governed through existing legal, regulatory, and institutional mechanisms, including HIPAA, GDPR, audit controls, data-minimization requirements, and access-management policies [10], [31]. Treating privacy as a separate principle would have introduced substantial conceptual overlap with broader safety and compliance considerations while adding limited analytical distinction. Integrating privacy within the governance dimension therefore preserved its importance without duplicating related regulatory constructs.

Adaptability and learning were also considered as potential standalone principles. However, the reviewed literature generally operationalized these characteristics through autonomy, planning, reasoning, or self-correction mechanisms rather than treating them as independently evaluated constructs. Similarly, iterative self-correction was recognized as an important feature of agentic behavior, but it was not retained as a separate principle because it typically appeared as an implementation mechanism embedded within planning, reasoning, multi-agent critique, or tool-mediated feedback loops. Retaining self-correction as an independent principle would therefore have risked double-counting capabilities already represented elsewhere in the framework.

The resulting framework integrates three complementary perspectives: what agentic systems can do, how their reasoning can be understood, and the governance conditions under

which they may be responsibly deployed. The capability principles capture the technical architecture and operational behaviors of agentic systems; the transparency principle captures clinical intelligibility and explainability; and the governance principles capture deployment-readiness requirements related to oversight, safety, accountability, and compliance. This structure distinguishes the present review from prior studies that primarily organize the literature according to application domains, architectural paradigms, or functional archetypes.

The framework is also broadly consistent with the seven-dimensional taxonomy proposed by Vatsal et al. [10]. Although the terminology and analytical emphasis differ, both approaches cover a similar conceptual space encompassing agentic capabilities, transparency considerations, and governance requirements. This alignment provides additional support for the conceptual coverage of the framework while maintaining its specific focus on healthcare deployment readiness. In practical terms, the framework enables the review to distinguish studies that primarily demonstrate agentic capabilities from those that provide evidence relevant to responsible clinical deployment, a distinction that becomes central to the subsequent synthesis and discussion.

Table 5. Derivation of the seven-principle framework from prior literature

Principle	Category	Technical Source	Regulatory Source	Healthcare Rationale
Autonomy & Goal-Directed Behavior	Capability	Russell & Norvig [23]; ReAct [22]	—	Foundational property of agentic systems and prerequisite for higher-level agent behaviors.
Planning & Multi-Step Reasoning	Capability	Chain-of-Thought [32]; ReAct [22]; Tree-of-Thought [33]	—	Clinical decision-making frequently requires sequential reasoning across multiple evidence sources.
Multi-Agent Collaboration	Capability	Multi-Agent Debate [25]	—	Healthcare delivery is inherently multidisciplinary, making coordinated agent behavior relevant to clinical workflows.
Tool Use & Orchestration	Capability	Toolformer [24]	—	Effective healthcare AI requires interaction with EHRs, laboratory systems, imaging platforms, and clinical knowledge resources.
Explainability & Transparency	Transparency	SHAP, LIME, Attention-Based Methods [29], [30]	EU AI Act Article 13; WHO AI Ethics Guidance	Clinical accountability, informed decision-making, and medicolegal defensibility require interpretable and auditable reasoning.
Human-in-the-Loop Oversight	Governance	RLHF; Constitutional AI	EU AI Act Article 14; FDA SaMD Guidance	Human review and intervention remain important safeguards in high-risk clinical contexts.
Safety, Alignment & Compliance	Governance	AI Alignment Research; Red Teaming	EU AI Act; HIPAA; FDA 21 CFR	Preventing harm, ensuring accountability, and maintaining regulatory compliance are essential requirements for clinical deployment.

Figure 2 presents the evidence-informed synthesis framework used throughout this review. The framework organizes the seven principles into three interconnected dimensions—technical capability, transparency, and governance—which collectively support deployment-readiness assessment. By integrating established concepts from agentic AI research, healthcare AI, and regulatory guidance, the framework provides a structured analytical lens for evaluating both the capabilities and the deployment implications of agentic healthcare systems.

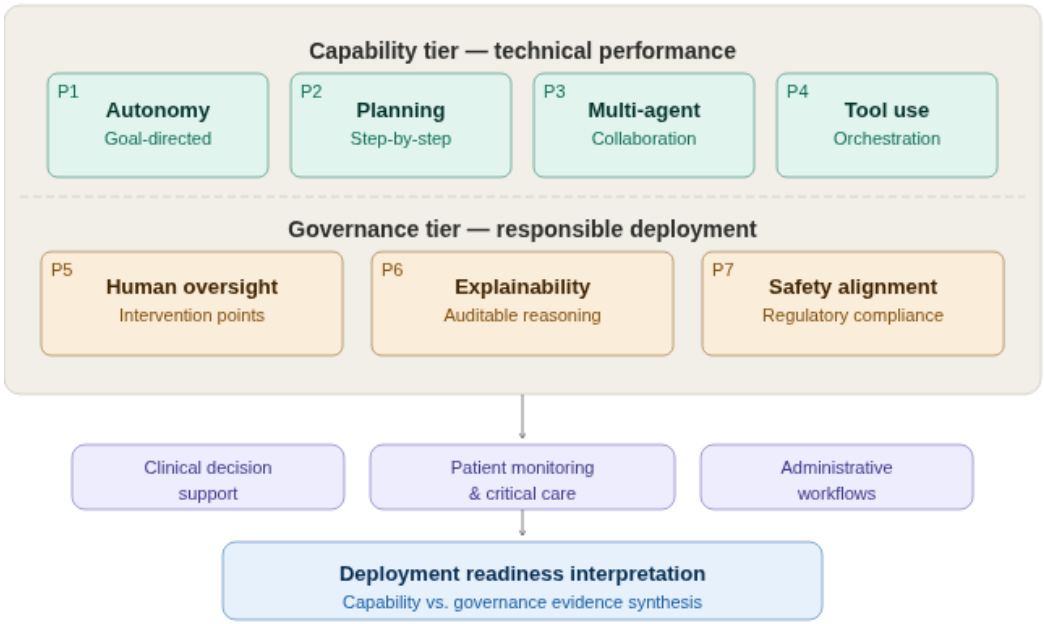

Figure 2. Evidence-informed seven-principle framework for healthcare deployment-readiness assessment.

Note: The framework integrates capability-oriented principles (Autonomy & Goal-Directed Behavior, Planning & Multi-Step Reasoning, Tool Use & Orchestration, and Multi-Agent Collaboration), the transparency principle (Explainability & Transparency), and governance-oriented principles (Human-in-the-Loop Oversight and Safety, Alignment & Compliance). The figure was created by the authors. Generative AI-assisted drafting was used solely for visual layout refinement; all conceptual content and final validation were completed by the authors.

4.5 Conceptual Boundaries and Analytical Value of the Framework

The seven-principle framework is grounded in three complementary bodies of literature that collectively inform the responsible deployment of agentic AI in high-stakes healthcare environments: technical research on agentic AI architectures [22], [32], [34], [35], AI safety and alignment literature [23], [36], and healthcare regulatory frameworks, including the EU AI Act, FDA guidance on Software as a Medical Device (SaMD), and the HIPAA Security Rule [26], [28], [37]. Drawing upon these sources, the framework organizes the principles into three analytically distinct dimensions: capability-oriented principles (Autonomy, Planning and Multi-Step Reasoning, Multi-Agent Collaboration, and Tool Use), a transparency-oriented principle (Explainability and Transparency), and governance-oriented principles (Human-in-the-Loop Oversight and Safety, Alignment, and Compliance). This structure reflects the widely recognized distinction between what AI systems are capable of doing and the safeguards required before those systems can operate in clinically consequential settings [38], [39].

Importantly, the three-tier structure is not intended as an arbitrary classification scheme. Rather, it reflects a distinction that is also embedded within contemporary regulatory and governance frameworks, where technical performance alone is considered insufficient for deployment without corresponding evidence of transparency, oversight, accountability, and safety controls. In this sense, the framework aligns technical evaluation with broader deployment-readiness considerations that are particularly relevant in healthcare environments.

Within each dimension, the principles are related but analytically separable, a distinction supported by the empirical distribution observed across the included studies. Among the 25 studies reviewed, none evaluated all seven principles simultaneously. Furthermore, the co-occurrence analysis presented in Section 5.3 (Table 10) shows that capability-oriented principles were frequently evaluated without corresponding governance-oriented assessment. For example, twelve studies evaluated autonomy without explicitly evaluating safety, alignment, or compliance mechanisms; four studies evaluated planning and reasoning without examining explainability; and six studies assessed tool-use capabilities without evaluating human

oversight mechanisms. These patterns are not merely conceptual distinctions but reflect observable differences in how research teams design and evaluate agentic healthcare systems.

The analytical value of the framework lies in its ability to reveal such gaps. A framework that merged these principles into broader categories would obscure important differences between technical capability and deployment-oriented safeguards. Existing domain-based reviews can identify where agentic AI is being applied, while architecture-based surveys can characterize how systems are constructed. However, neither approach directly reveals which deployment-relevant principles have accumulated empirical support and which remain comparatively underexplored. By making these distinctions visible, the framework helps identify areas where evidence remains limited and highlights a recurring pattern throughout the literature: empirical evaluation remains concentrated on capability-oriented principles, whereas governance-related dimensions—including human oversight, safety, compliance, and equity-sensitive validation—receive substantially less attention across healthcare domains and architectural approaches.

Table 6. Conceptual boundaries of the seven-principle framework

Principle	Core Analytical Question	Distinct Boundary
Autonomy & Goal-Directed Behavior	Can the system pursue a defined objective across multiple steps?	Distinct from planning because it concerns goal pursuit rather than the reasoning process used to achieve that goal.
Planning & Multi-Step Reasoning	Can the system decompose and sequence a clinical task into intermediate actions?	Distinct from autonomy because it focuses on task structure and reasoning logic.
Tool Use & Orchestration	Can the system interact with external resources, tools, or information systems?	Distinct from planning because it concerns execution and integration rather than reasoning.
Multi-Agent Collaboration	Can multiple specialized agents coordinate, critique, validate, or collaborate on a task?	Distinct from tool use because interactions occur among agents rather than between agents and external systems.
Explainability & Transparency	Can clinicians inspect, interpret, and challenge the system's reasoning process?	Distinct from performance because accurate outputs may still remain opaque or difficult to justify.
Human-in-the-Loop Oversight	Are formal mechanisms for human review, escalation, intervention, or approval incorporated into the workflow?	Distinct from explainability because explanations may support oversight but do not themselves guarantee human control.
Safety, Alignment & Compliance	Are mechanisms in place to address harm prevention, privacy, auditability, security, and regulatory obligations?	Distinct from HITL oversight because human review represents only one component of a broader governance and safety framework.

5. Characteristics of Included Studies

This section summarizes the characteristics of the 25 empirical studies that satisfied all PICOS eligibility requirements and quality-assessment criteria. The evidence base is examined across six dimensions: temporal distribution, publication venue, citation impact, authorship characteristics, healthcare-domain coverage, and the distribution of agentic AI principles. All descriptive statistics were derived directly from structured metadata extracted from the final corpus (N = 25). A summary of the principal quantitative characteristics is presented in Table 7.

5.1. Temporal Distribution

The included studies were published between January 2025 and April 2026, comprising 19 studies published in 2025 and 6 studies published during the first four months of 2026. When normalized by time, this corresponds to approximately 1.58 included studies per month in 2025 and between 1.50 and 1.71 studies per month in 2026, depending on whether April 2026 is treated as a complete or partial month. These values indicate a recently emerging and consistently active evidence base; however, they should not be interpreted as definitive

evidence of publication growth or acceleration within the field [40]. More broadly, the temporal distribution highlights the early developmental stage of empirical research on healthcare-focused agentic AI. As the literature continues to expand, future evidence syntheses may identify additional trends, deployment experiences, and governance-related evaluations that are not yet visible within the current corpus.

Table 7. Quantitative overview of included studies (N = 25)

Characteristic	Value	Details
Total Included Studies	N = 25	Met all PICOS eligibility and methodological quality criteria
Publication Years	2025–2026 (April)	19 studies (76%) in 2025; 6 studies (24%) in 2026
Publisher Families	3 groups	IEEE (n = 11, 44%); Nature/Springer Nature (n = 7, 28%); Elsevier (n = 7, 28%)
Citation Range	0–218	Median = 3; Mean = 20.0; Total citations = 500
Author Count	1–8 authors per study	Mean = 4.2; Median = 4
Healthcare Domains	3 primary domains	CDS (n = 15, 60%); Monitoring/ICU (n = 4, 16%); Administrative Workflows (n = 4, 16%); Cross-domain (n = 2, 8%)
Agentic Principles	7 principles	Mean = 3.1 principles per study; range = 2–7; Autonomy most frequently evaluated (68%)
Evaluation Contexts	5 categories	Laboratory/Benchmark (n = 15, 60%); Clinical/Retrospective (n = 5, 20%); Synthetic/Vignette (n = 3, 12%); Deployment-Adjacent (n = 1, 4%); Active Real-World Deployment (n = 1, 4%)

Note. The maximum value of the seven principles was observed in a taxonomy-oriented study that mapped healthcare LLM-agent literature across all seven framework dimensions. No primary agentic system evaluation comprehensively assessed all seven principles within a single clinical implementation.

5.2. Healthcare Domain Distribution

The 25 included studies were classified according to the primary healthcare domain described in their methods and evaluation sections. Clinical Decision Support (CDS) emerged as the dominant application area, whereas Patient Monitoring/ICU and Administrative Workflow applications were comparatively less represented. To avoid double counting, studies spanning multiple domains were categorized separately. The resulting distribution is presented in Table 8.

Table 8. Healthcare domain distribution of included studies (N = 25)

Healthcare Domain	Studies (n)	Percentage (%)
Clinical Decision Support (CDS)	15	60.0
Patient Monitoring & ICU	4	16.0
Administrative Workflows	4	16.0
Cross-Domain (All Three Domains)	1	4.0
CDS + Administrative Workflow	1	4.0
Total	25	100.0

Clinical Decision Support dominates the current evidence base, suggesting that agentic AI is presently being evaluated most frequently in settings where diagnostic reasoning, treatment support, and benchmark-oriented assessments are already relatively mature. In contrast, Patient Monitoring/ICU and Administrative Workflow applications remain less extensively studied, limiting the strength of conclusions regarding longitudinal monitoring, real-time surveillance, and operational deployment.

The relative scarcity of studies addressing administrative workflows is particularly noteworthy. Although such systems may offer near-term opportunities for implementation, they also introduce important challenges related to documentation quality, protected health information (PHI) handling, auditability, prior authorization, claims processing, and patient access.

Consequently, additional empirical evaluation is needed before broader deployment claims can be supported in these settings.

5.3. Agentic Principle Distribution

Each included study was coded according to the agentic AI principles that were explicitly evaluated, based on evidence reported in the methods, system architecture, evaluation design, and results sections. Because agentic systems often combine multiple capabilities, individual studies could receive more than one principle code. Across the 25 included studies, a total of 78 principle-level observations were identified, corresponding to a mean of 3.1 principles per study. The resulting frequency distribution is summarized in Table 9.

Table 9. Agentic principle frequency distribution across included studies (N = 25; Total Observations = 78)

Agentic Principle	Studies (n)	Frequency (%)	Rank
Autonomy & Goal-Directed Behavior	17	68.0	1st
Planning & Multi-Step Reasoning	11	44.0	2nd (tie)
Tool Use & Orchestration	11	44.0	2nd (tie)
Explainability & Transparency	11	44.0	2nd (tie)
Multi-Agent Collaboration	10	40.0	5th
Human-in-the-Loop (HITL) Oversight	8	32.0	6th
Safety, Alignment & Compliance	6	24.0	7th
Total Observations	78	—	—

Note. Because Vatsal et al. [10] represents a taxonomy-oriented empirical coding study rather than a primary healthcare agent evaluation, a sensitivity analysis was conducted excluding that study. The overall pattern remained unchanged. Capability-oriented principles remained more frequently represented than governance-oriented principles, and Safety, Alignment, and Compliance remained the least frequently evaluated principle. This suggests that the overall findings are not driven by the inclusion of the taxonomy-oriented study.

The distribution presented in Table 9 indicates that empirical attention is unevenly distributed across the seven-principle framework. Technical capability principles are substantially more prevalent than governance-oriented principles, suggesting that the current literature focuses more heavily on demonstrating what agentic systems can accomplish than on evaluating the safeguards required for responsible deployment. Although one taxonomy-oriented study addressed all seven principles simultaneously [10], no primary healthcare implementation evaluated all seven principles within a single clinical system or deployment context. Consequently, broad conceptual coverage should not be interpreted as evidence of comprehensive deployment readiness. To further examine how principles are evaluated together, Table 10 presents the co-occurrence patterns across the included studies.

Table 10. Principle co-occurrence matrix: frequency of joint evaluation across studies (N = 25)

Principle	P&R	Tool Use	XAI	MAS	HITL	Safety
Autonomy (n = 17)	10 (59%)	11 (65%)	8 (47%)	9 (53%)	7 (41%)	5 (29%)
Planning & Reasoning (n = 11)	—	9 (82%)	7 (64%)	7 (64%)	5 (45%)	4 (36%)
Tool Use (n = 11)	—	—	7 (64%)	7 (64%)	5 (45%)	4 (36%)
Explainability (n = 11)	—	—	—	7 (64%)	5 (45%)	4 (36%)
Multi-Agent (n = 10)	—	—	—	—	5 (50%)	3 (30%)
HITL (n = 8)	—	—	—	—	—	5 (63%)
Safety (n = 6)	—	—	—	—	—	—

Note. Cell values indicate the number of studies evaluating both the row principle and the column principle. Percentages are calculated relative to the total count of the row principle. P&R = Planning and Reasoning; XAI = Explainability and Transparency; MAS = Multi-Agent Collaboration.

The co-occurrence analysis reveals several notable patterns. The strongest association occurs between Planning and Multi-Step Reasoning and Tool Use & Orchestration, which co-occur in 9 of 11 planning-oriented studies (82%). This pattern is consistent with

contemporary LLM-agent architectures, where planning and reasoning are frequently implemented through tool-augmented workflows such as ReAct-style reasoning and external tool invocation [22], [24]. In practice, reasoning and tool interaction often function as tightly coupled capabilities rather than independent components.

A second notable pattern is the relationship between Human-in-the-Loop Oversight and Safety, Alignment, and Compliance, which co-occur in 63% of HITL studies. This suggests that governance-oriented evaluations are often framed around both oversight and safety considerations rather than addressing either dimension in isolation. More broadly, Table 10 indicates that capability-oriented principles tend to be evaluated together, whereas governance-oriented principles remain less frequently integrated into technical evaluations. The comparatively low representation of safety-related assessment is particularly noteworthy given its importance for clinical deployment. These findings reinforce a recurring theme across the reviewed literature: empirical evidence currently emphasizes agentic capability development more strongly than deployment-oriented governance and assurance mechanisms.

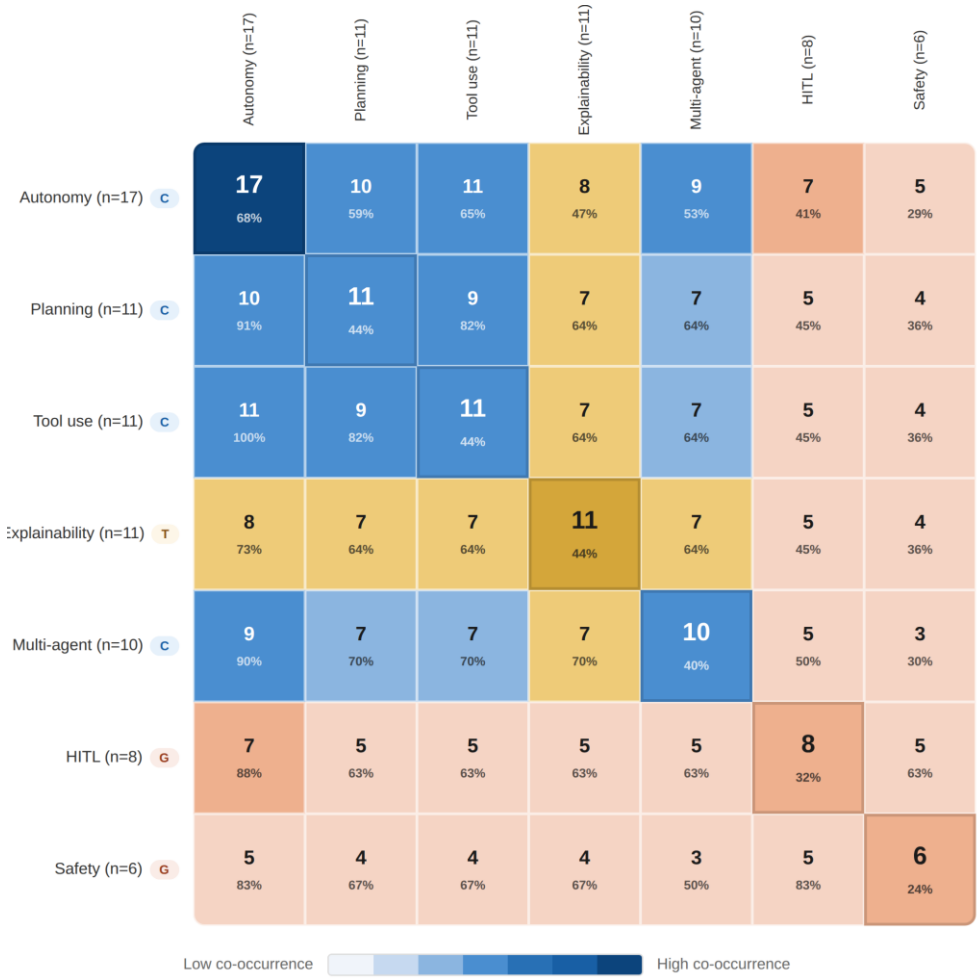

Figure 3. Co-Occurrence of agentic AI principles across included studies (N = 25).

Note. Capability-oriented principles exhibit relatively strong co-occurrence patterns, particularly between Planning and Multi-Step Reasoning and Tool Use & Orchestration. Governance-oriented principles appear less frequently and are often evaluated separately from technical capabilities. The strongest governance pairing is Human-in-the-Loop Oversight and Safety, Alignment, and Compliance. Overall, the figure highlights the concentration of empirical evaluation on technical capabilities relative to deployment-oriented safeguards.

5.4. Study Design and Evaluation Context

The included studies were further categorized according to their primary study design and evaluation context. Study-design categories reflect the principal methodological

contribution of each paper, whereas evaluation-context categories describe the environment, data source, and level of clinical realism used during empirical assessment. The resulting distribution is summarized in Table 11.

Table 11. Study design and evaluation context distribution (N = 25)

n	Category	Percentage (%)	Exemplar Studies
Study Design Type			
8	System Evaluation / Proof-of-Concept	32.0%	[40]–[47]
6	Framework Validation	24.0%	[10], [11], [37], [48]–[50]
4	Benchmark Evaluation (standardized dataset)	16.0%	[7], [51]–[53]
3	Comparative Evaluation vs. Baseline	12.0%	[13], [54], [55]
2	Retrospective Clinical Cohort / Deployment Pilot	8.0%	[20], [56]
1	Large-Scale Synthetic Vignette Study	4.0%	[57]
1	Proof-of-Concept with Real Clinical Cases	4.0%	[12]
Evaluation Context			
15	Laboratory / Benchmark Environment	60.0%	Public datasets; simulated clinical pipelines; benchmark corpora
5	Clinical / Retrospective Setting	20.0%	Real patient data; retrospective clinical records; clinical cohorts
3	Synthetic / Vignette-Based	12.0%	Computationally generated patient scenarios; clinical vignettes
1	Deployment-adjacent clinical evaluation	4.0%	Evaluated in a clinically realistic or workflow-linked setting, but not clearly active deployment
1	Deployment-adjacent clinical evaluation	4.0%	Evaluated in a clinically realistic or workflow-linked setting, but not clearly active deployment

The distribution shown in Table 11 suggests that the current evidence base remains substantially closer to technical feasibility assessment than to real-world clinical deployment. Most studies were classified as system evaluations, proof-of-concept implementations, framework validations, or benchmark-based assessments. In contrast, prospective clinical evaluations and active deployment studies were comparatively rare. Only one study clearly reported active real-world deployment, while one additional study was classified as deployment-adjacent because it was evaluated within a clinically relevant workflow but did not constitute routine operational use.

For interpretive purposes, laboratory/benchmark studies and synthetic/vignette-based studies can be viewed collectively as non-clinical or simulation-oriented evaluations. Together, these categories account for 18 of the 25 included studies (72%). The remaining seven studies incorporated some degree of clinical grounding through retrospective patient data, deployment-adjacent evaluation, or active real-world implementation.

This distribution has important implications for interpreting the evidence base. Current studies provide relatively strong support for technical feasibility and proof-of-concept performance but offer more limited evidence regarding clinical effectiveness, operational integration, patient outcomes, safety performance, or equity-related impacts. Consequently, prospective evaluations incorporating real-world workflows, patient-centered outcomes, safety monitoring, and equity-sensitive assessment remain among the most important gaps in the emerging literature on agentic AI in healthcare.

5.5. Summary of Descriptive Patterns

Taken together, the descriptive findings suggest that the empirical literature on agentic AI in healthcare remains active but is still at an early stage of translational maturity. Clinical Decision Support (CDS) is the most extensively studied domain, whereas Patient Monitoring/ICU and Administrative Workflow applications remain comparatively underrepresented. Across the seven-principle framework, capability-oriented principles are evaluated more frequently than governance-oriented principles, with autonomy, planning, tool use, and multi-agent collaboration appearing substantially more often than Human-in-the-Loop (HITL) oversight and Safety, Alignment, and Compliance.

The evaluation context provides the clearest indicator of evidence maturity. Of the 25 included studies, 18 were conducted in laboratory, benchmark, or synthetic settings, while only one study reported active real-world clinical deployment. Collectively, these patterns suggest that the current evidence base provides stronger support for technical feasibility and proof-of-concept performance than for clinical deployment, operational integration, or long-term governance readiness.

5.6. Characteristics of Included Studies

Table 12 summarizes the included studies according to healthcare domain, evaluated principles, and principal empirical outcomes. The table is provided primarily to support transparency and traceability of the synthesis rather than to indicate comparative evidence quality or methodological rigor.

Table 12. Characteristics of included studies

Reference	Domain	Principle(s)	Key Empirical Outcome
Ferber et al. [12]	CDS – Oncology	Autonomy, Planning, Tool Use	Autonomous multi-step oncology agent; proof-of-concept validation on real clinical cases
Zhao et al. [7]	CDS – Rare Disease	Autonomy, Planning, Reasoning	DeepRare: MIMIC-IV-Rare benchmark; performance edge maintained on in-house clinical dataset
Johri et al. [11]	CDS – Patient Interaction	HITL, Explainability, Planning	CRAFT-MD: clinical LLM evaluation framework; empirical patient interaction task benchmarks
Liu et al. [13]	CDS – Multi-Cancer	Multi-Agent, Planning, Autonomy	EvoMDT: self-evolving multi-agent; validated on lung adenocarcinoma, breast cancer, HCC
Hao et al. [20]	CDS – Patient Education	Tool Use, HITL, Autonomy	EHR-integrated LLM agent; real-world clinical deployment; patient usability evaluation
Omar et al. [57]	CDS – E-Medicine	HITL, Autonomy, Safety	1,000 synthetic e-visit vignettes; communication style impact on agentic AI advice quality
Freyer et al. [37]	CDS/Safety – Regulatory	Safety, Alignment, HITL	Regulatory barrier taxonomy: policy framework for AI agent implementation in clinical practice
Tzanis & Klont [51]	CDS – Radiology	Multi-Agent, Tool Use, Explainability	mAIstro: evaluated on 16 open-source datasets; automated radiomics pipeline
Linton-Reid et al. [56]	CDS – Oncology Imaging	Autonomy, Planning	Agentic AI radiomics: 89.6% therapy prediction accuracy in 500-patient clinical deployment
Wang et al. [58]	CDS – Hematology	Autonomy, HITL, Reasoning	Interactive guideline-based CDS; evaluated vs. two alternative approaches in hematology
Loaiza-Bonilla et al. [55]	CDS – Oncology	Multi-Agent, Tool Use, HITL	Neuro-symbolic multi-agent trial matching; superior accuracy vs. LLM-only baselines
Liu & Xiao [54]	CDS	Multi-Agent, Explainability	Argumentation-based multi-agent CDS: +13% accuracy over single-agent systems
Chen et al. [44]	CDS	Multi-Agent, Explainability, Safety	Multi-agent CDS with ethical AI governance; explainability-accuracy balance evaluation
Lee et al. [48]	CDS – Evidence Synthesis	Tool Use, Autonomy, Planning	A4SLR framework: F1 = 0.96–0.998 on article screening and data extraction tasks
Vatsal et al. [10]	Multi-domain	All 7 principles (taxonomy)	Seven-dimensional taxonomy; empirical labeling of 49 healthcare LLM-agent studies
Qiao et al. [45]	Patient Monitoring – Nursing	Autonomy, Multi-Agent, HITL	Parallel nursing: LLM agents for postoperative patient risk prediction and care coordination

Table 12 cont.

Sathya & Valaramathi [53]	Monitoring – Ophthalmology	Autonomy, Planning, Explainability	Agentic AI for diabetic retinopathy: superior benchmark performance vs. non-agentic baselines
Li et al. [42]	Monitoring – Diabetes	Autonomy, Planning, Tool Use	Agentic diabetes management framework; biomedical + lifestyle + psychosocial integration
Mukhtar et al. [46]	Monitoring – Neurology	Multi-Agent, HITL, Explainability	Multi-agent Parkinson's support: 95% clinical accuracy; critique agent factual validation
Shimgekar et al. [49]	Admin – Clinical Pipeline	Autonomy, Tool Use, Explainability	End-to-end medical data pipeline: ingestion, anonymization, feature extraction, inference
Patel et al. [50]	Admin – Cardiology Workflow	Autonomy, Tool Use, Planning	NLP and agentic AI for electrophysiology documentation and workflow automation
Mohapatra et al. [40]	Admin – Hospital Ops	Autonomy, Multi-Agent, Tool Use	Three-agent system: onboarding, documentation, and triage coordination agents

Note. Framework- and taxonomy-oriented studies were included only when they contained an empirical coding, validation, or assessment component. Purely conceptual taxonomies without empirical evaluation were excluded from the evidence synthesis.

While Table 12 provides study-level traceability, the broader evidence patterns become clearer when the findings are aggregated at the principle level. Table 13 therefore presents a consolidated evidence matrix that maps each principle to its empirical footprint, dominant evaluation context, evidence-maturity profile, and key research gaps across the 25 included studies.

Table 13. Core agentic AI principles evidence matrix (N = 25 Studies)

Principle	Studies (n)	Domains	Dominant Evaluation Context	Evidence Maturity / Quality Signal	Strongest Evidence	Key Gap Identified
Autonomy & Goal-Directed Behavior	n = 17	CDS, Monitoring, Admin	Mostly benchmark, proof-of-concept, and limited clinical evaluation	Moderate-to-high technical feasibility; limited real-world validation	[7], [12]	Long-horizon autonomous operation in live clinical settings with real patient outcomes
Planning & Multi-Step Reasoning	n = 11	CDS, Monitoring	Mostly benchmark and structured evaluation	Strong technical reasoning evidence; limited clinical verification of reasoning processes	[12], [13]	Causally grounded and clinically verifiable reasoning pathways
Multi-Agent Collaboration	n = 10	CDS, Monitoring, Admin	Comparative, benchmark, and proof-of-concept evaluations	Emerging comparative evidence; limited accountability and conflict-resolution assessment	[46], [54], [55]	Agent consensus, conflict resolution, and cross-specialty coordination
Tool Use & Orchestration	n = 11	CDS, Admin, Monitoring	Workflow, benchmark, and pipeline evaluations	Strong feasibility evidence; limited failure-recovery evaluation	[48], [49], [51]	Tool-failure recovery, hallucination under tool invocation, and FHIR interoperability
Human-in-the-Loop (HITL) Oversight	n = 8	All three domains	Framework, communication, and design-oriented evaluations	Widely recognized as important; limited empirical testing of oversight models	[11], [37], [57]	Determination of optimal intervention thresholds across clinical-acuity levels

Table 13 cont.

Explainability & Transparency	n = 11	CDS, Monitoring	Benchmark, architecture-level, and structured-reasoning evaluations	Moderate evidence of explanation availability; limited evidence of clinical usefulness	[44], [51], [52]	Clinician-centered XAI design and role-specific explanation strategies
Safety, Alignment & Compliance	n = 6	Admin, CDS, Monitoring	Governance, taxonomy, and design-oriented evaluations	Lowest empirical maturity; limited prospective safety validation	[10], [41], [59]	Prospective real-world safety validation and liability standards for autonomous clinical harm

Table 13 highlights a consistent pattern across the evidence base. Capability-oriented principles—including autonomy, planning, multi-agent collaboration, and tool use—show the strongest empirical support and are typically evaluated through benchmark, proof-of-concept, or workflow-oriented studies. In contrast, governance-oriented principles remain less mature, with relatively few studies evaluating oversight mechanisms, safety controls, regulatory readiness, or accountability frameworks through prospective empirical designs.

The evidence matrix also demonstrates that the nature of the research gaps differs across principles. For capability-oriented principles, the primary challenge is translation from technical feasibility to sustained clinical operation. For governance-oriented principles, the challenge is more fundamental: empirical evidence remains limited, and many proposed safeguards have yet to be evaluated under realistic deployment conditions. Explainability occupies an intermediate position, with a growing body of technical approaches but comparatively limited evidence regarding their usefulness for clinicians in real-world decision-making.

Viewed collectively, the matrix reinforces the central interpretation of this review: current evidence is strongest for demonstrating what agentic healthcare systems can do, while evidence regarding how such systems can be safely governed, supervised, and deployed remains comparatively underdeveloped. This imbalance represents one of the most significant barriers to broader clinical adoption and motivates several of the research priorities discussed in the following sections.

6. Results

The principle-level synthesis reveals a consistent pattern across the included studies: empirical evidence is currently stronger for technical capability than for deployment readiness. Capability-oriented principles, including autonomy, planning, tool use, and multi-agent collaboration, were evaluated more frequently than governance-oriented principles such as Human-in-the-Loop (HITL) oversight and Safety, Alignment, and Compliance. These principles also tended to co-occur more frequently, suggesting that agentic healthcare systems are increasingly being developed as integrated technical workflows, whereas governance mechanisms are being evaluated less consistently.

Evidence maturity varied substantially across evaluation settings. Most studies demonstrated feasibility in laboratory, benchmark, synthetic, or proof-of-concept environments, while only a limited number incorporated real-world clinical workflows or deployment contexts. To avoid treating principle frequency as a proxy for evidence strength, reported outcomes were grouped into six categories: diagnostic or predictive performance, workflow and operational efficiency, reasoning and explanation quality, human interaction and oversight, safety and compliance governance, and equity or subgroup performance. The resulting synthesis, summarized in Table 14, provides a complementary outcome-level perspective on the current evidence base.

Table 14. Summary of reported outcome types across included studies

Outcome Category	Included Examples	Typical Evidence Type	Interpretation for Deployment Readiness
Diagnostic or Predictive Performance	Accuracy, F1 score, therapy-response prediction, rare-	Benchmark, retrospective, clinical dataset, or	Supports technical feasibility; deployment strength

Table 14 cont.

	disease diagnosis, radiology VQA	proof-of-concept evaluation	depends on data source and validation context
Workflow or Operational Efficiency	Documentation support, literature screening, data-pipeline automation, prior authorization, administrative triage	Workflow evaluation, pipeline execution, or task-completion metrics	Supports operational feasibility but often lacks prospective workflow validation
Reasoning and Explanation Quality	Traceable reasoning chains, argumentation, ablation analysis, conversational reasoning quality	Structured evaluation or benchmark-based reasoning assessment	Supports auditability but does not necessarily validate clinical correctness
Human Interaction and Oversight	Patient-facing communication quality, critique agents, escalation protocols, clinician review points	Vignette studies, design-level evaluations, or regulatory analyses	Indicates emerging HITL mechanisms but provides limited evidence regarding optimal intervention thresholds
Safety, Compliance, and Governance	Regulatory barriers, safety constraints, privacy/security safeguards, auditability, compliance requirements	Policy analyses, taxonomy studies, and design-level safeguards	Identifies deployment requirements but rarely provides direct prospective validation
Equity and Subgroup Performance	Demographic-stratified accuracy, subgroup robustness, fairness analysis	Not reported in included studies	Major deployment-readiness gap

Table 14 further highlights the imbalance between capability-oriented and deployment-oriented evidence. Across the corpus, technical feasibility, task performance, and workflow execution are evaluated substantially more often than safety assurance, oversight effectiveness, equity considerations, or real-world implementation outcomes. The following subsections therefore examine each principle in greater detail, focusing on the available evidence, interactions with other principles, and the remaining challenges for responsible clinical deployment.

6.1 Principle 1: Autonomy and Goal-Directed Behavior

Autonomy and Goal-Directed Behavior appeared in 17 of the 25 included studies, making it the most frequently represented principle. Evidence was observed across a wide range of healthcare applications, including oncology decision support, rare-disease diagnosis, patient education, monitoring, documentation support, and administrative workflow automation [7], [12], [20]. Collectively, these studies demonstrate that agentic systems can pursue predefined clinical or operational objectives with limited human intervention across bounded tasks.

However, the evidence should be interpreted within the context of the evaluation setting. Most studies assessed autonomy in benchmark datasets, retrospective analyses, proof-of-concept systems, or narrowly scoped workflow environments rather than in sustained clinical operations. The strongest examples demonstrate autonomy within clearly defined operational boundaries. Similarly, administrative and workflow-oriented studies show that agents can coordinate documentation, anonymization, data processing, and inference tasks, but these evaluations generally remain task-specific rather than evidence of broad autonomous clinical practice.

Taken together, the literature supports bounded task-level autonomy rather than unrestricted clinical autonomy. Current studies show that agentic systems can pursue defined objectives across diagnostic, monitoring, and administrative contexts, yet relatively few evaluate how autonomous behavior performs under changing patient conditions, incomplete information, workflow interruptions, or competing clinical priorities [7], [12], [20]. The key challenge is therefore not whether autonomy can be implemented technically, but whether autonomous behavior can be safely bounded, monitored, audited, and escalated within real clinical environments. Future work should evaluate autonomy under conditions of uncertainty and explicitly assess escalation mechanisms, failure modes, and long-term operational behavior.

6.2 Principle 2: Planning and Multi-Step Reasoning

Planning and Multi-Step Reasoning appeared in 11 of the 25 included studies and frequently served as the mechanism connecting clinical objectives, evidence retrieval, tool use, and output generation. Co-occurrence analysis showed that planning was commonly evaluated alongside Tool Use & Orchestration (9 of 11 studies), Explainability & Transparency (7 of 11 studies), and Multi-Agent Collaboration (7 of 11 studies), indicating that planning is rarely implemented as an isolated capability in healthcare agentic AI.

The strongest evidence originated from studies addressing complex clinical tasks that required sequential reasoning rather than single-turn prediction [7], [13]. Several studies also examined reasoning quality directly through structured argumentation, traceable reasoning pathways, or clinician-interpretable decision processes [11], [43], [60]. Nevertheless, the existence of a reasoning chain should not be interpreted as evidence that the reasoning itself is clinically valid. Reasoning processes may remain incomplete, contain unsupported assumptions, or be optimized primarily for benchmark performance rather than real-world clinical decision-making.

Overall, the evidence suggests that multi-step reasoning has become a central component of agentic healthcare systems, particularly when tasks require integration of evidence retrieval, tool invocation, and recommendation generation. However, most studies evaluate whether reasoning pathways can be produced rather than whether they are medically sound, reproducible, auditable, and robust under uncertainty [7], [13], [61]. Future evaluations should therefore focus not only on generating interpretable reasoning chains but also on validating their clinical correctness, consistency, and resilience across diverse healthcare scenarios.

6.3. Principle 3: Multi-Agent Collaboration

Multi-Agent Collaboration appeared in 10 of the 25 included studies and was most commonly associated with tasks requiring distributed reasoning, critique, validation, or consensus formation. Rather than functioning as an isolated design feature, multi-agent architectures frequently reinforced other principles, particularly Planning and Multi-Step Reasoning, Explainability and Transparency, and Human-in-the-Loop Oversight. This pattern reflects the collaborative nature of healthcare decision-making, where complex clinical judgments often benefit from multiple perspectives and expertise domains. Agentic systems increasingly emulate this structure by assigning specialized roles to diagnostic, retrieval, critique, and validation agents [62].

The strongest evidence for multi-agent benefits came from studies that compared multi-agent architectures against single-agent or LLM-only baselines. These studies reported potential improvements in decision quality, robustness, throughput, or reliability when specialized agents were assigned complementary responsibilities such as argument generation, critique, consensus formation, or clinician escalation [54], [55]. Other studies primarily provided design-feasibility evidence, demonstrating how multi-agent architectures can support heterogeneous clinical reasoning, internal quality assurance, or domain-specific workflow coordination [13], [46], [59].

Overall, multi-agent collaboration appears promising but remains at an early stage of clinical maturity. The available evidence suggests that role specialization and critique mechanisms may improve robustness [13], [44], [46], [50]; however, multi-agent systems also introduce new challenges, including error propagation, responsibility attribution, coordination failures, and false confidence arising from apparent consensus. Future evaluations should therefore examine not only whether multi-agent systems outperform single-agent baselines, but also how they resolve disagreement, assign accountability, document reasoning processes, and support safe escalation when conflicting outputs are produced.

6.4 Principle 4: Tool Use and Orchestration

Tool Use and Orchestration appeared in 11 of the 25 included studies and served as the primary mechanism linking agentic reasoning with healthcare workflow execution. Across the reviewed literature, tool use extended well beyond information retrieval and included interaction with EHR systems, clinical knowledge repositories, imaging and radiomics pipelines, literature databases, anonymization tools, feature-extraction systems, and model-selection workflows. This helps explain its frequent co-occurrence with Planning and Multi-Step

Reasoning, as agents must not only determine appropriate actions but also select, sequence, and validate the tools required to execute those actions.

The strongest evidence emerged from studies implementing end-to-end workflows rather than isolated model outputs. Examples included radiomics and deep-learning pipeline automation, evidence-synthesis support systems, EHR-grounded patient education tools, and administrative workflow platforms supporting anonymization, semantic matching, inference generation, and explainability [20], [47], [48], [51]. Collectively, these studies illustrate how tool orchestration enables agentic AI to move beyond language generation toward operational healthcare tasks.

Overall, the evidence supports the technical feasibility of tool-supported agentic workflows. However, existing evaluations focus predominantly on successful tool execution rather than failure management. Relatively few studies examined incorrect tool selection, interoperability failures, incomplete or conflicting data, auditability, recovery mechanisms, or resilience when external services become unavailable. Future research should therefore evaluate not only whether agents can use tools effectively, but whether they can do so safely, transparently, and reliably under realistic clinical conditions.

6.5. Principle 5: Human-in-the-Loop (HITL) Oversight

Human-in-the-Loop (HITL) Oversight appeared in 8 of the 25 included studies and represents one of the clearest gaps between technical capability and deployment readiness. Across the evidence base, HITL was widely recognized as an important requirement, yet comparatively few studies operationalized it through explicit review points, escalation criteria, intervention thresholds, or responsibility frameworks.

The oversight mechanisms reported in the literature generally followed several recurring patterns. These included critique agents that screened outputs for factual or clinical relevance before user exposure, clinician-escalation mechanisms for higher-risk recommendations, patient-facing assessments of reasoning quality, and regulatory arguments supporting continuous monitoring of adaptive or multi-step agentic systems [11], [48], [51], [59], [63]. Together, these studies suggest that HITL is gradually evolving from a broad governance concept into a more concrete workflow-design principle.

Nevertheless, substantial evidence gaps remain. Few studies specified what degree of uncertainty should trigger human review, which outputs require clinician approval, how responsibility should be assigned after review, or how oversight requirements should differ between low-risk administrative tasks and high-risk clinical recommendations. This limitation is particularly important because human involvement alone does not automatically guarantee safety or accountability. Future work should therefore evaluate concrete oversight models, including escalation thresholds, reviewer workload, disagreement management, and the extent to which HITL mechanisms improve safety, accountability, and workflow feasibility under real clinical conditions.

6.6. Principle 6: Explainability and Transparency

Explainability and Transparency appeared in 11 of the 25 included studies and served as a critical link between technical performance and clinical trust. Across the reviewed literature, explainability was implemented primarily through three approaches: traceable reasoning, structured argumentation, and architectural transparency. These approaches move beyond traditional post-hoc explanation by embedding transparency directly into the design and operation of the system. Rather than merely explaining outputs after they are generated, the stronger studies sought to make reasoning processes visible, identify supporting evidence, and clarify how system components contributed to final recommendations [7], [51], [52], [54].

The available evidence suggests that explainability in agentic healthcare AI is gradually shifting from an auxiliary interpretive layer toward a more integrated form of transparency. Traceable reasoning pathways support auditability, structured argumentation makes recommendations contestable, and ablation-based or workflow-level analyses help clarify the contribution of individual agents or system components to overall performance.

However, most studies focused on whether explanations could be generated rather than whether they were useful in practice. Technical availability does not necessarily translate into clinical understanding, trust, or appropriate use. Consequently, future research should

evaluate explanation quality directly with clinical users and compare different explanation formats across healthcare roles, workflow settings, and risk levels.

6.7 Principle 7: Safety, Alignment, and Regulatory Compliance

Safety, Alignment, and Regulatory Compliance was the least frequently evaluated principle, appearing in only 6 of the 25 included studies. The nature of the evidence in this area also differed substantially from that supporting capability-oriented principles. Whereas autonomy, planning, tool use, and multi-agent collaboration were commonly evaluated through task performance, benchmark accuracy, workflow efficiency, or comparative studies, safety and compliance were more often addressed through regulatory analyses, taxonomy-oriented assessments, or design-level safeguards rather than direct evaluation of safety outcomes.

This distinction is important when interpreting deployment readiness. Regulatory and governance studies help clarify the conditions required for responsible clinical deployment, including accountability mechanisms, post-market monitoring, privacy protection, auditability, and compliance obligations [41]. However, such analyses do not demonstrate that a specific agentic system can prevent harm under real-world clinical conditions. Similarly, design-level safeguards indicate that safety considerations can be incorporated into system architectures, but they are not equivalent to prospective validation of unsafe-output detection, escalation effectiveness, bias monitoring, auditability, or post-deployment surveillance [46].

Taken together, the literature demonstrates broad recognition that safety, alignment, and compliance are essential requirements for healthcare deployment. However, empirical validation remains limited. Current evidence primarily identifies regulatory barriers, governance challenges, and architectural safeguards rather than directly evaluating clinical safety outcomes. As a result, Safety, Alignment, and Compliance remain the least mature principle within the framework. Future studies should treat safety, privacy and security, auditability, bias monitoring, escalation effectiveness, alignment mechanisms, and post-deployment surveillance as measurable outcomes rather than background assumptions. This transition from conceptual requirement to empirically validated practice represents one of the most important challenges for the next generation of agentic healthcare AI research.

6.8 Evidence from Healthcare Domains

To complement the principle-level analysis, the included studies were also examined according to their primary healthcare application domain. Table 15 summarizes the evidence across Clinical Decision Support (CDS), Patient Monitoring and ICU-related applications, Administrative Workflows, and cross-domain studies. This domain-level perspective helps clarify where agentic AI has progressed most rapidly and where important evidence gaps remain.

Table 15. Evidence summary by healthcare domain (N = 25 Studies)

Domain	n	Dominant Principles	Key Empirical Evidence
Clinical Decision Support	15	Autonomy, Planning & Reasoning, Multi-Agent Collaboration, Explainability	Autonomous oncology agent validated on real clinical cases [12]; DeepRare evaluated on MIMIC-IV-Rare and in-house datasets [7]; EvoMDT for multi-cancer decision support [13]; argumentation-based multi-agent CDS with +13% accuracy improvement [54]; 89.6% therapy prediction accuracy in a 500-patient cohort [56]; mAIstro evaluated across 16 radiomics datasets [51]; neuro-symbolic clinical trial matching [55]
Patient Monitoring & ICU	4	Autonomy, Planning, Multi-Agent Collaboration, Explainability	Multi-agent Parkinson's support system achieving 95% clinical accuracy [46]; agentic diabetic retinopathy screening outperforming non-agentic baselines [53]; parallel nursing agents for postoperative risk prediction and care coordination [45]; diabetes management framework integrating

Table 15 cont.

			biomedical, lifestyle, and psychosocial information [42]
Administrative Workflows	4	Autonomy, Tool Use & Orchestration, Safety & Compliance, Multi-Agent Collaboration	End-to-end clinical data pipeline from ingestion to inference [49]; electrophysiology workflow automation [50]; three-agent hospital operations system supporting onboarding, documentation, and triage [40]; healthcare insurance automation with empirical compliance evaluation [47]
Cross-Domain / Multi-Domain	2	Mixed principle coverage	Studies spanning multiple healthcare domains or providing taxonomy-level evidence across agentic AI applications

Table 15 shows that empirical evidence remains concentrated in Clinical Decision Support, while Patient Monitoring and Administrative Workflow applications are comparatively less represented. Across all domains, technical capabilities are generally better supported than governance-related aspects, reinforcing the broader pattern identified throughout the review.

6.8.1 Clinical Decision Support

Clinical Decision Support (CDS) was the most represented domain, accounting for 15 of the 25 included studies. Compared with other domains, the evidence base for CDS is relatively more mature because studies more frequently employed benchmark datasets, retrospective clinical records, comparative evaluations, or clinically grounded validation settings. These characteristics provide stronger evidence of technical capability and task-level effectiveness than is currently available in other healthcare domains.

Nevertheless, most CDS studies remain focused on bounded decision-support scenarios rather than prospective clinical implementation. The literature demonstrates that agentic systems can assist with diagnosis, treatment planning, evidence retrieval, and patient-specific recommendation generation, but substantially less evidence exists regarding their integration into routine clinical workflows. Consequently, the primary limitation is not the absence of agentic capability, but the lack of evidence showing that these systems can operate safely and effectively within real clinical environments that require oversight, auditability, accountability, and patient-outcome evaluation. Future research should therefore move beyond performance validation toward prospective assessments of clinical effectiveness, workflow integration, and deployment readiness.

6.8.2 Patient Monitoring and ICU

Patient Monitoring and ICU-related applications accounted for 4 of the 25 included studies. The available evidence suggests that agentic AI is emerging primarily in structured or semi-structured monitoring scenarios, including chronic disease management, neurological support [46], ophthalmic screening [53], postoperative surveillance [45], and diabetes management [42]. These applications generally involve more predictable data streams and narrower evaluation targets than high-acuity ICU environments, which may partly explain the current distribution of evidence.

Overall, the literature supports early-stage feasibility rather than mature deployment. Several studies reported encouraging task-level performance and clinically relevant support functions, but prospective validation remains limited. In particular, there is little evidence regarding the performance of agentic monitoring systems in dynamic environments where patient conditions change rapidly and escalation decisions are time-sensitive. Future studies should evaluate whether agentic monitoring systems improve deterioration detection, escalation accuracy, clinician workload, patient outcomes, and safety when deployed within real clinical workflows.

6.8.3 Administrative Workflows

Administrative Workflow applications also accounted for 4 of the 25 included studies and were predominantly evaluated through proof-of-concept or constrained workflow settings [59]. The evidence suggests that agentic AI may support a range of operational tasks, including documentation automation [50], data processing, workflow routing, prior authorization, anonymization, and administrative decision support.

However, the current evidence remains closer to operational feasibility than deployment readiness. Although administrative systems are often perceived as lower risk than diagnostic or treatment-support systems, they can still influence important aspects of healthcare delivery, including documentation integrity, referral pathways, insurance coverage, claims processing, patient access, PHI management, auditability, and institutional compliance. Consequently, the central challenge is not only technical performance but also the ability of these systems to operate within healthcare's regulatory, legal, and accountability requirements [64].

Future research should therefore evaluate administrative agentic AI under real operational conditions, with particular attention to error propagation, compliance, privacy and security protection, audit-trail integrity, staff workload, and downstream effects on patient access and continuity of care. Compared with clinical decision support, administrative workflows may provide a more accessible pathway toward near-term deployment; however, substantial governance and validation challenges remain before widespread adoption can be justified.

7. Discussion and Limitations

This review synthesized recent empirical evidence on agentic AI in healthcare through a deployment-readiness perspective. The findings indicate that progress across the field is uneven. While a growing number of studies demonstrate bounded technical capabilities such as autonomy, planning, tool use, and multi-agent collaboration, substantially less evidence exists for the governance conditions required for clinical translation. In particular, Human-in-the-Loop (HITL) oversight, safety and compliance, privacy and security safeguards, equity-sensitive evaluation, and post-deployment monitoring remain underrepresented relative to the pace of technical development. Taken together, the evidence suggests that agentic healthcare AI is becoming increasingly feasible in controlled settings, yet its readiness for safe, accountable, and equitable clinical deployment remains insufficiently validated.

These findings should be interpreted in light of several limitations. First, the review protocol was not prospectively registered, limiting external verification of protocol adherence. Second, the 16-month review window was intentionally selected to capture the most recent empirical evidence, but this choice reduces historical contextualization and may exclude rapidly emerging studies published after the search period. Third, Google Scholar searching and citation chasing improved retrieval breadth but are inherently less reproducible than structured database searches; however, peer-review status and eligibility criteria were verified before study inclusion. Fourth, data extraction, quality appraisal, and principle coding were not independently duplicated with a formal inter-rater reliability assessment. Fifth, citation counts were used solely as descriptive metadata and should be interpreted cautiously because many included studies were recently published. Sixth, substantial heterogeneity in study design, outcome reporting, and evaluation context precluded quantitative meta-analysis and necessitated narrative synthesis. Finally, the limitations of the underlying literature also constrain the conclusions of this review. Most studies remain laboratory, benchmark, proof-of-concept, or retrospective evaluations, and none reported demographic-stratified performance outcomes.

The most important finding is the gap between capability-oriented evidence and deployment-readiness evidence. Capability-related principles appeared more frequently and were typically supported by measurable outcomes such as benchmark performance, workflow efficiency, reasoning quality, or comparison against baseline systems. In contrast, governance-oriented principles, particularly HITL oversight and Safety, Alignment, and Compliance, were evaluated less frequently and were more often discussed through policy analyses, design recommendations, or taxonomy-oriented assessments rather than direct empirical validation. This distinction is particularly important because agentic AI systems differ from conventional clinical AI tools. Their behavior may emerge from multi-step planning, interaction with external tools, coordination among specialized agents, and evolving intermediate outputs. Consequently, strong task performance alone is insufficient to establish deployment readiness. A system may achieve high benchmark accuracy while still lacking validated escalation procedures, auditability mechanisms, privacy and security safeguards, subgroup robustness, or post-deployment monitoring processes. The evidence therefore supports cautious optimism regarding technical feasibility, but it does not yet justify broad claims of clinical effectiveness or deployment safety.

These findings extend prior reviews that primarily classify agentic AI according to application domains, architectural paradigms, or functional categories [15]-[19]. Such classifications are valuable for mapping the field, but they do not directly address whether agentic systems are prepared for regulated healthcare deployment. By applying an evidence-informed seven-principle framework, the present review shifts the focus from the question of where agentic AI is being used to the question of which deployment-relevant principles are supported by empirical evidence. This distinction provides additional insight beyond application- or architecture-based taxonomies. Reviews centered on system categories can identify growth in LLM agents, multi-agent architectures, or workflow assistants, but they may obscure whether those systems incorporate explainability, oversight, safety controls, privacy protections, or equity-sensitive validation. The current synthesis demonstrates that these governance-related dimensions are not progressing at the same rate as technical capabilities. As a result, the framework offers a deployment-readiness perspective that is less visible in conventional taxonomic or scoping reviews.

7.1. Meaning of the Laboratory-to-Clinic Gap

The evaluation setting strongly influences how current evidence should be interpreted. Most included studies were conducted in laboratory, benchmark, synthetic, proof-of-concept, or retrospective environments. These studies are valuable because they establish whether agentic mechanisms can be implemented, measured, and evaluated. However, they provide limited evidence regarding how such systems behave within live clinical workflows, where patient information may be incomplete, priorities change over time, clinicians disagree, and institutional constraints shape implementation decisions.

The current literature appears to occupy an early stage of the translational pathway. Benchmark and proof-of-concept studies primarily establish technical feasibility, whereas retrospective clinical studies offer greater ecological validity. Active deployment studies provide the strongest evidence for workflow integration, user behavior, accountability, safety, and patient impact, yet such studies remain uncommon. Consequently, the strongest conclusion supported by the current evidence is not that agentic AI is clinically ready, but rather that the field has matured sufficiently to justify prospective deployment-oriented evaluation. The next phase of research should therefore focus on validating performance under real-world conditions rather than continuing to rely predominantly on controlled experimental environments.

7.2. Interdependence of the Seven Principles

The seven principles should not be interpreted as independent checklist items. The observed co-occurrence patterns suggest that agentic systems operate as integrated sociotechnical architectures in which multiple principles interact simultaneously. Planning frequently depends on tool use; tool use requires reliable orchestration and auditability; multi-agent collaboration can support critique, validation, and consensus formation while simultaneously creating accountability challenges; and autonomy becomes clinically acceptable only when bounded by oversight and safety mechanisms.

This interdependence has important implications for evaluation. Assessing individual capabilities in isolation may fail to reveal risks that emerge when multiple capabilities interact. For example, a tool-using agent may perform well when each component is tested independently but fail when tool outputs conflict, data are incomplete, workflow interruptions occur, or escalation procedures are poorly defined. Similarly, strong autonomy without corresponding oversight may increase operational efficiency while simultaneously increasing governance risk. Future evaluations should therefore assess not only whether individual principles are present, but also whether they function together in ways that are clinically safe, explainable, accountable, and governable.

7.3. Broader Implications for Research, Practice, and Regulation

For researchers, the findings suggest that the next stage of agentic healthcare AI evaluation should move beyond task completion and final-answer accuracy. Future studies should examine reasoning validity, tool-failure recovery, escalation thresholds, clinician workload, privacy and security safeguards, demographic subgroup performance, auditability, and post-deployment monitoring. These outcomes are essential if agentic systems are to progress from technical prototypes toward clinically responsible technologies.

For healthcare organizations, the findings support a cautious and evidence-based implementation strategy. Agentic AI may provide value across clinical decision support, patient education, monitoring, documentation, and administrative workflows, but deployment decisions should remain aligned with the maturity of the supporting evidence. Systems validated only through benchmark or proof-of-concept evaluations should generally be regarded as experimental or assistive technologies rather than clinically validated autonomous systems.

For regulators and policymakers, the review highlights the need for evaluation frameworks capable of addressing dynamic, multi-step, tool-using, and potentially adaptive AI systems. Existing healthcare AI governance frameworks already emphasize transparency, human oversight, safety, accountability, and lifecycle monitoring [26]-[28]. The evidence synthesized in this review suggests that these requirements become even more important in the context of agentic AI because system behavior may evolve across reasoning steps, tool interactions, and agent-to-agent coordination. Consequently, future governance approaches may need to evaluate not only final outputs but also the processes through which agentic systems arrive at those outputs, particularly in high-stakes clinical environments.

8. Conclusion

This systematic review makes four contributions beyond prior work. First, it synthesizes recent peer-reviewed empirical evidence on agentic AI across Clinical Decision Support (CDS), Patient Monitoring and ICU-related applications, and Administrative Workflows. Second, it introduces an evidence-informed seven-principle synthesis framework that integrates established concepts from agentic AI, healthcare governance, and clinical AI ethics into a unified deployment-readiness perspective. Third, it maps the distribution and co-occurrence of these principles across the empirical literature, revealing how technical and governance-related dimensions are represented in current research. Fourth, it identifies specific deployment-readiness gaps that should be addressed before agentic AI systems can be responsibly translated into clinical practice.

The findings reveal a clear imbalance in the current evidence base. Capability-oriented principles, including Autonomy and Goal-Directed Behavior, Planning and Multi-Step Reasoning, Tool Use and Orchestration, and Multi-Agent Collaboration, are comparatively well represented and frequently evaluated together. Across multiple healthcare domains, studies report encouraging evidence of technical feasibility and task-level performance, including applications in oncology, rare-disease diagnosis, radiomics-supported therapy prediction, patient monitoring, and multi-specialty clinical reasoning. In contrast, governance-oriented principles receive substantially less empirical attention. Human-in-the-Loop (HITL) Oversight appears in only 32% of included studies, while Safety, Alignment, and Compliance appears in only 24%. Furthermore, none of the included studies reported demographic-stratified performance outcomes. Collectively, these findings suggest that technical capabilities are advancing more rapidly than the governance evidence required to support safe, accountable, and equitable deployment.

Several research priorities emerge directly from these findings. First, future studies should place greater emphasis on prospective clinical evaluations that assess patient outcomes rather than relying primarily on task-level performance metrics. Second, demographic subgroup analysis should become a routine component of evaluation to support fairness and equity assessment. Third, greater methodological consistency is needed through standardized evaluation protocols that facilitate comparison across studies and application domains. Fourth, more rigorous investigation of HITL oversight mechanisms is required, including escalation thresholds, intervention strategies, and responsibility allocation across different levels of clinical risk. Fifth, future work should develop and validate safety-assessment methodologies that align with emerging healthcare AI governance requirements, including those reflected in the EU AI Act and evolving FDA guidance for adaptive AI and machine learning systems [26]-[28].

Overall, the evidence indicates that the central question is no longer whether agentic AI systems can perform healthcare-related tasks with increasing autonomy. Rather, the more pressing challenge is whether the field can generate the governance, safety, oversight, and validation evidence necessary to support responsible deployment. The seven-principle framework presented in this review provides a structured basis for assessing that progress and for guiding future research toward clinically meaningful and deployment-oriented evaluation.

Author Contributions: Conceptualization: C.P.; Methodology: C.P., A.S., and M.L.; Writing - original draft preparation: C.P. and A.S.; Writing - review and editing: C.P., M.L., and A.S.; Visualization: C.P.; Supervision: M.L.; Project administration: C.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new dataset was created. This article synthesizes published studies listed in the references. Supplementary materials can be found on the GitHub page here - c-prakash/Agentic-AI-Seven-Principle-Framework.

Acknowledgments: During the preparation of this research article, the author(s) used Grammarly and ChatGPT to enhance the readability by improving sentence structure, transitions, and grammar. After using these services, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] S. Balakrishna and V. Kumar Solanki, "A Comprehensive Review on AI-Driven Healthcare Transformation," *Ing. Solidar.*, vol. 20, no. 2, pp. 1–30, May 2024, doi: 10.16925/2357-6014.2024.02.07.

[2] J. Qiu *et al.*, "LLM-based agentic systems in medicine and healthcare," *Nat. Mach. Intell.*, vol. 6, no. 12, pp. 1418–1420, Dec. 2024, doi: 10.1038/s42256-024-00944-1.

[3] A. Alnattah, M. Jajroudi, S. A. N. Fadafen, M. N. Manzari, and S. Eslami, "Artificial Intelligence in Clinical Decision-Making: A Scoping Review of Rule-Based Systems and Their Applications in Medicine," *Cureus*, Aug. 2025, doi: 10.7759/cureus.91333.

[4] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[5] R. Hirani *et al.*, "Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities," *Life*, vol. 14, no. 5, p. 557, Apr. 2024, doi: 10.3390/life14050557.

[6] M. M. Raza, K. P. Venkatesh, and J. C. Kvedar, "Generative AI and large language models in health care: pathways to implementation," *npj Digit. Med.*, vol. 7, no. 1, p. 62, Mar. 2024, doi: 10.1038/s41746-023-00988-4.

[7] W. Zhao *et al.*, "An agentic system for rare disease diagnosis with traceable reasoning," *Nature*, vol. 651, no. 8106, pp. 775–784, Mar. 2026, doi: 10.1038/s41586-025-10097-9.

[8] H. Eguia, C. L. Sánchez-Bocanegra, F. Vinciarelli, F. Alvarez-Lopez, and F. Saigí-Rubió, "Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review," *J. Med. Internet Res.*, vol. 26, p. e55315, Sep. 2024, doi: 10.2196/55315.

[9] M. H. Shahin, S. Goswami, S. Lobentanzer, and B. W. Corrigan, "Agents for Change: Artificial Intelligent Workflows for Quantitative Clinical Pharmacology and Translational Sciences," *Clin. Transl. Sci.*, vol. 18, no. 3, Mar. 2025, doi: 10.1111/cts.70188.

[10] 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*, vol. 14, pp. 4840–4863, 2026, doi: 10.1109/ACCESS.2026.3651218.

[11] S. Johri *et al.*, "An evaluation framework for clinical use of large language models in patient interaction tasks," *Nat. Med.*, vol. 31, no. 1, pp. 77–86, Jan. 2025, doi: 10.1038/s41591-024-03328-5.

[12] D. Ferber *et al.*, "Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology," *Nat. Cancer*, vol. 6, no. 8, pp. 1337–1349, Jun. 2025, doi: 10.1038/s43018-025-00991-6.

[13] Q. Liu *et al.*, "EvoMDT: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer," *npj Digit. Med.*, vol. 9, no. 1, p. 124, Jan. 2026, doi: 10.1038/s41746-025-02304-8.

[14] S. Bedi *et al.*, "Testing and Evaluation of Health Care Applications of Large Language Models," *JAMA*, vol. 333, no. 4, p. 319, Jan. 2025, doi: 10.1001/jama.2024.21700.

[15] B. G. Collaco *et al.*, "The role of agentic artificial intelligence in healthcare: a scoping review," *npj Digit. Med.*, vol. 9, no. 1, p. 345, Mar. 2026, doi: 10.1038/s41746-026-02517-5.

[16] M. Abou Ali, F. Dornaika, and J. Charafeddine, "Agentic AI: a comprehensive survey of architectures, applications, and future directions," *Artif. Intell. Rev.*, vol. 59, no. 1, p. 11, Nov. 2025, doi: 10.1007/s10462-025-11422-4.

[17] B. Njei *et al.*, "Artificial intelligence agents in healthcare research: A scoping review," *PLoS One*, vol. 21, no. 2, p. e0342182, Feb. 2026, doi: 10.1371/journal.pone.0342182.

[18] P. N. Srinivasu, G. L. Aruna Kumari, S. Ahmed, and A. Alhumam, "Exploring Agentic AI in Healthcare: A Study on Its Working Mechanism," *Front. Med.*, vol. 12, Jan. 2026, doi: 10.3389/fmed.2025.1753443.

[19] S. Hosseini and H. Seilani, "The role of agentic AI in shaping a smart future: A systematic review," *Array*, vol. 26, p. 100399, Jul. 2025, doi: 10.1016/j.array.2025.100399.

[20] Y. Hao *et al.*, "Personalizing prostate cancer education for patients using an EHR-Integrated LLM agent," *npj Digit. Med.*, vol. 8, no. 1, p. 770, Dec. 2025, doi: 10.1038/s41746-025-02166-0.

[21] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[22] S. Yao and others, "ReAct: Synergizing Reasoning and Acting in Language Models," *OpenReview.net*. 2023. [Online]. Available: https://openreview.net/forum?id=WE_vluYUL-X
[23] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, Hoboken, 2021.
[24] T. Schick and others, "Toolformer: Language Models Can Teach Themselves to Use Tools," in *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://dl.acm.org/doi/10.5555/3666122.3669119
[25] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving Factuality and Reasoning in Language Models through Multiagent Debate," in *ICML'24: Proceedings of the 41st International Conference on Machine Learning*, May 2024, pp. 11733–11763. [Online]. Available: http://arxiv.org/abs/2305.14325
[26] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)." 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
[27] WHO guidance and World Health Organization, "Ethics and governance of artificial intelligence for health," *World Health Organization*, 2021. https://www.who.int/publications/i/item/9789240029200
[28] G. Joshi, A. Jain, S. R. Araveeti, S. Adhikari, H. Garg, and M. Bhandari, "FDA-Approved Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: An Updated Landscape," *Electronics*, vol. 13, no. 3, p. 498, Jan. 2024, doi: 10.3390/electronics13030498.
[29] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Inform.*, vol. 113, p. 103655, Jan. 2021, doi: 10.1016/j.jbi.2020.103655.
[30] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
[31] S. Mbonihankuye, A. Nkunzimana, and A. Ndagijimana, "Healthcare Data Security Technology: HIPAA Compliance," *Wirel. Commun. Mob. Comput.*, vol. 2019, pp. 1–7, Oct. 2019, doi: 10.1155/2019/1927495.
[32] M. Wooldridge and N. R. Jennings, "Intelligent Agents: Theory and Practice," *Knowl. Eng. Rev.*, vol. 10, no. 2, pp. 115–152, 1995.
[33] S. Yao *et al.*, "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, 2023, pp. 11809–11822. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3666122.3666639
[34] Z. Xi *et al.*, "The rise and potential of large language model based agents: a survey," *Sci. China Inf. Sci.*, vol. 68, no. 2, p. 121101, Feb. 2025, doi: 10.1007/s11432-024-4222-0.
[35] L. Wang *et al.*, "A survey on large language model based autonomous agents," *Front. Comput. Sci.*, vol. 18, no. 6, p. 186345, Dec. 2024, doi: 10.1007/s11704-024-40231-1.
[36] I. Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds Mach.*, vol. 30, no. 3, pp. 411–437, Sep. 2020, doi: 10.1007/s11023-020-09539-2.
[37] O. Freyer, S. Jayabalan, J. N. Kather, and S. Gilbert, "Overcoming regulatory barriers to the implementation of AI agents in healthcare," *Nat. Med.*, vol. 31, no. 10, pp. 3239–3243, Oct. 2025, doi: 10.1038/s41591-025-03841-1.
[38] L. Floridi *et al.*, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds Mach.*, vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/s11023-018-9482-5.
[39] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.
[40] B. Mohapatra, B. Walia, and S. Dash, "A Modular Multi-Agent Framework for Clinical Documentation and Hospital Operations," in *2025 8th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)*, Dec. 2025, pp. 1–5. doi: 10.1109/ACAI68217.2025.11406271.
[41] S. AlZu'Bi, M. Almiani, and Y. Jararweh, "Agentic AI for Healthcare: Solutions to Intelligent Patient Care," in *2025 12th International Conference on Information Technology (ICIT)*, May 2025, pp. 33–38. doi: 10.1109/ICIT64950.2025.11049267.
[42] Q. Li, K. R. Amat, and J. Li, "From Knowledge to Action: An Agentic AI Framework for Diabetes Management," in *2025 IEEE 16th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, Oct. 2025, pp. 395–401. doi: 10.1109/UEMCON67449.2025.11267605.
[43] L. Murali, A. Thankachan, and D. M. Viswanathan, "Agentic AI for Severity Extraction in Clinical Notes: Enhancing Disease Diagnosis Beyond Rule-Based and Traditional ML Models," in *2025 International Conference on Power, Instrumentation, Control, and Computing (PICC)*, Oct. 2025, pp. 1–6. doi: 10.1109/PICC67314.2025.11291255.
[44] Y.-J. Chen, A. Albarqawi, and C.-S. Chen, "Enhancing Clinical Decision-Making: Integrating Multi-Agent Systems with Ethical AI Governance," in *2025 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Aug. 2025, pp. 1–7. doi: 10.1109/CIBCB66090.2025.11177136.
[45] X. Qiao *et al.*, "Parallel Nursing: Enhancing Postoperative Nursing With LLM Agent Systems," *IEEE Trans. Comput. Soc. Syst.*, vol. 13, no. 2, pp. 2002–2011, Apr. 2026, doi: 10.1109/TCSS.2025.3605582.
[46] A. Mukhtar, G. E. Arzu, W. T. Toor, and U. Ali, "Collaborative multi-agent conversational artificial intelligence for clinical support in Parkinson disease," *Parkinsonism Relat. Disord.*, p. 108292, Mar. 2026, doi: 10.1016/j.parkreldis.2026.108292.
[47] R. R. Pai, "Agentic AI Implementation in Healthcare Insurance Industry: A Comprehensive Framework for Automated Claims Processing and Risk Assessment," in *2025 IEEE Madhya Pradesh Section Conference (MPCON)*, Aug. 2025, pp. 812–817. doi: 10.1109/MPCON66082.2025.11256764.
[48] K. Lee *et al.*, "A4SLR: An Agentic Artificial Intelligence-Assisted Systematic Literature Review Framework to Augment Evidence Synthesis for Health Economics and Outcomes Research and Health Technology Assessment," *Value Heal.*, vol. 28, no. 11, pp. 1655–1664, Nov. 2025, doi: 10.1016/j.jval.2025.08.002.

[49] S. R. Shimgekar, S. Vassef, A. Goyal, K. Saha, P. Zonooz, and N. Kumar, "Agentic AI Framework for End-to-End Medical Data Inference," in *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2025, pp. 7807–7810. doi: 10.1109/BIBM66473.2025.11356498.

[50] A. Patel, S. Joseph, C. Bailey, A. Sakharpe, and M. Devarapalli, "Transforming electrophysiology workflows with natural language processing and agentic artificial intelligence," *Hear. Rhythm O2*, vol. 6, no. 10, pp. 1613–1620, Oct. 2025, doi: 10.1016/j.hroo.2025.07.013.

[51] E. Tzanis and M. E. Klontzas, "mAIstro: An open-source multi-agent system for automated end-to-end development of radiomics and deep learning models for medical imaging," *Eur. J. Radiol. Artif. Intell.*, vol. 4, p. 100044, Dec. 2025, doi: 10.1016/j.ejrai.2025.100044.

[52] Z. Yi, J. Liu, M. V. Albert, and T. Xiao, "A Multi-Agent System for Complex Reasoning in Radiology Visual Question Answering," in *2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Dec. 2025, pp. 139–147. doi: 10.1109/JCDL67857.2025.00025.

[53] R. Sathya and A. Valaramathi, "Detection and diagnosis of diabetic retinopathy in retinal fundus images using agentic AI approaches," *Sci. Rep.*, vol. 16, no. 1, p. 3780, Dec. 2025, doi: 10.1038/s41598-025-34016-0.

[54] P. Liu and L. Xiao, "Improving Clinical Decision Support: Architecture Design of a Multi-Agent System based on an Argument Quality Assessment Ontology," in *2025 IEEE 22nd International Conference on Software Architecture (ICSA)*, 2025, pp. 313–323. doi: 10.1109/ICSA65012.2025.00039.

[55] A. Loaiza-Bonilla *et al.*, "Transforming oncology clinical trial matching through neuro-symbolic, multi-agent AI and an oncology-specific knowledge graph: a prospective evaluation in 3804 patients," *ESMO Real World Data Digit. Oncol.*, vol. 12, p. 100706, Jun. 2026, doi: 10.1016/j.esmorw.2026.100706.

[56] K. Linton-Reid, M. Chen, M. B. Martell, J. M. Posma, and E. O. Aboagye, "Radiomics in clinical radiology: advances, challenges, and future directions," *Clin. Radiol.*, vol. 92, p. 107165, Jan. 2026, doi: 10.1016/j.crad.2025.107165.

[57] M. Omar *et al.*, "Impact of patient communication style on agentic AI-generated clinical advice in E-medicine," *Am. J. Med.*, vol. 139, no. 4, pp. 437–444, Apr. 2026, doi: 10.1016/j.amjmed.2025.12.027.

[58] J. Wang, K. Arora, D. Swoboda, and A. Nazha, "Artificial intelligence system for delivering interactive and verifiable guideline-based clinical decision support in hematology," *Blood*, vol. 146, no. Supplement 1, pp. 2578–2578, Nov. 2025, doi: 10.1182/blood-2025-2578.

[59] O. Freyer *et al.*, "The regulation of artificial intelligence in intensive care units: from narrow tools to generalist systems," *npj Digit. Med.*, vol. 9, no. 1, p. 246, Mar. 2026, doi: 10.1038/s41746-026-02535-3.

[60] A. J. Goodell, S. N. Chu, D. Rouholiman, and L. F. Chu, "Large language model agents can use tools to perform clinical calculations," *npj Digit. Med.*, vol. 8, no. 1, p. 163, Mar. 2025, doi: 10.1038/s41746-025-01475-8.

[61] S. Schmidgall *et al.*, "AgentClinic: a multimodal benchmark for tool-using clinical AI agents," *npj Digit. Med.*, Apr. 2026, doi: 10.1038/s41746-026-02674-7.

[62] B. Babic, I. Glenn Cohen, A. D. Stern, Y. Li, and M. Ouellet, "A general framework for governing marketed AI/ML medical devices," *npj Digit. Med.*, vol. 8, no. 1, p. 328, May 2025, doi: 10.1038/s41746-025-01717-9.

[63] M. Moritz, E. Topol, and P. Rajpurkar, "Coordinated AI agents for advancing healthcare," *Nat. Biomed. Eng.*, vol. 9, no. 4, pp. 432–438, Apr. 2025, doi: 10.1038/s41551-025-01363-2.

[64] J. C. L. Ong *et al.*, "Innovating global regulatory frameworks for generative AI in medical devices is an urgent priority," *npj Digit. Med.*, vol. 9, no. 1, p. 364, Mar. 2026, doi: 10.1038/s41746-026-02552-2.