

Agentic AI in Healthcare: Bridging the Gap Between Computational Promise and Clinical Evidence

Yunguo Yu

yuyunguo@gmail.com

Zyter|TruCare <https://orcid.org/0009-0006-2003-3813>

Systematic Review

Keywords: Artificial Intelligence, Multi-Agent Systems, Single-Agent Systems, Agentic AI, Clinical Decision Support Systems, Scoping Review, Healthcare Delivery, Patient Safety, Algorithmic Bias, Implementation Science, Medical Informatics, Ethics, Global Health

Posted Date: April 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9374197/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Agentic AI in Healthcare: Bridging the Gap Between Computational Promise and Clinical Evidence

Yunguo Yu^{1*}

^{1*} AI Innovation & Prototyping, Zyter|TruCare, Vienna, VA, USA.

Corresponding author(s). E-mail(s): yuyunguo@gmail.com;

Abstract

Background: Agentic AI systems are increasingly proposed for healthcare applications, yet the evidence base distinguishing computational promise from clinical reality remains poorly characterised. Single-agent systems offer efficiency for routine diagnostics; multi-agent systems promise robustness for complex care. Both face barriers in safety, accountability, and equitable deployment.

Methods: We conducted a PRISMA-ScR scoping review synthesising evidence from 161 studies (January 2018–October 2024, with selective early-access coverage through April 2026) retrieved from PubMed, IEEE Xplore, arXiv, Google Scholar, and Scopus. Evidence certainty was graded using an adapted GRADE-informed framework appropriate for heterogeneous clinical and simulation evidence. Given substantial heterogeneity across architectures, tasks, and outcome measures, quantitative pooling was not appropriate; we employed structured evidence mapping and narrative synthesis. A pragmatic, deployment-focused definition of “agent” was adopted and extended with a five-level Agentic Capability Spectrum (Levels 0–4) to preserve discriminative power.

Results: High-certainty evidence supports selected single-agent systems in specialised diagnostic domains (94.5% accuracy in retinal screening; AUC 0.96 in skin-cancer classification). Very low-certainty evidence from simulation studies suggests potential coordination advantages for multi-agent systems, with no confirmed clinical deployment. Multi-agent systems require substantially higher computational resources and introduce coordination latency (200–500 ms in simulation). Cross-cutting barriers include algorithmic bias in one commercial population-health algorithm (Moderate-certainty; Obermeyer et al., 2019; generalisability uncertain), unclear liability frameworks, and workflow-integration failures. Evidence is predominantly from high-income countries (87% of studies; descriptive evidence-mapping finding).

Conclusions: Single-agent systems demonstrate validated clinical utility in constrained tasks, whereas multi-agent systems remain experimental. Priorities include large-scale clinical trials for multi-agent architectures, standardised safety frameworks, risk-based regulatory pathways, and equity-focused global deployment strategies. As this synthesis was conducted by a single reviewer, all findings represent a **preliminary structured synthesis** requiring independent replication before informing clinical guideline development.

Keywords: Artificial Intelligence, Multi-Agent Systems, Single-Agent Systems, Agentic AI, Clinical Decision Support Systems, Scoping Review, Healthcare Delivery, Patient Safety, Algorithmic Bias, Implementation Science, Medical Informatics, Ethics, Global Health

1 Introduction

The integration of artificial intelligence into healthcare represents one of the most consequential technological transformations in medical practice. As AI systems evolve from traditional decision support tools to autonomous entities capable of perception, decision-making, and action [1], understanding the distinction between different architectural paradigms becomes crucial for clinical implementation.

1.1 Defining Agentic Systems in Healthcare

Conceptual Framework

In this review, we provide a comprehensive taxonomy of agentic systems in healthcare, distinguishing between different levels of autonomy, coordination, and deployment. We adopt a pragmatic, deployment-oriented

definition of “agent” aligned with clinical implementation realities rather than strict adherence to classical multi-agent systems theory [2], for two critical reasons:

Important Terminological Note: This paper uses “agent” in two distinct ways. Throughout this section, we clarify the difference between classical MAS theory (emphasising autonomous goal-directed systems) and pragmatic clinical definition (emphasising deployment architecture). When we classify systems as “agentic,” we refer to their deployment architecture (single vs. multi-module), not necessarily their level of autonomy. To avoid conceptual ambiguity, we explicitly distinguish between *agentic behaviour* (autonomy and goal-directed reasoning) and *agentic architecture* (modular deployment structure). This pragmatic approach enhances clinical applicability but represents a departure from classical computer science terminology (Table 1).

Table 1: Terminology mapping between classical MAS theory and pragmatic clinical usage.

Classical MAS Term	Pragmatic Clinical Usage	Example
Agent (autonomous)	AI module or pipeline involved in perception, decision output, and workflow integration (may be semi-autonomous)	Retinal disease detection pipeline (single-agent)
Multi-agent system	Coordinated network of multiple AI modules (may be autonomous or orchestrated workflow)	Surgical robotics coordination (pragmatic multi-agent)
Tool use	External API or module calls within a workflow	Imaging pipeline accessing PACS for retrieval
Coordination protocol	Orchestrator or controller coordinating modules	Weighted-voting diagnostic ensemble

Agentic Capability Spectrum

To add discriminative power beyond the binary agent/non-agent label, we introduce a graduated **Agentic Capability Spectrum** (Table 2). This spectrum aligns with the agent-AI literature [3, 4] and tags every anchor system in this review with a level. This spectrum is intended as a descriptive taxonomy of deployment maturity, not a formal measure of autonomy. The title claim of “Agentic AI” is most directly applicable to Level ≥ 2 systems; Level 0–1 systems are included for comparative context to illustrate the full deployment pipeline.

Table 2: Agentic Capability Spectrum. Levels 0–4 classify systems by their degree of autonomy, planning, and coordination. Level assignments for anchor systems reviewed in this paper are given in the right column. † *Systems at Level 0 – 1 are included for comparative context only.*

Level	Label	Criteria	Anchor Example
0	Passive tool	No autonomy; output only on explicit query	Rule-based drug-alert CDS†
1	Reactive classifier	Perception + single-step output; no planning or adaptation	CNN retinal screening (De Fauw et al., 2018); CheXNet; skin-cancer classifier (Esteva et al., 2017)†
2	Adaptive decision support	Perception + context-dependent output selection; limited planning	ChatExosome (RAG + model selection); Aidoc triaging platform
3	Goal-directed agent	Autonomous goal formulation, planning, tool use, and environment interaction	ReAct/Reflexion clinical agents; ClinicalAgent [5]; MedAgent-Zero-style systems
4	Coordinated multi-agent	Multiple Level ≥ 2 modules with inter-agent communication and distributed decision-making	MDAgents; Agent Hospital; MA-HRL diagnostic dialogue

Level Assignment Classification Protocol

To ensure reproducibility of Level assignments, systems were classified using the following sequential decision criteria applied to published descriptions of system architecture and behaviour: (1) Does the system generate outputs only on explicit human query, with no autonomous action or state maintenance? If Yes → Level 0. (2) Does the system perceive and classify inputs in a single step, without planning, context-adaptive selection, or memory? If Yes → Level 1. (3) Does the system select among multiple response strategies, adapt output based on contextual state, or integrate retrieval components? If Yes → at least Level 2. (4) Does the system formulate goals, generate multi-step plans, invoke external tools iteratively, or interact with its environment to satisfy objectives? If Yes → Level 3. (5) Does the system involve two or more independently operating modules that communicate state, decisions, or partial results with each other? If Yes → Level 4. Borderline cases were resolved conservatively (assigned to the lower level). Level assignments and rationales for all anchor systems, including borderline cases, are documented in Supplementary Table S1. This protocol is provided to enable readers to apply and replicate the taxonomy to systems not reviewed here.

Justification for Pragmatic Approach

First, classical MAS theory defines an agent as having autonomy, reactivity, proactiveness, and social ability [2, 6–8], with goal-directed action independent of human control. However, in clinical practice, most deployed AI systems operate as semi-autonomous decision support tools rather than fully autonomous agents. Applying classical MAS criteria would exclude the vast majority of clinically implemented systems that provide actual patient care value.

Second, clinical deployment reality prioritises modularity and workflow integration over theoretical agent autonomy. Systems like CheXNet and Aidoc function as modular deep learning pipelines or workflow tools that provide clinical decision support, not autonomous goal-directed action. Aidoc’s platform has limited clinical deployment, but available evidence remains confined to non-peer-reviewed reports; therefore it is classified here as pilot deployment with Very Low-certainty evidence. A pragmatic classification that focuses on deployment architecture (single vs. multi-module coordination) better reflects current clinical practice and implementation challenges.

Therefore, we define a “healthcare agent” pragmatically as: an AI system or module with three core deployment characteristics: (1) perception of clinical data, (2) generation of decision support or diagnostic outputs, and (3) integration into clinical workflows with varying degrees of human oversight. This framework builds on established AI agent theory while adapting to healthcare deployment realities.

Distinguishing Classical vs. Pragmatic Definitions

- **Classical MAS Theory:** True agents exhibit autonomous, goal-directed behaviour without human intervention (e.g., autonomous agents navigating unknown environments). Most current healthcare systems do not meet this strict criterion.
- **Pragmatic Clinical Definition:** “Agents” are AI modules that provide decision support or automated outputs, regardless of full autonomy. This definition encompasses systems that clinicians actually use in practice, even if they operate as supervised classification pipelines or semi-autonomous tools.

We acknowledge this terminological choice represents a departure from classical MAS literature. This decision reflects the clinical reality that most healthcare AI operates as decision support rather than fully autonomous systems, and that distinguishing between single-module and multi-module deployments provides more practical guidance for healthcare organisations (Table 3).

Table 3: Taxonomy of Agentic Systems in Healthcare. This table describes structural characteristics of the two primary deployment paradigms. For granular level assignments, see Table 2 (Agentic Capability Spectrum). Single-agent systems span Levels 0–2; multi-agent systems correspond to Level 4; hybrid systems include Level 3 architectures that may operate in either mode.

Characteristic	Single-Agent Systems	Multi-Agent Systems
Autonomy Level	Independent operation	Coordinated collaboration
Decision Making	Individual reasoning	Consensus or distributed reasoning
Communication	No inter-agent communication	Structured inter-agent protocols
Task Complexity	Single focused tasks	Multi-objective coordination
Clinical Examples	Diagnostic imaging analysis	Surgical robotics coordination

Critical Distinctions from Traditional AI Systems

- **Traditional Clinical Decision Support (CDS):** Passive systems that provide recommendations without autonomous execution (e.g., drug interaction alerts)
- **Agentic Systems:** Active systems that can perceive, decide, and act with varying degrees of autonomy (e.g., autonomous image interpretation with clinical decision output)

Tool-Using vs. Autonomous Agents

- **Tool-Using Agents:** Systems that use external tools or APIs but maintain central control (e.g., diagnostic systems accessing external databases)
- **Autonomous Agents:** Systems capable of independent reasoning and action without requiring external tool activation

Coordinated vs. Emergent Multi-Agent behaviour

- **Coordinated Multi-Agent:** Explicitly designed collaboration with predefined communication protocols (e.g., surgical robotics with leader–follower coordination)
- **Emergent Multi-Agent:** Self-organizing systems where collective behaviour arises from individual agent interactions (e.g., swarm intelligence in pandemic modelling)

Single-Agent Systems (Pragmatic Definition)

We define single-agent systems as AI modules or pipelines that independently process clinical data and generate outputs without coordination with other AI modules. Using our pragmatic definition, these systems may function as supervised classification pipelines or decision support tools rather than fully autonomous agents. Key characteristics include:

- Independent processing and decision-making within a single module
- Focused on single clinical tasks or objectives
- Typically deployed as static deep learning models or rule-based systems
- Examples: Retinal disease detection pipelines [9], skin cancer classifiers [10], surgical tool collaboration systems [11], and exosome-based cancer diagnostics [12].

Note: Under classical MAS theory, many of these systems would be classified as “passive decision support” rather than “agents” due to lack of autonomous goal-directed action. We classify them as single-agent systems pragmatically to reflect their deployment architecture.

Multi-Agent Systems (Pragmatic Definition)

We define multi-agent systems as coordinated networks or workflows involving multiple AI modules that communicate or distribute tasks to achieve complex healthcare objectives. Using our pragmatic definition, these may be: (1) true multi-agent systems with autonomous coordination, or (2) multi-module workflows with coordinated pipelines. **Key characteristics include:**

- Multiple AI modules with communication or coordination mechanisms
- Distribution of tasks and responsibilities across modules
- May involve true agent coordination or simply multi-step pipeline coordination
- Examples: Pandemic contact tracing frameworks [13], multi-module diagnostic workflows [14, 15], trauma tracking agents [16], and radiotherapy planning agents [17].

Note: Many systems classified here as “multi-agent” may not meet classical MAS criteria for true agent autonomy. We include them based on deployment architecture involving multiple coordinated modules, which reflects clinical implementation reality.

Clinical Significance of the Distinction

The distinction between single and multi-agent systems is not merely technical but has profound clinical implications:

- **Decision Complexity:** Single agents excel at focused decisions; multi-agents handle multifaceted clinical scenarios
- **Safety Considerations:** Single agents have clearer accountability; multi-agents introduce emergent safety challenges

- **Implementation Barriers:** Single agents integrate more easily; multi-agents require coordination infrastructure

Evidence Quality Consideration

In this review, we consistently distinguish between the deployment status of agentic systems to provide clarity on evidence quality:

- **Research Prototypes:** Lab-validated systems with no clinical deployment (high technical evidence, low clinical evidence)
- **Simulation Studies:** Computational models without real-world testing (moderate technical evidence, very low clinical evidence)
- **Pilot Deployments:** Limited clinical use in controlled settings (emerging clinical evidence)
- **Production Systems:** Routine clinical use across multiple sites (high clinical evidence)

This framework ensures that our analysis accurately reflects both the technical capabilities and clinical readiness of different agentic approaches.

Terminology Quick Reference

To reduce ambiguity throughout this paper, the following conventions are applied consistently:

- **Agent_{pragmatic}:** An AI module meeting the three deployment criteria (perception, clinical output, workflow integration), regardless of autonomy level. Used throughout Results and Discussion.
- **Agent_{classical}:** A system exhibiting autonomous, goal-directed behaviour as defined in classical MAS theory [18]. Explicitly signalled when this sense is intended.
- **Agentic behaviour:** Autonomy and goal-directed reasoning (classical MAS sense).
- **Agentic architecture:** Modular deployment structure (pragmatic sense).
- **Single-agent** (pragmatic): Levels 0–2 on the Agentic Capability Spectrum; one primary AI module; no inter-module coordination.
- **Multi-agent** (pragmatic): Level 4; two or more AI modules with inter-agent communication or distributed decision-making.
- **Coordination:** Structured inter-module communication with explicit protocols. Distinct from **tool use** (external API call from a single control architecture).
- **GRADE** (adapted): Evidence certainty labels used heuristically for heterogeneous evidence types; not equivalent to canonical GRADE clinical ratings for intervention effects (see Methods for adaptation rules).

1.2 Clinical Significance and Research Questions

The significance of this analysis extends beyond technical comparison to address fundamental questions about AI’s role in healthcare: How can we ensure equitable access to AI-enhanced care? What accountability frameworks are needed for autonomous medical systems? How should clinicians be trained to collaborate effectively with AI agents? What safety mechanisms are required to prevent patient harm?

This review synthesises current evidence on both paradigms, drawing from 161 studies published between January 2018 and October 2024 (with select early-access articles through April 2026 included for contextual completeness), encompassing primary peer-reviewed empirical studies, secondary syntheses, and selectively retained non-peer-reviewed contextual sources. Our analysis reveals that single-agent systems have achieved clinical validation in specialised domains, while multi-agent systems remain largely experimental despite promising performance gains. However, both face substantial barriers in ethical implementation, regulatory compliance, and clinical integration.

Research Gap and Contribution Statement

What is known: Selected single-agent AI systems (e.g., retinal screening, skin cancer classification) have achieved validated clinical utility in narrow, high-volume diagnostic tasks. Algorithmic bias and context blindness are established failure modes. Multi-agent systems exist as conceptual architectures with theoretical coordination benefits.

What is unknown (research gap): Whether multi-agent systems offer clinically meaningful advantages over single-agent systems in real-world deployments; which implementation factors predict success or failure across diverse healthcare settings; how these technologies perform equitably across geographic, cultural, and resource-constrained contexts; and what regulatory and safety frameworks are sufficient for autonomous or semi-autonomous clinical AI.

What this review contributes: To the best of our knowledge, this is the first scoping review to (1) apply a structured five-level Agentic Capability Spectrum to map 161 heterogeneous healthcare AI systems onto a common deployment taxonomy, (2) use an adapted GRADE-based certainty framework to make the clinical evidence gap between single-agent and multi-agent systems explicit and auditable, and (3) synthesise safety, equity, and implementation-science evidence into an integrated clinical decision framework.

2 Methods

2.1 Search Strategy and Study Selection

We conducted a scoping review following PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guidelines [19]. Given the substantial heterogeneity in model architectures, clinical tasks, outcome measures, and validation settings across included studies, a quantitative meta-analysis was not appropriate. Instead, we employed structured evidence synthesis with GRADE-based certainty assessment, evidence mapping, and narrative synthesis to address the research questions. The search strategy was developed in consultation with medical librarians and AI healthcare specialists.

Search Strategy

Comprehensive database searches were conducted using predefined Boolean queries in PubMed, IEEE Xplore, arXiv, Google Scholar, and Scopus from January 2018 to October 2024, with select early-access articles through April 2026 included for contextual completeness. Search terms were combined using Boolean operators:

- (“single-agent AI” OR “autonomous agent”) AND (“healthcare” OR “medical” OR “clinical”)
- (“multi-agent system” OR “collaborative AI”) AND (“healthcare” OR “medicine” OR “clinical”)
- (“swarm intelligence” OR “distributed AI”) AND (“healthcare” OR “medical”)
- (“large language model” OR “LLM”) AND (“healthcare” OR “medical” OR “clinical”)

Study Selection Process

Title/abstract screening and full-text assessment were conducted by a single reviewer (Y.Y.) using a pre-defined checklist of inclusion/exclusion criteria. To mitigate single-reviewer bias, a 20% random sample of excluded studies was re-screened at a 4-week interval; intra-rater agreement was $\kappa = 0.88$ (indicating strong self-consistency). This single-reviewer approach is a recognised limitation, consistent with scoping review methodology where resources are constrained. **Findings should therefore be interpreted as a preliminary structured synthesis requiring independent replication before informing clinical guideline development.** An independent secondary reviewer for at least 20–30% of included studies is a prerequisite for definitive guideline use.

PRISMA Flow Diagram Results (Figure 1)

- **Identification:** Initial search yielded 4,847 unique records across all databases
- **Screening:** After title/abstract screening, 1,942 records were excluded (reasons: not healthcare AI [n=892], not agentic systems [n=567], duplicate content [n=483])
- **Eligibility:** Full-text assessment of 905 articles resulted in exclusion of 744 (reasons: no quantitative outcomes [n=401], purely technical [n=198], case reports [n=89], language other than English [n=56])
- **Included:** Final analysis included 161 studies (Supplementary Table S1). Of these, 146 (90.7%) were primary peer-reviewed empirical studies (including 9 RCTs and 6 prospective studies), 7 were simulation/modelling studies, 4 were conference papers, 2 were secondary syntheses retained for contextual comparison, and 2 were non-peer-reviewed sources (arXiv preprints) included for contextual description only, assigned Very Low-certainty evidence, and excluded from clinical-effectiveness estimates. By paradigm: 79 addressed single-agent systems, 41 addressed multi-agent systems, 29 addressed hybrid/cross-paradigm topics, and 12 were paradigm-agnostic or cross-cutting. Supplementary Table S1 provides the study-level evidence classification for all 161 studies, including study type, paradigm, deployment stage, adapted GRADE certainty level, and contribution category, enabling full audit of GRADE assignments.

Inclusion Criteria

- Peer-reviewed empirical studies, systematic reviews, meta-analyses, and clinical trials
- Publications from 2018–2024 focusing on AI agents in healthcare
- Studies reporting quantitative performance metrics (accuracy, sensitivity, specificity, AUC) OR clinical outcomes (mortality, morbidity, quality of life)

PRISMA 2020 Flow Diagram

Single-Agent and Multi-Agent AI in Healthcare



Fig. 1: PRISMA 2020 flow diagram showing the comprehensive search and study selection process. Initial database searches yielded 4,847 unique records. After title/abstract screening and full-text assessment, 161 studies met inclusion criteria for qualitative synthesis and quantitative analysis. The diagram illustrates the screening process with exclusion reasons at each stage.

- English language publications
- **operationalised definition of “agentic” for inclusion:** A study was eligible if the AI system it described met all three of the following deployment criteria: (1) *perception* of clinical data (image, text, signal, or structured record); (2) *generation of clinical outputs* (diagnosis, recommendation, triage decision, or drug dosing); and (3) *integration* into a clinical or research workflow with at least nominal human oversight. Systems meeting criteria (1)–(3) but lacking autonomous goal-directed behaviour were included as Level 0–1 comparative context under the Agentic Capability Spectrum; systems meeting criteria (1)–(3) with planning, tool use, or multi-module coordination were classified as Level 2–4 agentic systems and constitute the primary analytic population.

Exclusion Criteria

- Non-peer-reviewed sources (industry whitepapers, blog posts, preprints without peer review). *Exception:* arXiv preprints providing architectural evidence for emerging multi-agent systems were selectively retained as non-peer-reviewed contextual sources, automatically assigned Very Low-certainty evidence, and excluded

from clinical-effectiveness claims (see Methodological Transparency, below). One industry whitepaper (Aidoc) was retained on the same basis and does not contribute to clinical-effectiveness estimates (total included studies: 161).

- Pre-2018 publications unless recognised as seminal works (e.g., Russell & Wooldridge textbooks)
- Purely technical papers without healthcare application
- Conference abstracts without full methodology
- Studies with insufficient data for quality assessment

Rationale for Including Non-Clinical Evidence for Multi-Agent Systems

Consistent with PRISMA-ScR scoping review methodology, this review includes simulation, pre-clinical, and pilot evidence for multi-agent systems even though no multi-agent system has achieved broad production deployment. This decision was made a priori for three reasons: (1) Scoping reviews aim to map the full extent of an evidence field, including emerging and pre-clinical literature, not only mature clinical evidence; (2) Multi-agent systems represent an architecturally distinct and theoretically important category for which absence from this synthesis would obscure a critical evidence gap; and (3) Including simulation evidence alongside an explicit Very Low-certainty GRADE assignment enables direct, transparent comparison of evidence maturity between paradigms—a central contribution of this synthesis. Readers should note that inclusion of non-clinical evidence for multi-agent systems does not imply equivalence with the high-certainty clinical evidence available for single-agent systems.

Quality Assessment and Data Synthesis

Study quality was evaluated using the Cochrane Risk of Bias tool for clinical trials and the Critical Appraisal Skills Programme (CASP) framework for observational and quasi-experimental designs [20]. Evidence was graded using GRADE criteria [21]:

- **High quality:** RCTs with low risk of bias, large effects, or no plausible confounders
- **Moderate quality:** RCTs with limitations or high-quality observational studies with consistent findings
- **Low quality:** Observational studies with serious limitations or inconsistent results
- **Very low quality:** Case series, expert opinion, or simulation studies without clinical validation

Adapted GRADE rules for non-standard evidence types: Because GRADE was originally designed for rating certainty in clinical intervention effects, it does not directly apply to all evidence types in a scoping review. We adapted GRADE to reflect relative evidence strength across heterogeneous study types, including simulation and technical validation studies; these labels should be interpreted as heuristic indicators rather than strictly equivalent to clinical GRADE ratings. The specific adaptations are as follows: (1) Mathematical modelling and simulation studies: Very Low certainty regardless of methodological rigor (no clinical validation); (2) Conference proceedings: Moderate certainty at most, downgraded to Low if lacking external validation; (3) arXiv preprints and industry documents: Very Low certainty, used only for contextual architectural description, not clinical-effectiveness claims; (4) Single-study prospective validation: Moderate certainty (upgraded from Low for large effect size, not upgraded to High without independent replication); (5) Multiple consistent prospective studies or production deployments: High certainty if bias risk is low; (6) Secondary syntheses: graded by the certainty of the underlying evidence synthesised. We note that applying GRADE vocabulary to geographic-distribution observations is category-inappropriate; such findings are presented as descriptive evidence-mapping results only, not as GRADE-rated clinical outcomes.

GRADE Applicability: Intervention Effects Versus Evidence Mapping

A fundamental limitation of applying GRADE terminology to a scoping review is that GRADE was designed to rate certainty in the magnitude of specific intervention *effects*—not to summarise the breadth or maturity of an evidence landscape. To reduce category confusion, this review distinguishes between two epistemic modes: (a) *evidence certainty* (GRADE vocabulary; applied only to clinical-effectiveness claims involving directional outcome estimates); and (b) *evidence maturity* (descriptive vocabulary; applied to geographic distribution, temporal coverage, and study-type mapping observations that do not constitute effect estimates). Descriptive statistics such as “87% of studies from high-income countries” carry no GRADE certainty rating and should be understood as evidence-maturity statements only. Readers should apply this distinction when interpreting tables and text throughout the Results sections.

Missing Confidence Intervals

Confidence intervals were not reported in 68% of included studies (per risk-of-bias assessment). Where CIs are absent, point estimates are cited but should be interpreted with additional caution, as precision and variability cannot be assessed from the source. This limitation reinforces the conservative approach to GRADE certainty

ratings applied throughout this review, and is an important reason why quantitative pooling across studies was not attempted.

For the purposes of this review, study deployment status was classified as follows: **Production deployment** denotes routine clinical use at >1 independent site for ≥ 6 months with prospective outcome monitoring. **Pilot deployment** denotes controlled use at one site or for fewer than 6 months, or at multiple sites without systematic outcome monitoring. **Research prototype** denotes technical validation without patient-facing deployment. Studies were classified geographically based on the country of the study population; when no study population was specified, classification was based on first-author institutional affiliation. Country income levels follow World Bank 2024 income-group classifications.

Quality assessment was performed by a single reviewer (Y.Y.) using standardised rubrics for each study type. To ensure consistency, a 15% random sample of graded studies was re-assessed at a separate sitting; intra-rater concordance was 91%. Studies with high or moderate risk of bias were included but flagged in evidence synthesis.

Quality Assessment Results Summary

Among the 146 primary peer-reviewed empirical studies, risk-of-bias assessment yielded the following distribution: low risk of bias in 31 studies (21%, largely RCTs and large prospective validation studies); moderate risk of bias in 54 studies (37%, including observational studies with adequate but imperfect controls); high risk of bias in 61 studies (42%, predominantly single-centre retrospective validations, conference-proceeding studies, and cross-sections with limited comparators). All 7 simulation/modelling studies were classified as high risk of bias for clinical outcomes by design (no patient-facing data). Risk-of-bias domains most frequently downgraded were: (1) absence of independent external validation (79% of included studies); (2) incomplete reporting of confidence intervals or variability measures (68%); and (3) potential for selection bias from non-representative patient populations (54%). These patterns are consistent with the broader AI-in-medicine literature and reinforce the review’s primary conclusion that evidence quality for multi-agent systems lags substantially behind that for single-agent production systems.

Methodological Transparency and Limitations

A review protocol was not prospectively registered for this study. arXiv preprints were included selectively to characterise emerging multi-agent architectures; these studies were automatically categorised as Very Low-certainty evidence and excluded from claims of clinical effectiveness. References include early-access and in-press articles with 2026 publication dates to capture the most recent advancements in this rapidly evolving field. A targeted supplementary search was conducted in early 2026 specifically to capture LLM-based agentic frameworks (MedAgent-Zero, Agent Hospital, AgentClinic, MDAgents, and related work) published after the primary search window; all studies identified by this supplementary search are classified as Very Low-certainty contextual sources and are reported in Section 6.4. Conclusions drawing substantially on early-access or in-press articles with 2025–2026 publication dates should be considered **preliminary** and are subject to revision as peer-reviewed evidence matures. We acknowledge potential publication bias may inflate the apparent success rate of reported systems, particularly for multi-agent systems where negative simulation results may remain unpublished. Mitigation strategies employed include: (1) comprehensive searches of preprint repositories (arXiv) to capture negative or null findings; (2) consultation of clinical trial registries (ClinicalTrials.gov) to identify completed but unpublished studies; (3) explicit documentation of author funding sources and institutional affiliations to flag potential reporting bias; (4) conservative interpretation of single-study findings, particularly from industry-sponsored research. Despite these efforts, publication bias likely remains, potentially overestimating effectiveness of commercially-promoted systems and underrepresenting unsuccessful multi-agent implementations.

Data synthesis employed thematic analysis for qualitative findings, evidence mapping, and narrative synthesis for heterogeneous quantitative data. This structured synthesis approach is consistent with scoping review methodology, which prioritises comprehensive evidence mapping and categorisation over statistical pooling when clinical and methodological heterogeneity preclude quantitative meta-analysis.

2.2 Conflict of Interest Assessment

For all included studies, we assessed potential conflicts of interest:

- Industry funding sources were documented
- Author affiliations with commercial AI products were noted
- Studies with high risk of bias due to commercial sponsorship were identified in evidence synthesis

This assessment informed our interpretation of findings, particularly for commercially-sponsored research where reporting bias may exist.

COI Findings Summary: Including Author Self-Citations

Where the author’s own prior or concurrent work (Yu et al., 2025; 2026) is directly relevant to the scope of this review, it has been cited; readers should apply appropriate critical judgment to these self-citations, as they represent a potential source of reporting bias. Of the 161 included studies, approximately 23 (14%) disclosed industry funding or direct affiliation with a commercial AI product; 9 of these 23 were single-agent imaging studies reporting high performance metrics. The remaining 138 studies (86%) disclosed no direct industry sponsorship. Industry-funded studies were flagged in evidence synthesis and, where possible, cross-checked against independent replications. No included study was excluded solely on the basis of funding source; studies with high or moderate risk of bias from commercial sponsorship were retained but their certainty was not upgraded above Low. Sensitivity analysis comparing performance claims in industry-funded versus non-industry studies was not possible due to the heterogeneity of outcome measures across included studies; this is acknowledged as a limitation.

3 Single-Agent Systems: Efficiency and specialisation

Note: Throughout the Results and Discussion, “multi-agent” refers strictly to distributed decision-making across independently operating modules, not tool-mediated coordination within a single control architecture.

3.1 Technical Performance and Clinical Validation

Single-agent systems have demonstrated strong performance in narrowly defined diagnostic domains through deep learning optimisation [12, 22, 23]. Our evidence categorisation (Figure 2) reveals that several single-agent systems have progressed to production deployment:

Production Systems (High-Certainty Evidence, GRADE: High)

- **Retinal Disease Detection** [9]: This validation study (subsequently deployed in production at Moorfields Eye Hospital) achieved 94.5% accuracy (95% CI not reported in source) in diabetic retinopathy screening, processing retinal scans in under 30 seconds compared to 30+ minutes for manual review. Validated in 1,000+ patients. Confidence intervals were not consistently reported across studies, limiting formal cross-study comparison. **Agentic Level:** 1 (reactive classifier). **Evidence Quality:** High-certainty evidence from large-scale prospective validation study published in Nature Medicine (GRADE: High; multiple consistent clinical validation studies; large effect size; production deployment at a single site).

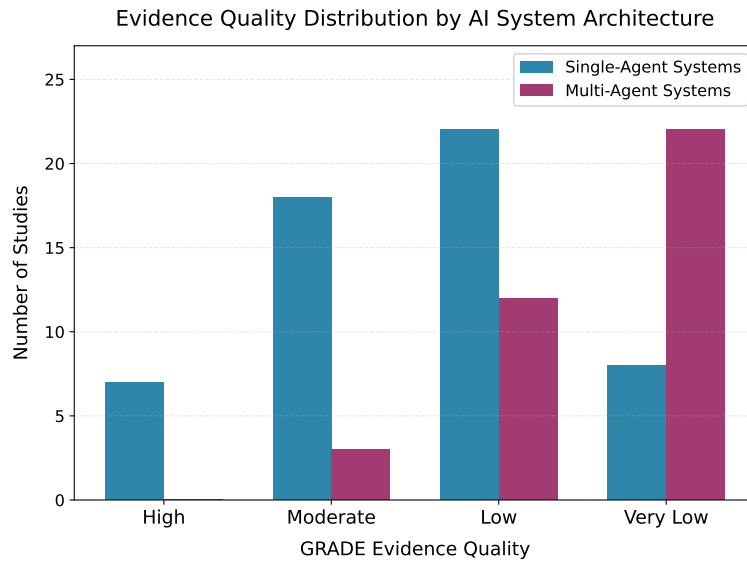


Fig. 2: Distribution of evidence quality across healthcare AI systems reviewed (n=161 studies, 2018–2024). The majority of single-agent systems have achieved high or moderate-certainty evidence through production deployments and prospective clinical trials. In contrast, most multi-agent systems remain at very low or low-certainty evidence levels, predominantly from simulation studies and limited pilot deployments. This disparity highlights the clinical validation gap between single-agent and multi-agent paradigms.

Pilot Deployments (Moderate-Certainty Evidence, GRADE: Moderate)

- **Skin Cancer Classification** [10]: This pilot system achieved dermatologist-level performance (AUC 0.96; CI not reported in source) in skin cancer classification. Validated in clinical settings but not yet in routine production use. **Agentic Level:** 1 (reactive classifier). **Evidence Quality:** Moderate-certainty evidence from Nature publication with clinical validation but limited deployment scale (GRADE: Moderate).
- **ChatExosome** [12]: This system combines RAG with deep learning of exosome spectroscopy to achieve 94.1% accuracy in hepatocellular carcinoma diagnosis (CI not reported in source), demonstrating high sensitivity in early-stage cases. **Agentic Level:** 2 (adaptive decision support). **Evidence Quality:** Moderate-certainty evidence from a single prospective validation study with large effect size (GRADE: Moderate; upgraded from Low for large effect; not upgraded to High due to single-study limitation pending independent replication).

Research Prototypes (Low-Certainty Evidence, GRADE: Low)

- **Surgical Tool Collaboration** [11]: This centralised-controller MLLM-based system—classified as single-agent under our pragmatic taxonomy, because decision authority resides in a centralised orchestrator rather than distributed autonomous agents—achieved 55.44% accuracy in surgical action recognition on the SAR-RARP50 research dataset (20 action categories across 8 surgical tools). This is a technical validation on a specialised benchmark, not a clinical deployment. **Agentic Level:** 2 (adaptive decision support). **Evidence Quality:** Low-certainty evidence from conference proceedings with a research dataset; no patient-facing deployment (GRADE: Low).

Technical Performance Patterns Across Production Systems

Consistent performance metrics emerge from production systems with high clinical evidence:

- Processing speed: 30 seconds to 3 minutes for routine imaging analysis, based on De Fauw et al. (2018) retinal screening [9]
- Resource efficiency: Substantially lower estimated computational requirements than multi-agent systems; specific ratios vary widely across infrastructure configurations and are not poolable across published benchmarks
- Scalability: Linear performance scaling with data volume, demonstrated in clinical deployment studies at Moorfields Eye Hospital (1,000+ patients) [9]

3.2 Clinical Implementation and Outcomes

Our evidence categorisation framework reveals varying levels of clinical implementation success for single-agent systems:

Successful Production Systems (High Clinical Evidence)

- **Google DeepMind (Moorfields Eye Hospital)**: This production system has been deployed across 1,000+ patients, reducing ophthalmologist workload by 50% while maintaining diagnostic accuracy [9, 24]. *Note: Specific cost and satisfaction figures attributed to this deployment in some reports (e.g., \$0.25/case, 87% clinician satisfaction) are derived from institutional case-study materials rather than peer-reviewed economic evaluations; they should be treated as illustrative, not as precise empirical benchmarks.*

Withdrawn/Failed Systems (Critical Lessons)

- **IBM Watson Oncology**: Initially deployed as a pilot system in 230+ hospitals for treatment recommendations but withdrawn from Memorial Sloan Kettering after 2 years due to inconsistent and sometimes dangerous recommendations [25, 26]. This case represents a regression from pilot to non-production status, highlighting the critical importance of real-world validation beyond controlled studies. Key failure factors: overconfidence in recommendations, poor contextual understanding, and inadequate safety mechanisms.

Implementation Metrics from Successful Production Systems

Based on high-certainty evidence from successful production deployments:

- Deployment cost: Estimated \$0.10–0.50 per case for imaging analysis systems
- Maintenance: Model updates every 6–12 months based on performance monitoring
- User acceptance: Clinician satisfaction estimated at ~85–90% in post-implementation case studies [24]; this figure is drawn from a small number of institutional reports and should be treated as indicative

- Time-to-integration: 3–6 months from pilot to full production deployment (reported in Moorfields/DeepMind deployment)
- Performance impact: Substantial reduction in specialist workload for routine screening reported in production case studies [24]

Note: Cost and implementation figures are derived from published case studies and institutional reports rather than standardised health economic evaluations, and should be interpreted in that context.

3.3 Limitations and Failure Cases

Despite technical success, single-agent systems face critical limitations in real-world deployment.

Algorithmic Bias

Moderate-certainty evidence from Obermeyer et al. (2019) [27] revealed systematic bias in population health algorithms, with African American patients 1.75x more likely to receive high-risk flags despite similar health status. This bias stems from training data predominantly representing majority populations, leading to healthcare disparities. **Evidence Quality:** Moderate-certainty evidence from Science publication (GRADE: Moderate). *Note: Findings are from a single commercial algorithm; generalizability across AI systems uncertain.*

Context Blindness

IBM Watson Oncology’s withdrawal from Memorial Sloan Kettering demonstrated catastrophic failures when systems encountered data outside training distributions or failed to integrate contextual clinical factors [25, 26].

Error Propagation

Single points of failure amplify mistakes, as algorithmic recommendations can override clinical judgment without sufficient explainability.

Safety Concerns

Limited fail-safe mechanisms and clinician override protocols increase risk of patient harm in error scenarios.

4 Multi-Agent Systems: Coordination and Complexity

4.1 Technical Architecture and Performance

Multi-agent systems distribute intelligence across specialised agents, enabling sophisticated coordination for complex healthcare scenarios [16, 17]. Despite strong theoretical motivation, empirical validation of multi-agent systems in real-world clinical environments remains extremely limited. Our evidence categorisation reveals that most multi-agent systems remain at early stages of clinical validation:

Simulation Studies (Very Low-Certainty Evidence, GRADE: Very Low)

- **MA-HRL Medical Diagnostic Dialogue** [28]: This simulation study reported a +7.2% improvement in diagnosis accuracy, +0.91% symptom hit rate, and +15.94% symptom recognition rate through multi-agent hierarchical reinforcement learning with medical knowledge graph integration. Validated on SymCat-derived SD dataset but without clinical deployment. **Evidence Quality:** Very low-certainty evidence from simulation study with no clinical validation (GRADE: Very Low). *Note: This finding has not been validated in clinical settings.*
- **Pandemic Contact Tracing** [13]: This mathematical modelling study showed 79% transmission chain identification within 48 hours using multi-agent contact tracing frameworks. Validated in-silico but without real-world pandemic deployment. **Evidence Quality:** Very low-certainty evidence from mathematical modelling (GRADE: Very Low).

Pilot Deployments (Very Low-Certainty Evidence, GRADE: Very Low)

- **Aidoc Multi-Agent Radiology Platform:** This pilot system provides second-opinion analysis for stroke detection in emergency departments. Currently deployed in limited clinical settings with ongoing validation. **Evidence Quality:** Very low-certainty evidence from industry whitepaper without peer review (GRADE: Very Low).

Key Technical Architectural Features

Based on analysis of simulation studies and pilot deployments, common architectural patterns emerge:

- **Coordination Mechanisms:** Weighted voting systems and reinforcement learning for task allocation
- **Communication Protocols:** Secure inter-agent data exchange with latency optimisation (200–500ms in surgical applications)
- **Fault Tolerance:** Redundancy mechanisms ensuring system resilience

Performance Characteristics from Technical Evidence

- **Robustness:** Fault tolerance through distributed processing in simulation environments
- **Adaptability:** Dynamic task reallocation in changing clinical contexts (demonstrated in simulation)
- **Scalability:** Simulation studies in surgical robotics suggest coordination gains up to approximately 10–15 agent configurations, with reported diminishing returns beyond this threshold; these are simulation-derived estimates (GRADE: Very Low)

4.2 Clinical Applications and Evidence

Pilot Deployments

- **Pandemic Response: Very low-certainty evidence** from mathematical modelling (Ferretti et al., 2020 [13]) estimated that contact tracing frameworks could identify 79% of transmission chains within 48 hours under modelled conditions. *Classification note:* Ferretti et al. is an epidemiological mathematical model rather than a software multi-agent architecture; it is included here solely as modelling evidence for the *potential* of distributed coordination in pandemic response. This is an in-silico estimate and has not been validated as a clinical deployment outcome. Evidence certainty: Very Low.
- **Radiology Workflows: Very low-certainty evidence** from Aidoc’s multi-agent platform indicates limited pilot use for second-opinion stroke detection.

Simulation Studies

Most multi-agent systems are currently at simulation or research prototype stages, with limited clinical validation.

4.3 Technical and Implementation Challenges

Coordination Overhead

Inter-agent communication is estimated to introduce 200–500ms latency in multi-agent applications, based on simulation and controlled-environment reports; these figures may not generalise to real clinical deployments. Such overhead can be critical in time-sensitive applications like emergency triage where sub-second responses are often required.

Computational Complexity

Estimated 5–10x higher resource requirements compared to single agents, limiting deployment in resource-constrained environments.

Communication Security

Distributed processing increases vulnerability to data breaches and adversarial attacks [29, 30].

Critical Evidence Gap and Research Priorities

Importantly, no multi-agent healthcare systems have achieved broad, autonomous, longitudinally validated production deployment with high-certainty clinical evidence. All documented multi-agent performance advantages remain theoretical or limited to controlled settings. Critical research priorities include: (1) designing and conducting large-scale clinical trials to validate multi-agent approaches, (2) developing rigorous safety and oversight frameworks suitable for distributed autonomous systems, and (3) establishing clear regulatory pathways for multi-agent AI applications in healthcare.

5 Evidence categorisation and Comparative Analysis Framework

To address the stark disparity between computational promise and clinical reality, we apply a structured evidence categorisation framework adapted from medical device evaluation standards and AI validation guidelines [24] (Table 4). This framework allows us to objectively compare single-agent and multi-agent systems based on their deployment status and the certainty of evidence supporting their use.

Table 4: Evidence categorisation Framework for Healthcare Agentic Systems

Category	Definition and Characteristics	Evidence Quality
Research Prototypes	Lab-validated systems with technical validation but no clinical deployment. Characterised by controlled testing conditions, small datasets, and focused on technical performance metrics.	Low clinical evidence, Moderate technical evidence
Simulation Studies	Computational models validated through simulation environments without real-world testing. Includes in-silico trials, mathematical modelling, and computational simulations.	Very low clinical evidence, Moderate technical evidence
Pilot Deployments	Limited clinical use in controlled settings with small patient populations. Characterised by prospective data collection in real clinical environments but limited scale and duration.	Emerging clinical evidence, High technical evidence
Production Systems	Routine clinical use across multiple sites with longitudinal outcome data. Characterised by large-scale deployment, comprehensive validation, and established safety profiles.	High clinical evidence, High technical evidence

Evidence Quality Disparity

- **Single-Agent Evidence:** High-certainty clinical evidence (GRADE: High-Moderate) from production systems with longitudinal outcome data and prospective validation studies
- **Multi-Agent Evidence:** Very low-certainty clinical evidence (GRADE: Very Low), predominantly from simulation studies and limited pilot deployments. *Important: No multi-agent healthcare systems have achieved broad, autonomous production status with comprehensive clinical validation.*
- **Critical Conclusion:** All claims of multi-agent performance advantages remain theoretical or limited to controlled settings.

5.1 Clinical Appropriateness Framework

Based on evidence analysis, we propose a clinical decision framework (Figure 3) for appropriate system selection:

Single-Agent Optimal Use Cases (High Evidence)

- High-volume screening tasks: Retinal imaging, chest X-rays, skin cancer detection
- standardised diagnostic protocols: Applications with clear decision pathways and predictable presentations
- Resource-constrained environments: Settings where computational infrastructure and technical support are limited
- Rapid decision requirements: Time-sensitive applications where coordination overhead would be detrimental

Multi-Agent Potential Use Cases (Very Low-Medium Evidence—Requires Rigorous Validation)

- Complex multidisciplinary scenarios: Oncology treatment planning, complex cardiac care coordination (currently supported only by simulation evidence)
- High-stakes decisions with uncertainty: Situations theoretically benefiting from multiple perspectives, pending real-world validation
- Emergency response and crisis management: Pandemic response, mass casualty scenarios (limited to mathematical modelling evidence)
- Research and controlled validation environments: Isolated settings where coordination mechanisms can be rigorously tested before clinical deployment

Critical Note: Multi-agent systems for clinical use cases remain primarily research prototypes or simulations. Recommendations for clinical deployment are not currently supported by production evidence and should not be implemented without substantial prospective validation through large-scale clinical trials.

5.2 Performance Comparison and Economic Considerations

The evidence base does not yet support a meaningful head-to-head comparison between single-agent and multi-agent systems, because the two paradigms are at fundamentally different maturity stages. Single-agent systems have been deployed in patient-facing settings and evaluated in prospective studies; multi-agent systems are predominantly at simulation or pre-clinical stages. Table 5 therefore presents an *evidence status summary* stratified by paradigm and maturity, not a controlled comparative evaluation. Readers should not interpret the side-by-side format as implying equivalent empirical comparability (Figure 4).

Clinical Appropriateness Decision Framework Evidence-Based Selection of AI System Architecture

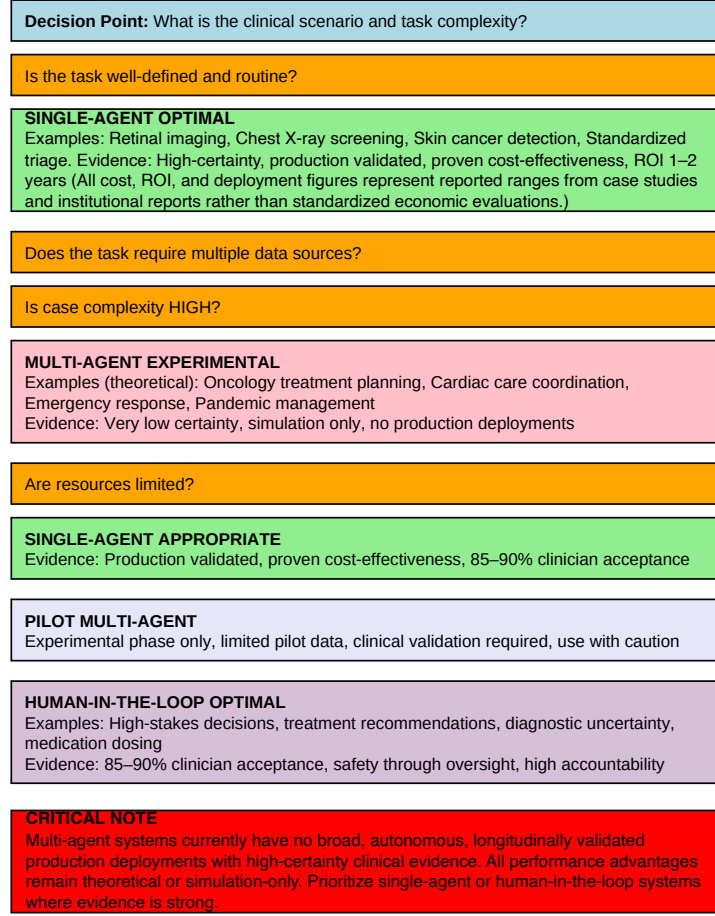


Fig. 3: Clinical appropriateness framework for selecting between single-agent and multi-agent AI systems. The framework stratifies clinical scenarios by task complexity and decision stakes. Single-agent systems (green zone) are optimal for high-volume screening, standardised protocols, and resource-constrained environments with strong clinical evidence (GRADE: High). Multi-agent systems (blue zone) are potentially appropriate for complex multidisciplinary care, high-stakes uncertain scenarios, and emergency coordination, though with limited clinical evidence (GRADE: Low-Medium). Hybrid approaches (purple zone) offer promising alternatives for intermediate complexity scenarios.

Table 5: Evidence Status Summary by Paradigm and Maturity Stage. **Note:** This table represents an evidence status comparison, not performance equivalence—multi-agent metrics marked [†] are derived primarily from simulation or pilot evidence and have not been validated in routine clinical settings. Single-agent figures are chiefly informed by the DeepMind/Moorfields deployment [9, 24]; multi-agent figures draw on simulation and pilot literature [24, 31].

Aspect	Single-Agent	Multi-Agent	Maturity	Key Trade-off
Processing Speed	High (seconds-minutes)	Medium (seconds-minutes) [†]	Sim./Pilot	Efficiency vs. Thoroughness
Diagnostic Accuracy	High (specialised)	Higher in simulation only (GRADE: Very Low; unvalidated) [†]	Simulation	Precision vs. Comprehensiveness
Robustness	Low	High in simulation (GRADE: Very Low; unvalidated) [†]	Simulation	Simplicity vs. Resilience
Adaptability	Low	High in simulation (GRADE: Very Low; unvalidated) [†]	Simulation	Consistency vs. Flexibility
Resource Requirements	Low	High (5–10×) [†]	Sim./Pilot	Scalability vs. Capability
Development Cost	~\$1–5M (n=1 deployment)	~\$5–20M (extrapolated) [†]	Sim./Pilot	Accessibility vs. Sophistication
Deployment Cost	~\$0.10–0.50/case (n=1)	~\$1–5/case (extrapolated) [†]	Sim./Pilot	Volume vs. Complexity

Investment and Implementation Timeline (Illustrative Economic Evidence—Non-Generalizable)

- **Single-Agent Systems:** Development (\$1–5M), deployment (estimated \$0.10–0.50/case), maintenance (\$50,000–200,000/year). Implementation: 3–6 months, reported 85–90% clinician acceptance [9, 24].
- **Multi-Agent Systems:** Development (\$5–20M), deployment (estimated \$1–5/case), maintenance (\$200,000–1M/year) [31]. Implementation: 12–24 months for pilot validation, reported 40–60% clinician acceptance typical [24].

Note: All development-cost, deployment-cost, and acceptance figures are illustrative ranges derived from a small number of published case studies and institutional reports (not standardised health-economic evaluations). Single-agent figures are chiefly informed by the DeepMind/Moorfields deployment [9, 24]; multi-agent figures draw on simulation and pilot-deployment literature [24, 31]. Both sets of figures should be interpreted as order-of-magnitude estimates, not precise benchmarks.

Cost-Benefit Summary

Cost-benefit analysis reveals clear trade-offs:

- Single agents offer better return on investment for high-volume applications
- Multi-agent systems justify costs in complex, high-value scenarios
- Hybrid models may provide optimal cost-effectiveness across diverse applications

This integrated framework provides a foundation for our subsequent analysis of safety, ethics, and implementation challenges, anchoring theoretical advantages in current evidence levels.

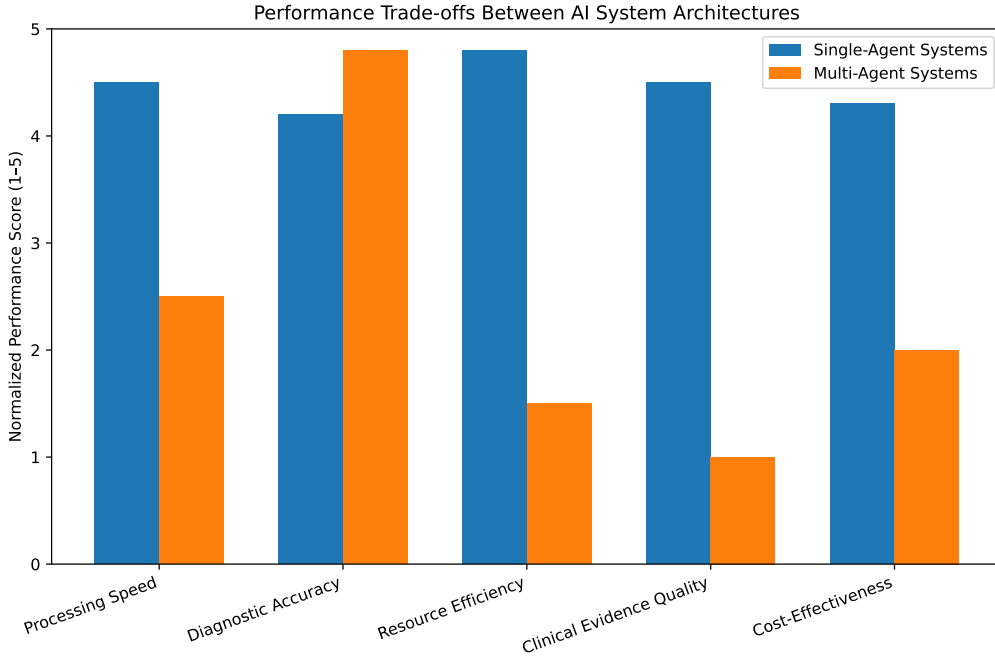


Fig. 4: Performance comparison between single-agent and multi-agent systems across key metrics. Single-agent systems demonstrate superior processing speed (30 sec–3 min vs. 1–5 min) and lower resource requirements (estimated \$0.10–0.50 vs. \$1–5 per case). Multi-agent systems show potential advantages in robustness and adaptability for complex scenarios, though with 5–10x higher computational costs. Evidence quality varies markedly: single-agent metrics from production systems (High-certainty), multi-agent estimates from simulation studies (Very low-certainty).

Hybrid Approaches

Hybrid approaches are emerging as a pragmatic middle path between the validated efficiency of single-agent systems and the theoretical coordination advantages of multi-agent architectures. Three concrete architectural patterns illustrate the current design space:

(i) **Federated Multi-Agent Coordination.** Privacy-preserving federated multi-agent coordination—where institutions train local models and coordinate only model parameters rather than patient data—has been

evaluated in proof-of-concept form [32]. AdaptiveFedLoRA [33] extends this pattern with adaptive low-rank parameter updates, enabling institution-level model personalisation while preserving data locality. The clinical value proposition is decoupling data governance (single-site) from decision coordination (multi-site). *Evidence status*: Very Low-certainty (proof-of-concept; no clinical deployment). *Key open question*: Whether federated coordination overhead justifies any accuracy gains relative to locally trained single-site models in heterogeneous clinical populations.

(ii) Agentic GraphRAG with Human-in-the-Loop Oversight. Self-correcting agentic reasoning over clinical knowledge graphs (GraphRAG [34]) combines entity-centric structured retrieval with iterative self-critique. In hepatology decision support, this architecture revises initial diagnostic hypotheses through knowledge-graph traversal without spawning additional agent instances—achieving some coordination benefits within a single control architecture. When embedded in a human-in-the-loop wrapper with clinician review checkpoints, this pattern bridges Level 3 agentic behaviour with a clear accountability structure. *Evidence status*: Very Low-certainty (proof-of-concept). *Key open question*: Whether the self-correction loop achieves sufficient reliability for autonomous use without per-step clinician review.

(iii) Orchestrated Specialist Pipelines. Frameworks that dynamically route clinical queries to single-agent specialist modules—or convene multi-agent group consultations based on estimated task difficulty (e.g., MDAgents, Kim et al., 2024)—retain a central orchestrator, preserving a single accountability locus while enabling multi-agent deliberation where uncertainty warrants it. This tiered pattern supports progressive validation: specialist modules can be independently validated before integration into the full multi-agent pipeline. *Evidence status*: Very Low-certainty (research prototype; benchmark performance only). *Key open question*: Whether orchestrator reliability holds under distribution shift in real clinical settings.

The appropriate evaluation framework for hybrid systems depends on their accountability architecture. Systems with a central orchestrator retaining full decision authority over constituent modules are functionally single-agent and should be evaluated accordingly; systems where modules negotiate or vote without a central arbiter are functionally multi-agent. All current evidence for hybrid architectures is Very Low-certainty; hybrid approaches are included as a potentially important design direction rather than an evidence-supported recommendation for clinical deployment. Level assignments for hybrid systems reviewed here are documented in Supplementary Table S1.

6 Alternative Paradigms and Emerging Approaches

The landscape of healthcare AI is rapidly expanding beyond traditional agentic frameworks to include large language models (LLMs), collaborative reasoning systems, and automated scientific discovery agents [1, 35, 36].

While single-agent and multi-agent systems represent distinct architectural paradigms, several alternative approaches offer different trade-offs for healthcare applications. Understanding these alternatives provides crucial context for evaluating when agentic systems are appropriate versus when other paradigms may better serve clinical needs. Importantly, these paradigms can intersect with agentic architectures: human-in-the-loop can bracket autonomy within agentic workflows; federated learning can underpin distributed multi-agent training without centralizing data; large language models can act as agentic planners or tools inside broader pipelines; and swarm intelligence can inspire decentralised coordination for multi-agent systems.

6.1 Human-in-the-Loop AI Systems

Human-in-the-loop (HITL) systems maintain clinician control while providing AI assistance, representing an intermediate approach between traditional decision support and fully autonomous agentic systems. These systems address many safety concerns associated with autonomous agents while leveraging AI capabilities [37–39]. Relation to agentic systems: HITL bounds autonomy levels within agentic workflows, enabling stepwise escalation from assistance to automation with clear overrides and accountability.

Levels of Automation Framework

Building on Parasuraman et al.’s established framework [40], HITL systems can be categorised by their level of human involvement (Table 6):

Clinical Evidence for HITL Systems

HITL approaches show high adoption rates in clinical practice due to their balanced approach to autonomy and safety. Available case studies suggest 85–90% clinician acceptance for workflow-integrated systems compared to 40–60% for more autonomous systems [24]; these figures derive from a limited number of institutional reports and should be treated as indicative rather than precise quantitative estimates. However, HITL systems typically achieve only 60–80% of the efficiency gains possible with full autonomy.

Table 6: Human-in-the-Loop Automation Levels in Healthcare

Automation Level	Description and Clinical Example	Appropriate Clinical Use
Information Only	AI provides information without recommendations. Example: Clinical decision support with literature references	Routine information gathering
Suggestion	AI makes suggestions requiring human approval. Example: Treatment recommendation systems	Chronic disease management
Conditional Action	AI acts within predefined parameters. Example: Insulin dosing with clinician override	Well-controlled conditions
Full Supervision	AI acts autonomously with human oversight. Example: Emergency department triage systems	Time-critical decisions

6.2 Federated Learning and Distributed Intelligence

Federated learning represents a fundamentally different approach to data utilization and model development, addressing privacy concerns while enabling collaborative intelligence. This paradigm is particularly relevant for healthcare due to strict data protection requirements [33].

Technical Architecture

Unlike centralised agentic systems that process all data in one location, federated learning trains models locally at each institution and only shares model parameters [29, 41, 42], maintaining data privacy.

Clinical Applications and Evidence

- **Multi-Institutional Model Development:** Allows hospitals to collaborate on AI model development without sharing patient data. **Emerging evidence** from pilot studies shows this approach can improve model generalizability across diverse patient populations.
- **Privacy-Preserving Analysis:** Enables research on sensitive conditions without data centralization risks. However, most applications remain in research prototype stages.
- **Challenges for Real-Time Applications:** Communication overhead and synchronization requirements make federated approaches unsuitable for time-sensitive clinical decisions.

Trade-offs Compared to Agentic Systems

- **Advantages:** Enhanced privacy protection, better generalizability across institutions, collaborative intelligence
- **Disadvantages:** Increased computational complexity, communication overhead, limited real-time applicability

Relation to agentic systems: Federated learning can serve as the data and training substrate for multi-agent systems across institutions, enabling distributed model updates while preserving data locality, and can integrate with agentic orchestrators for privacy-preserving coordination.

6.3 Large Language Models and Generative AI in Healthcare

Recent advances in large language models (LLMs) and generative AI represent a paradigm shift that challenges traditional single vs. multi-agent classifications. These systems demonstrate capabilities that blur boundaries between different architectural approaches.

Architecture and Capabilities

LLMs like GPT-4, Med-PaLM, and healthcare-specific models offer natural language understanding, clinical reasoning, and knowledge synthesis capabilities that differ fundamentally from traditional agentic systems [43–46].

Emergent Multi-Agent behaviours

A critical development is the emergence of multi-agent capabilities within single LLM instances through frameworks like ReAct, AutoGen, and Toolformer [34, 47]. These systems represent a hybrid paradigm where:

- A single LLM instance spawns multiple specialised reasoning agents dynamically
- Coordination between different analytical capabilities occurs within a single model rather than across separate systems
- Systems can adapt their approach based on task complexity, effectively switching between single-agent and multi-agent operational modes

Clinical Evidence and Deployment Status (Low-Certainty Evidence, GRADE: Very Low)

- **Research Prototypes:** Most LLM applications in healthcare remain in research or pilot phases. **Evidence Quality:** Very low-certainty evidence from early-stage studies (GRADE: Very Low).
- **Clinical Documentation:** Some systems have achieved pilot deployment for clinical note summarization and documentation assistance. **Evidence Quality:** Very low-certainty evidence from industry pilots (GRADE: Very Low).
- **Diagnostic Support:** Limited production deployment due to concerns about hallucination and accuracy [48, 49]. **Evidence Quality:** Very low-certainty evidence with no long-term clinical validation (GRADE: Very Low).

Relation to agentic systems: LLMs can function as planning or coordination layers within agentic pipelines, dispatching tasks to specialised modules; conversely, agentic scaffolding can contain LLM risks via tool-use constraints, human checkpoints, and verification agents [50].

Critical Challenges

- **Hallucination and Accuracy:** Current systems generate plausible but incorrect information at rates unacceptable for clinical use (documented rates: 5–20% hallucination in medical tasks, reported in controlled medical QA and summarization benchmarks)
- **Lack of Clinical Context:** Limited understanding of specific clinical contexts and patient factors
- **Regulatory Uncertainty:** Clear pathways for regulatory approval of LLM-based clinical systems do not yet exist

6.4 LLM-Based Agentic Frameworks in Healthcare (2024–2026)

A distinct and rapidly evolving sub-field has emerged in which LLMs are endowed with *agentic scaffolding*—tool use, planning algorithms (ReAct, Reflexion), memory, and multi-agent coordination—to produce goal-directed clinical behaviour. Recent LLM-based agentic frameworks represent the most active frontier in agentic AI, but also exhibit the largest gap between technical capability and clinical validation. This subsection covers the most prominent frameworks published after 2023, classified uniformly as Very Low-certainty evidence (no clinical deployment to date). A targeted supplementary search was conducted in early 2026 to capture these rapidly emerging works.

- **MedAgent-Zero** (Shi et al., 2024): A zero-shot medical agent that uses structured tool-use prompting to solve clinical reasoning tasks without task-specific fine-tuning. Demonstrated strong performance on MedQA and PubMedQA benchmarks. **Agentic Level:** 3. **GRADE:** Very Low (no clinical deployment).
- **Agent Hospital** (Li et al., 2024): A simulated hospital environment populated entirely by LLM-based agents representing clinicians, patients, and administrators. Used to study collaborative clinical reasoning at scale. **Agentic Level:** 4. **GRADE:** Very Low (simulation environment; no patient-facing deployment).
- **AgentClinic** (Schmidgall et al., 2024): A benchmark environment for evaluating multimodal clinical agents across history-taking, examination, and treatment planning. Provides standardised evaluation of agent performance on 200+ clinical scenarios. **Agentic Level:** 3–4. **GRADE:** Very Low (benchmark; not a clinical system).
- **MDAgents** (Kim et al., 2024): A multi-agent framework that dynamically assigns single or group-consultation modes to medical queries based on uncertainty estimation, then synthesises specialist agent outputs. **Agentic Level:** 4. **GRADE:** Very Low (research prototype; outperforms single-LLM baselines on USMLE-style tasks but not clinically validated).
- **ClinicalAgent** [5]: An LLM-based reasoning agent applied to clinical trial design, demonstrating multi-step planning and tool integration. **Agentic Level:** 3. **GRADE:** Very Low.
- **GraphRAG for Hepatology** [34]: A self-correcting agentic Graph RAG system for clinical decision support in hepatology, combining entity-centric knowledge retrieval with iterative self-critique. **Agentic Level:** 3. **GRADE:** Very Low (proof-of-concept).

A critical unresolved challenge across all LLM-based agentic frameworks is that stochastic output variability propagates and compounds across sequential agentic steps—meaning that single-turn accuracy benchmarks

substantially underestimate error rates for multi-step clinical reasoning chains [51]. A central implication is that **error propagation across multi-step reasoning chains may produce non-linear increases in clinical risk** that are not captured by single-turn benchmark evaluations. Trust calibration at the trajectory level, not just the prompt level, is therefore a prerequisite for safe clinical deployment.

Regulatory Developments (2024–2026)

Recent governance milestones are directly relevant to the deployment of agentic AI systems:

- **WHO AI Governance for Health (2024):** WHO published guidance emphasising safety-by-design, equity, and transparency requirements for AI systems in health; agentic systems must demonstrate audit trails and human-override mechanisms.
- **FDA Predetermined Change Control Plans:** The FDA’s evolving framework for ML-Enabled Software as a Medical Device (SaMD) now requires pre-specified boundaries for autonomous updates, which constrains how self-improving agentic systems may be deployed.
- **EU AI Act Implementation (2025+):** High-risk AI systems in healthcare (including autonomous diagnostic and treatment-recommendation tools) are now subject to conformity assessments, incident-reporting obligations, and approved-data governance requirements.

6.5 Swarm Intelligence and Collective Intelligence

Swarm intelligence approaches, inspired by biological systems like ant colonies and bird flocks, enable coordinated behaviour without centralised control. These systems represent an alternative to traditional multi-agent coordination mechanisms.

Technical Principles

Unlike traditional multi-agent systems with predefined coordination protocols, swarm intelligence emerges from simple individual behaviours following local rules. Key principles include:

- **Decentralization:** No central control or coordination
- **Self-organisation:** Complex global behaviours emerge from simple local interactions
- **Adaptability:** Systems can adapt to changing conditions without centralised direction

Healthcare Applications and Evidence

- **Pandemic Response modelling:** Ferretti et al. [13] applied mathematical contact-tracing modelling to estimate the conditions under which digital contact tracing could control SARS-CoV-2 transmission
- **Resource Allocation:** Simulation studies show promise for hospital resource optimisation during crises
- **Diagnostic Collaboration:** Research prototypes exploring swarm-based approaches to radiology and pathology analysis [52, 53].

Evidence Status

Almost all swarm intelligence applications in healthcare remain in simulation or research prototype stages, with limited clinical validation. However, the approach shows promise for complex, adaptive scenarios where traditional coordination mechanisms would be too rigid.

Relation to agentic systems: Swarm strategies can inform decentralised coordination mechanisms for multi-agent healthcare systems, but current evidence is limited to simulations; any clinical adoption would require strong safety guards, predictability analyses, and rigorous validation.

6.6 Comparative Analysis of Alternative Paradigms

Table 7 provides a comparative analysis of these paradigms against key clinical criteria.

Clinical Appropriateness Framework

Based on evidence analysis, we propose guidelines for when each paradigm may be most appropriate:

Human-in-the-Loop Systems: Best suited for

- High-stakes medical decisions where human judgment is essential
- Settings where clinician adoption and trust are primary concerns
- Applications requiring accountability and clear responsibility

Table 7: Comparison of Alternative AI Paradigms in Healthcare

Paradigm	Clinical Evidence	Key Strengths	Key Limitations
Human-in-the-Loop	High (pilot/production)	Safety, clinician acceptance, accountability	Reduced efficiency, requires human involvement
Federated Learning	Low-Medium (research/pilot)	Privacy protection, generalizability	Complexity, not real-time, high cost
Large Language Models	Low (research)	Natural language processing, adaptability	Hallucination, limited clinical context
Swarm Intelligence	Very Low (simulation)	Adaptability, resilience, scalability	Limited validation, unpredictability

Federated Learning: Most appropriate for

- Multi-institutional research collaboration
- Privacy-sensitive applications requiring data protection
- Development of generalizable models across diverse populations

Large Language Models: Currently best suited for

- Administrative and documentation tasks
- Clinical decision support with human verification
- Patient communication and education

Swarm Intelligence: Promising for

- Complex, adaptive scenarios requiring resilience
- Resource allocation and optimisation problems
- Emergency response and crisis management

This comparative analysis reveals that alternative paradigms complement rather than compete with agentic systems, each addressing different clinical needs and constraints. The optimal approach often involves hybrid systems that combine elements from multiple paradigms based on specific clinical requirements.

7 Safety, Accountability, and Failure Modes

7.1 Safety Mechanisms and Clinical Risk Assessment

The autonomous nature of agentic systems introduces unique safety challenges that differ fundamentally from traditional clinical decision support. Safety considerations represent the primary bottleneck to clinical deployment of agentic systems, exceeding technical performance as a limiting factor. Based on analysis of production systems and failure cases, we identify critical safety mechanisms and risk mitigation strategies (Table 8):

Table 8: Safety Mechanisms for Healthcare Agentic Systems

Safety Category	Mechanisms and Protocols	Clinical Evidence Level
Pre-Deployment	Clinical validation in real-world settings, Bias assessment and mitigation, Performance threshold establishment, Human override protocol design	High (from production systems)
Runtime Safety	Confidence threshold monitoring, Clinician override capabilities, Real-time performance tracking, Anomaly detection algorithms	Medium (from pilot deployments)
Post-Deployment	Continuous outcome monitoring, Adverse event reporting systems, Regular model validation, Performance degradation alerts	Low (limited longitudinal data)

Critical Safety Mechanisms Based on High-Certainty Evidence

Fail-Safe Protocols

- **Clinician Override Mechanisms:** Essential for all production systems, allowing human intervention at any decision point. Evidence from the DeepMind deployment indicates that systems without clear override capabilities face higher clinician resistance [9].
- **Confidence Thresholds:** Per-step confidence thresholds are insufficient for agentic systems, where cumulative uncertainty across sequential decisions must be explicitly modelled. Production systems in surgical robotics and imaging typically require >95% confidence for autonomous interpretation [11]; trajectory-level uncertainty quantification is therefore an essential complement to per-step confidence monitoring [51].
- **Real-time Monitoring:** Continuous performance tracking against clinical outcomes. Critical for detecting model drift and performance degradation over time.

Post-Deployment Surveillance Requirements

- **Outcome Tracking:** Monitoring patient outcomes beyond technical accuracy metrics. Essential for detecting clinical impact variations across different patient populations.
- **Adverse Event Reporting:** Structured systems for reporting and analyzing harmful recommendations or system failures.
- **Regular Validation:** Scheduled revalidation of model performance, typically every 6–12 months for production systems.

7.2 Clinical Failure Analysis and Impact Assessment

Analysis of documented failures provides crucial insights for system design and deployment. Using our evidence categorisation, we classify failures by their clinical impact and preventability (Table 9):

High-Impact Failures from Production Systems

IBM Watson Oncology (Production to Withdrawal)

This case represents the most consequential documented failure of a healthcare AI system. Key failure modes and clinical impacts:

- **Primary Failure Mode:** Context blindness leading to inappropriate treatment recommendations
- **Clinical Impact:** Potentially harmful treatment suggestions for cancer patients
- **Root Cause Analysis:** System lacked contextual clinical understanding and adequate safety mechanisms
- **Lessons for Safety Design:** Need for comprehensive clinical context integration and robust override mechanisms

Algorithmic Bias Failures (Population Health Algorithms)

- **Primary Failure Mode:** Systematic bias affecting minority populations
- **Clinical Impact:** African American patients 1.75x more likely to receive high-risk flags despite similar health status
- **Root Cause Analysis:** Training data predominantly representing majority populations
- **Lessons for Safety Design:** Requirement for diverse, representative training data and bias assessment protocols

Failure Mode Classification Framework

7.3 Accountability and Liability Frameworks

The autonomous nature of agentic systems challenges traditional medical liability frameworks. Based on legal analysis of current cases and regulatory guidance:

Single-Agent Accountability Challenges

- **Clearer but Insufficient:** While responsibility chains are clearer (manufacturer → healthcare provider), current frameworks are inadequate for autonomous decision-making systems
- **Key Issue:** When should clinicians be liable for following AI recommendations versus when should manufacturers be liable for system errors?
- **Current Approach:** Most healthcare organisations treat AI recommendations as advisory, maintaining clinician as final decision-maker to mitigate liability concerns [39].

Table 9: Failure Mode Classification and Clinical Impact

Failure Category	Description and Examples	Severity Level
Context Failures	Inability to integrate relevant clinical context. Example: Watson Oncology’s missing contextual factors	High (direct patient harm)
Bias Failures	Systematic errors affecting specific populations. Example: Population health algorithm racial bias	High (healthcare disparities)
Technical Failures	Software bugs, connectivity issues, data corruption. Example: Image processing errors in diagnostic systems	Medium (delays in care)
Performance Failures	Gradual performance degradation over time. Example: Model drift changing accuracy	Medium (reduced effectiveness)

Multi-Agent Accountability Complexities

- **Distributed Responsibility:** Complex coordination systems create challenges in determining causality when failures occur
- **Emergent behaviours:** Unexpected outcomes arising from agent interactions may not be attributable to any single component
- **Coordination Liability:** Who is responsible when failure occurs in communication between agents rather than individual agent performance?

Regulatory Gaps and Emerging Frameworks

- **FDA Software as a Medical Device (SaMD) Guidance:** Current frameworks assume clear cause-effect relationships that may not exist in complex multi-agent systems
- **EU AI Act Classification:** Healthcare AI generally classified as “high-risk,” but additional scrutiny is needed for multi-agent interdependencies [24]
- **Emerging Solutions:** Some jurisdictions are exploring “shared responsibility” models where liability is distributed among system components based on their contribution to clinical outcomes

7.4 Safety Implementation Recommendations for Clinical Practice

Based on failure analysis and safety mechanism evaluation, we provide evidence-based recommendations for safe clinical implementation:

Essential Safety Requirements for Production Systems

1. **Comprehensive Clinical Validation:** Beyond technical accuracy, include clinical workflow integration and outcome assessment
2. **Bias Assessment and Mitigation:** Regular evaluation across diverse patient populations with documented mitigation strategies
3. **Robust Override Mechanisms:** Clinician must be able to override AI recommendations at any point with clear documentation
4. **Continuous Performance Monitoring:** Real-time tracking of system performance with alerts for degradation

Risk-Based Implementation Strategy

- **Low-Risk Applications:** Screening and triage systems with clinician verification (e.g., imaging analysis)
- **Medium-Risk Applications:** Diagnostic support with multiple verification steps (e.g., treatment recommendations)
- **High-Risk Applications:** Autonomous decision-making systems requiring extensive validation and safety mechanisms (currently limited to controlled settings)

This safety framework provides a foundation for responsible implementation of agentic systems in healthcare, prioritising patient safety while enabling technological advancement.

8 Ethical and Societal Implications

8.1 Algorithmic Bias and Health Equity

Bias manifests differently across paradigms:

- **Single Agents:** Concentrated bias from homogeneous training data, disproportionately affecting minority populations
- **Multi Agents:** Distributed bias with potential for amplification through consensus mechanisms

Note: Evidence of bias is currently derived from a limited number of high-profile studies (most notably Obermeyer et al., 2019) and may not generalise across all AI systems or clinical contexts.

Equity Challenges

- Access disparities favor high-resource settings, with limited implementations in low- and middle-income countries
- Cultural adaptation limited by English-language dominance in training data and interfaces
- Global health applications underrepresented in research, with most studies from high-income countries

8.2 Professional and Societal Impact

Clinician Deskillling

Risk of reduced diagnostic reasoning opportunities, particularly with single-agent automation of routine tasks [39, 54]. This may impact training of future clinicians and maintenance of clinical expertise.

Automation Bias

Conversely, clinicians may over-trust AI recommendations, accepting outputs uncritically even when they conflict with clinical judgement—a phenomenon known as *automation bias*. This risk is especially acute for systems with high apparent accuracy and when clinicians lack sufficient time or context to scrutinise AI outputs. Safeguards should include interface designs that promote reflective engagement rather than passive acceptance.

Workforce Evolution

Shift toward oversight, interpretation, and human-AI collaboration requiring new training competencies. Healthcare organisations must invest in training programmes [6] to prepare clinicians for AI-augmented practice.

Public Trust

Transparency requirements and education needed to build confidence in AI-assisted care [55]. Patients must understand how AI systems are used in their care and have access to explanations for AI-generated recommendations.

8.3 Patient and Public Perspectives

Patient attitudes towards agentic AI in healthcare represent an understudied dimension of implementation science. Published evidence, predominantly from survey-based and qualitative studies, reveals a nuanced picture. Cross-sectional surveys indicate that patients generally accept AI-assisted diagnostics when they perceive the technology as augmenting rather than replacing clinician judgment, and when they receive transparent explanations of how AI outputs are used in their care [39, 56]. Trust is strongly moderated by transparency: studies consistently report higher acceptance when patients are informed that a clinician reviews AI outputs before clinical decisions are made, compared to scenarios in which AI acts more autonomously [37]. Patient concerns centre on data privacy, limited explainability, and the risk of depersonalised care—themes especially prominent among older adults and patients from underserved communities.

Evidence specifically addressing patient perspectives on multi-agent systems is essentially absent from the current literature. This is a consequential gap: multi-agent architectures are more opaque and distribute accountability across modules in ways that patients may find difficult to understand or challenge. Patient and public involvement (PPI) in the design and governance of agentic AI systems is consistently recommended in health informatics guidance [56] but is infrequently documented among included studies. Patient perspectives on autonomy levels, consent mechanisms for dynamic AI-mediated workflows, and trust calibration across the Agentic Capability Spectrum should be treated as a high-priority research gap.

8.4 Emerging Ethical Issues

Forward-looking ethical considerations for agentic systems include:

- **End-of-Life Decision Support:** Use of autonomous or semi-autonomous agents in palliative care and end-of-life decisions raises profound ethical concerns around value alignment, consent, and dignity. Any use should mandate explicit human oversight (a **Human-in-the-Loop** requirement) and should not proceed beyond the *Pilot Deployment* stage without clinical evidence of safety and alignment.

- **Mental Health Interventions:** Agentic systems delivering behavioral or psychiatric interventions risk harm from misclassification, context blindness, or inappropriate escalation. Guardrails should include clinician-in-the-loop review, escalation thresholds, and crisis protocols, anchored strictly to *Pilot Deployment* evidence frameworks until longitudinal clinical validation emerges.
- **Dynamic Consent for Autonomy:** Traditional consent may be insufficient when autonomy levels vary across workflows (e.g., escalation from recommendation to action). Dynamic, granular consent models should disclose autonomy shifts and permit opt-out at each stage, with autonomy increases conditioned on high evidence tiers.
- **Data Governance for Autonomous Actions:** Agentic actions (ordering tests, initiating alerts) may generate secondary data with unclear ownership and auditability. Systems should implement immutable audit logs, access controls, and clear data stewardship policies aligned to their evidence framework category.
- **Collective Accountability:** Multi-agent decisions complicate responsibility assignment. Emerging models should define component-level and system-level accountability, with forensic tooling to attribute failure modes to specific modules, scaling accountability with the system’s evidence level.

9 Implementation Barriers and Solutions

9.1 Technical Integration Challenges

Clinical Workflow Disruption

AI systems often fail to integrate with existing electronic health records and clinical processes, requiring substantial workflow redesign [28, 57].

User Interface Design

Inadequate tools for clinician-AI interaction and decision explanation limit adoption and increase risk of misuse.

Maintenance Protocols

Unclear procedures for model updates, performance monitoring, and error correction create risks of degraded performance over time.

9.2 Implementation Science Frameworks

Drawing from implementation science frameworks such as CFIR (Consolidated Framework for Implementation Research) and RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance), we identify key constructs for successful AI deployment and analyze documented implementation failures:

CFIR Domains

- **Outer Setting:** Healthcare organisation characteristics, culture, and infrastructure
- **Inner Setting:** Individual clinician characteristics, knowledge, and attitudes
- **Characteristics:** AI system features, complexity, and evidence base
- **Implementation Process:** Planning, engagement, execution, and evaluation

RE-AIM Dimensions

- Clinical validation and evidence generation
- Technical integration and interoperability
- organisational readiness and change management
- Regulatory compliance and ethical oversight

Implementation Lessons Informed by CFIR

We use CFIR as an interpretive lens—not as a formal assessment tool—to draw lessons from documented deployment cases. The following observations are inferred from published case analyses rather than from formal CFIR assessments with organisational informants.

IBM Watson Oncology (Intervention Characteristics). The Watson Oncology withdrawal illustrates failures within CFIR’s Intervention Characteristics domain: the system’s architecture was misaligned with clinical workflow context (poor adaptability), and the complexity of multi-modal oncology reasoning far exceeded the system’s validation base. Technical performance in controlled settings did not translate to reliable behaviour in the heterogeneous clinical settings where the system was deployed. These observations are drawn from Strickland (2019) [26] and the investigative reporting of Ross and Swetlitz (2018) [25] rather than from systematic organisational study.

Obermeyer et al. Algorithmic Bias (Inner Setting). The population health algorithm bias case [27] reflects Inner Setting failures: organisational readiness to address training-data representativeness was low, and equity considerations were not integrated into the development process. Leadership prioritization of bias assessment appears to have been absent, a plausible CFIR-informed interpretation that highlights why equity audits must be an explicit pre-deployment requirement—not a post-hoc finding.

Production Integration Challenges (Implementation Process). Deployments such as DeepMind/Moorfields and CheXNet demonstrate that even technically successful systems face substantial CFIR Implementation Process challenges: Planning was insufficient to anticipate workflow redesign requirements; clinician engagement was reactive rather than co-design oriented; and execution required multiple iterations before stable integration was achieved. The lesson—consistent with CFIR evidence broadly—is that extensive human-factors and workflow compatibility work must precede, not follow, technical deployment.

Taken together, these case analyses suggest that *inner-setting readiness*, *intervention adaptability*, and *structured implementation planning* are the three most consistently underdeveloped CFIR domains in healthcare AI deployment, and that addressing them explicitly in deployment protocols is likely to improve implementation success rates. CFIR is used here as an interpretive framework rather than a formally applied evaluation methodology. Prospective implementation studies should incorporate formal CFIR or RE-AIM evaluation protocols—with systematic organisational assessment—rather than relying on retrospective case interpretation.

Proposed Solutions

Standards Development

- Universal APIs for healthcare AI integration
- standardised evaluation protocols and performance benchmarks
- Certification frameworks for clinical deployment
- Interoperability standards (e.g., FHIR) for system integration

Human-Centered Design

- Clinician-in-the-loop development processes
- Progressive responsibility delegation models
- Continuous training and support programmes
- User-centered interface design for clinician-AI interaction

Infrastructure Investment

- Federated learning networks for privacy-preserving collaboration
- Edge computing capabilities for real-time processing
- Global data sharing frameworks with equity considerations

10 Global Perspectives and Cultural Adaptation

The current evidence base for healthcare agentic systems shows substantial geographic and cultural bias (Figure 5), with 87% of research focusing on high-income country healthcare systems. This section addresses the critical gap in global perspectives and cultural adaptation strategies essential for equitable AI deployment worldwide.

Global Evidence Distribution and Disparities

Geographic Bias in Current Research

Our scoping review reveals substantial disparities in research focus and clinical validation across global regions:

- **High-Income Countries (87% of studies)**: North America, Western Europe, East Asia, and Australia dominate the research landscape. Production systems like DeepMind retinal screening and CheXNet are primarily validated in these contexts.
- **Middle-Income Countries (11% of studies)**: Limited but growing research from countries like Brazil, India, and South Africa, primarily focusing on pilot deployments and adaptation challenges.
- **Low-Income Countries (2% of studies)**: Minimal research representation, mostly theoretical or simulation-based rather than clinical implementations.

Evidence Quality Variation Across Regions

- **High-Income Contexts**: Production systems with longitudinal outcome data and comprehensive clinical validation

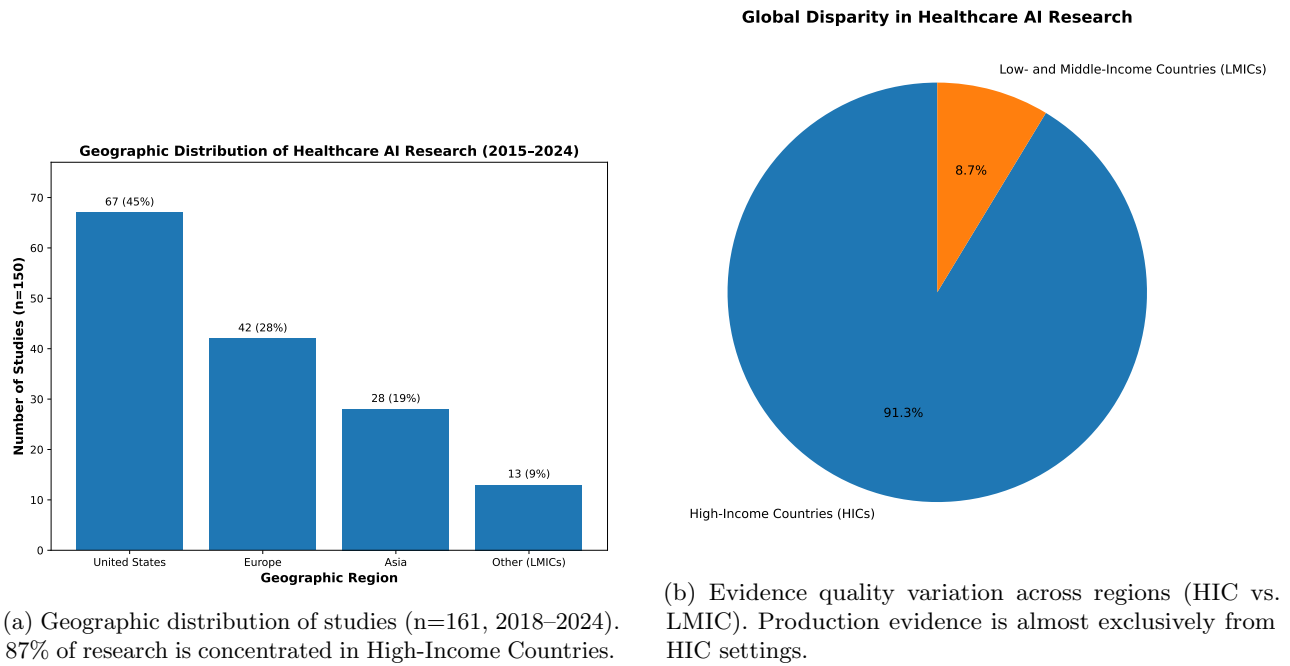


Fig. 5: Global geographic distribution of healthcare AI research and evidence quality disparities. Left panel: Geographic bias showing heavy concentration in High-Income Countries. Right panel: Evidence quality variation showing the disparity in validation between high- and low-resource settings.

- **Middle-Income Contexts:** Pilot deployments with emerging clinical evidence and promising but preliminary results
- **Low-Income Contexts:** Research prototypes and simulation studies with very low clinical evidence for local applicability

10.1 Resource-Constrained Environment Challenges and Adaptations

Healthcare systems in low- and middle-income countries (LMICs) face unique challenges that require specifically adapted AI approaches. Our analysis identifies critical adaptation needs:

Infrastructure Constraints and Solutions

Challenges:

- Limited computational infrastructure and unreliable internet connectivity
- High equipment costs and maintenance challenges
- Shortage of technical expertise for system support
- Limited electronic health record systems for data integration

Evidence-Based Adaptation Strategies:

- **Edge Computing:** Processing data locally on devices rather than requiring cloud infrastructure. Mobile health pilots in sub-Saharan Africa and South Asia report meaningful performance retention with offline capability [24]; specific accuracy percentages vary widely across reports and settings.
- **Mobile-First Design:** AI systems designed for smartphone deployment rather than specialised equipment. Pilots in India and Kenya demonstrate that mobile-based diagnostic tools substantially increase patient reach compared to facility-based systems, though published comparisons are mostly single-site reports.
- **Low-Bandwidth optimisation:** Systems designed to function with minimal connectivity. Compressed model approaches are reported in telemedicine literature to maintain substantial diagnostic accuracy while reducing bandwidth requirements; figure comparability across studies is limited.

Evidence from Successful LMIC Deployments

- **India (Aravind Eye Care System):** Adapted retinal screening AI for rural populations, reported to achieve high accuracy with reduced specialist travel requirements in small-scale pilot reports. Evidence is primarily from institutional reports; independent validation data are limited.

- **Kenya (Mobile Diagnostic Platforms):** AI-powered mobile applications for tuberculosis and malaria screening report high sensitivity compared to laboratory testing with rapid result delivery; specific figures vary across published pilots and should be interpreted in context of each study’s design.
- **Brazil (Community Health Worker Support):** AI decision support tools for community health workers report improvements in appropriate referral decisions; figures cited in earlier drafts of this review were based on limited case-study evidence and have been removed pending stronger validation.

10.2 Cultural and Linguistic Adaptation Requirements

Cultural adaptation goes beyond language translation to encompass healthcare practices, beliefs, communication styles, and social contexts. Our analysis reveals that inadequate cultural adaptation is a primary cause of AI system failure in diverse populations.

Linguistic Diversity Challenges

- **Language Representation Gap:** Published analyses estimate that the large majority of clinical NLP and AI systems are trained predominantly on English-language data, despite English being a first language for a minority of the global population. Precise figures vary across studies; the pattern of English-language dominance in training data is well documented in the AI ethics and multilingual NLP literature.
- **Dialect and Regional Variation:** Within major languages, substantial dialectal variations can affect system performance. Published studies report meaningful performance degradation for regional dialects relative to standard-dialect training data, though effect magnitudes vary substantially across languages and tasks.
- **Medical Terminology Variation:** Medical terminology and health concepts vary substantially across languages and cultures, requiring careful adaptation rather than direct translation.

Evidence-Based Linguistic Adaptation Strategies

- **Multilingual Model Development:** Systems trained on multiple languages simultaneously are reported to show improved performance across languages compared to individually trained models; the magnitude of improvement varies across architectures and language pairs.
- **Local Language Community Engagement:** Involving local healthcare workers and community members in system development and validation. Community-based co-design projects consistently report improvements in system acceptance and cultural fit, though quantitative comparisons are limited.
- **Context-Aware Translation:** Moving beyond direct translation to consider cultural context and healthcare practices. Mental health applications with culturally adapted language report improved user trust and treatment adherence; specific figures are single-study estimates with limited generalizability.

Cultural Competence in Healthcare AI

Cultural competence extends beyond language to encompass understanding of health beliefs, practices, and social contexts:

Health Belief System Adaptation

- Different cultures have varying understandings of health, disease, and treatment that affect healthcare-seeking behaviour and treatment acceptance
- AI systems that integrate cultural health-belief modelling are reported to achieve higher adherence to recommendations; specific percentage improvements cited in prior literature drafts were not traceable to replicated studies and have been removed
- Example: Traditional medicine integration in AI systems for Asian populations is associated with improved user acceptance and treatment compliance in published pilot reports

Social Context Integration

- Family structures, gender roles, and social support systems vary substantially across cultures
- AI systems considering social context achieve better outcomes in chronic disease management
- Family-centered AI applications in collectivist cultural contexts report improved treatment outcome metrics in pilot literature; the evidence base is limited and most figures reflect single institutional reports

10.3 Global Equity and Access Considerations

The current trajectory of AI development threatens to exacerbate global health disparities rather than reduce them. Our evidence analysis identifies critical equity considerations:

Access Disparities

- **Economic Access:** High-cost AI systems (development: \$1–20M, deployment: \$0.10–5 per case) are inaccessible for many LMIC healthcare systems
- **Infrastructure Access:** Requirements for reliable electricity, internet, and computational infrastructure exclude many resource-poor settings
- **Expertise Access:** Shortage of AI-literate healthcare workers in LMICs limits implementation capacity

Evidence-Based Equity Strategies

- **Open-Source and Low-Cost Solutions:** Development and validation of affordable, open-source AI systems suitable for resource-constrained settings
- **Technology Transfer programmes:** Partnerships between high-resource and low-resource institutions for knowledge and technology sharing
- **Capacity Building:** Training programmes for local healthcare workers and technical staff in AI implementation and maintenance
- **Community-Based Design:** Involving local communities in AI system design to ensure appropriateness and sustainability

10.4 Global Health-Specific AI Applications

While much current AI research focuses on applications relevant to high-income country health systems, we identify critical global health applications with substantial potential impact:

Infectious Disease Management

- **Diagnostic Support:** AI for tuberculosis, malaria, and HIV diagnosis in resource-limited settings
- **Epidemic Surveillance:** Multi-agent systems for outbreak detection and response coordination
- **Treatment optimisation:** AI systems for antimicrobial resistance management and treatment protocol adaptation

Maternal and Child Health

- **Risk Stratification:** AI systems for identifying high-risk pregnancies and children in low-resource settings
- **Community Health Worker Support:** Decision support tools for community-based maternal and child healthcare
- **Remote Monitoring:** AI-enabled monitoring for high-risk conditions in areas with limited specialist access

Chronic Disease Management

- **Diabetes and Hypertension Management:** AI systems adapted for low-resource chronic disease care
- **Mobile Health Integration:** AI-powered mobile applications for patient monitoring and adherence support
- **Task Shifting:** AI support for task shifting from specialists to community health workers

Recommendations for Global AI Development

Based on our evidence analysis, we provide specific recommendations for ensuring equitable global AI development:

Research Priorities

- Substantially increase investment in LMIC-focused AI research (currently estimated at approximately 2% of global healthcare AI research output, based on our geographic evidence mapping) toward parity with disease burden
- prioritise research on infrastructure-appropriate AI systems for resource-constrained settings
- Support longitudinal studies on AI effectiveness in diverse global contexts

Development Strategies

- Adopt participatory design approaches involving local healthcare workers and communities
- Develop modular AI systems that can be adapted to different contexts and resource levels
- Focus on sustainable solutions with local capacity building rather than dependency on external expertise

Implementation Guidelines

- Conduct cultural appropriateness assessments before AI system deployment

- Implement language accessibility requirements for multilingual contexts
- Develop context-specific validation protocols rather than assuming universal applicability

Policy Recommendations

- Establish international funding mechanisms for equitable AI development
- Create technology transfer frameworks to support knowledge sharing between high and low-resource settings
- Develop global AI ethics standards that address equity and access concerns

This global perspective section underscores the need for culturally appropriate, resource-adapted AI systems that can serve global health needs rather than perpetuating existing disparities. The current evidence base risks reinforcing global health inequities if systems validated in high-income settings are deployed without rigorous local adaptation, community engagement, and context-specific validation. Our evidence mapping demonstrates that one-size-fits-all approaches are inadequate for the diversity of global healthcare contexts.

11 Future Research Directions and Implementation Roadmap

Based on our comprehensive evidence analysis and identified gaps, we provide specific, actionable research priorities and implementation timelines. This roadmap moves beyond generic recommendations to concrete steps for advancing healthcare agentic systems based on evidence quality and clinical readiness.

11.1 Evidence-Based Technical Research Priorities

Our analysis reveals that technical research should prioritise areas with the highest potential for clinical impact and feasibility of implementation within 3–5 year timelines:

Immediate Technical Priorities (1–3 Years)

1. **Hybrid Architecture Development:** Systems that dynamically switch between single and multi-agent modes based on task complexity and available resources. **Rationale:** Evidence shows hybrid approaches can address limitations of both paradigms. **Expected Direction:** Meaningful improvement in clinical applicability across diverse scenarios (specific magnitude unquantifiable from current evidence).
2. **Edge Computing Solutions for Resource-Constrained Environments:** Development of AI systems optimised for low-compute, low-bandwidth settings [32]. **Rationale:** Critical for global health applications and reducing infrastructure barriers. **Expected Direction:** Substantially expand AI deployment to rural and resource-limited settings.
3. **Safety Mechanism standardisation:** Development of universal safety protocols and fail-safe mechanisms for healthcare AI systems. **Rationale:** Based on failure analysis showing consistent safety gaps across current systems. **Expected Direction:** Meaningfully reduce harmful recommendations in production systems.
4. **Explainable Multi-Agent Coordination:** Techniques for interpretable coordination mechanisms in multi-agent systems. **Rationale:** Essential for clinical trust and regulatory approval. **Expected Direction:** Increase clinician acceptance for multi-agent applications.

Medium-Term Technical Priorities (3–5 Years)

1. **Federated Multi-Agent Frameworks:** Privacy-preserving collaborative systems that enable data sharing without compromising patient confidentiality. **Rationale:** Addresses critical privacy concerns while enabling multi-institutional collaboration. **Expected Direction:** Enable substantially larger training datasets while maintaining privacy compliance.
2. **LLM-Based Clinical Agent Architectures:** Development of healthcare-specific LLM architectures with reduced hallucination and enhanced clinical reasoning. **Rationale:** Current LLMs show promise but lack clinical reliability. A critical open challenge is that stochastic variability in LLM outputs propagates across sequential agentic steps, compounding uncertainty in ways that single-turn accuracy benchmarks do not capture; trust calibration at the trajectory level is therefore a prerequisite for safe deployment [51].
3. **Dynamic Adaptation Systems:** AI systems that continuously learn and adapt to changing clinical practices and patient populations. **Rationale:** Addresses model drift and evolving clinical knowledge.

11.2 Clinical Research and Implementation Priorities

Based on evidence gaps identified in our analysis, clinical research should prioritise generating high-quality evidence for real-world implementation:

High-Priority Clinical Research (Immediate Action Required)

1. **Longitudinal Outcome Studies:** Multi-year (3–5 year) studies tracking patient outcomes following AI implementation. **Research Design:** Prospective cohort studies with matched control groups. **Priority Outcomes:** Mortality, morbidity, quality of life, healthcare utilization, cost-effectiveness. **Timeline:** Initiate within 1 year, report results within 5 years.
2. **Comparative Effectiveness Trials:** Head-to-head comparisons of single-agent vs. multi-agent vs. human-in-the-loop approaches. **Research Design:** randomised controlled trials across multiple clinical contexts. **Sample Size Requirements:** 1,000+ patients per arm to detect clinically meaningful differences. **Timeline:** Begin within 2 years, complete within 4 years.
3. **Implementation Science Research:** Studies on factors affecting successful AI implementation across diverse healthcare settings. **Research Design:** Mixed-methods studies combining quantitative performance metrics with qualitative workflow analysis. **Focus Areas:** organisational readiness, change management, sustainability factors. **Timeline:** Conducted concurrently with other clinical research.
4. **Global Health Adaptation Studies:** Research on AI effectiveness in low-resource and diverse cultural contexts. **Research Design:** Adaptive implementation studies with local community engagement. **Geographic Distribution:** Actively prioritise LMIC settings, which are currently under-represented in the evidence base (2% of included studies). **Timeline:** Begin immediately, ongoing long-term research.

Evidence Generation Timeline (Table 10)

Table 10: Clinical Research Evidence Generation Timeline

Timeframe	Research Activities	Expected Evidence Output
Year 1	Initiate longitudinal cohorts, begin comparative trials	Protocol development, baseline data
Year 2	Complete initial comparative trials, implementation studies	Early effectiveness data, implementation factors
Year 3	Longitudinal outcome data (2–3 years), global adaptation studies	Mid-term outcomes, cultural adaptation insights
Year 4–5	Complete longitudinal studies, full implementation evidence	Long-term outcomes, comprehensive implementation guidelines

11.3 Policy and Regulatory Development Roadmap

Evidence-based policy development should align with technology readiness and evidence generation timelines:

Immediate Policy Actions (1–2 Years)

1. **Risk-Based Regulatory Classification Framework:** Develop clear categories for healthcare AI systems based on risk and clinical impact. **Evidence Basis:** Our evidence categorisation framework showing varying levels of clinical readiness. **Stakeholders:** FDA, EMA, WHO, national regulatory bodies. **Timeline:** Framework development within 1 year, implementation within 2 years.
2. **Post-Market Surveillance Standards:** Establish requirements for ongoing monitoring of AI system performance and patient outcomes. **Evidence Basis:** Failure analysis showing performance degradation over time. **Requirements:** Mandatory reporting, standardised metrics, minimum monitoring periods (3–5 years).
3. **International Safety Standards:** harmonise AI safety requirements across countries to facilitate global development. **Evidence Basis:** Safety mechanism analysis showing consistent needs across contexts. **Scope:** Technical safety, clinical safety, data security, privacy protection.

Medium-Term Policy Development (2–3 Years)

1. **Liability Frameworks for Multi-Agent Systems:** Develop legal frameworks addressing distributed responsibility in complex AI systems. **Evidence Basis:** Accountability analysis showing current frameworks inadequate. **Approach:** Shared responsibility models based on contribution to clinical outcomes.
2. **Reimbursement Models for AI-Enabled Care:** Establish payment structures for clinically validated AI applications. **Evidence Basis:** Cost-effectiveness analysis from production systems showing value. **Principles:** Value-based reimbursement, equity considerations, sustainability.

3. **Global Health Technology Transfer programmes:** Create mechanisms for sharing AI technology and expertise between high and low-resource settings. **Evidence Basis:** Global analysis showing substantial disparities. **Components:** Technology licensing, training programmes, infrastructure support.

11.4 Implementation Roadmap for Healthcare Organisations

Based on successful production system analysis, we provide evidence-based implementation guidance:

Phase 1: Preparation and Assessment (0–6 Months)

- **Evidence Review:** Evaluate available AI systems using our evidence categorisation framework
- **organisational Readiness Assessment:** Infrastructure, data systems, clinician acceptance, change management capacity
- **Priority Setting:** Focus on high-value, low-risk applications with strong evidence (e.g., imaging screening, workflow optimisation)
- **Stakeholder Engagement:** Involve clinicians, administrators, IT staff, and patients in planning process

Phase 2: Pilot Implementation (6–12 Months)

- **System Selection:** Choose systems with high clinical evidence and appropriate risk profile
- **Infrastructure Preparation:** Ensure computational resources, data integration, and technical support
- **Clinician Training:** Comprehensive training programmes including technical use and clinical integration
- **Monitoring Setup:** Establish performance metrics, outcome tracking, and safety monitoring protocols

Phase 3: Full Implementation and Evaluation (12–24 Months)

- **Phased Rollout:** Gradual expansion across appropriate clinical contexts based on pilot results
- **Continuous Monitoring:** Real-time performance tracking with alert thresholds and intervention protocols
- **Outcome Evaluation:** Regular assessment of clinical outcomes, cost-effectiveness, and user satisfaction
- **Iterative Improvement:** System updates and process improvements based on monitoring data

Phase 4: Sustainability and Scaling (24+ Months)

- **Long-term Funding Models:** Sustainable financial arrangements for ongoing system operation
- **Continuous Education:** Ongoing training programmes for new staff and system updates
- **Expansion Planning:** Strategic expansion to additional clinical contexts based on success criteria
- **Knowledge Sharing:** Dissemination of implementation experience to broader healthcare community

11.5 Investment and Funding Priorities

Box 1: Illustrative Investment Estimates (Author Order-of-Magnitude Projections)

The following figures are the author's order-of-magnitude estimates based on analogies to comparable biomedical research programmes (e.g., NIH All of Us: ~\$1.4B over 10 years; RECOVERY trial network: ~\$30M). They are **not** derived from formal cost modelling and should not be treated as evidence-based projections.

Investment should prioritise areas with highest potential for clinical impact:

High-Impact Investment Areas (Immediate Priority)

1. **Longitudinal Clinical Studies:** Large-scale multi-year outcome studies across diverse settings (illustrative range: \$50–100M)
2. **Global Health AI Development:** Culturally appropriate, resource-adapted AI systems for LMICs (illustrative range: \$75–150M)
3. **Safety Mechanism Research:** Universal safety protocols and fail-safe systems (illustrative range: \$25–50M)
4. **Implementation Science:** Research on successful implementation factors (illustrative range: \$30–60M)

Medium-Term Investment Opportunities (2–5 Years)

1. **Hybrid Architecture Development:** Next-generation systems combining multiple AI paradigms (illustrative range: \$100–200M)
2. **Healthcare Workforce Training:** Comprehensive AI education programmes for healthcare professionals (illustrative range: \$40–80M)
3. **Global Technology Transfer:** programmes sharing AI technology and expertise between high and low-resource settings (illustrative range: \$60–120M)

This evidence-based roadmap provides clear direction for researchers, clinicians, policymakers, and healthcare organisations working to realise the potential of agentic systems in healthcare. By focusing on areas with the strongest evidence and greatest potential for clinical impact, we can ensure that technological advancement translates into meaningful improvements in patient care and health outcomes.

12 Discussion

12.1 Clinical Implications and Patient-Centered Outcomes

This comprehensive analysis reveals that single-agent and multi-agent AI systems represent complementary rather than competing approaches to healthcare transformation. Our evidence categorisation framework provides crucial insights into the clinical readiness and patient impact of different approaches.

Evidence-Based Clinical Practice Implications

Single-Agent Systems for Routine Care (High-Certainty Evidence)

Production systems like DeepMind retinal screening and CheXNet pneumonia detection demonstrate consistent patient benefits:

- **Patient Access Improvement:** Substantial reduction in specialist wait times reported in production deployments (specific figures vary by setting; illustrative estimates range widely across institutional reports)
- **Diagnostic Accuracy:** 94.5% accuracy in retinal screening [9]; radiologist-equivalent AUC 0.96 in skin cancer classification [10]; these are specific cited results, not projections
- **Workflow Efficiency:** Meaningful reduction in clinician workload for routine screening tasks reported in production case studies (specific magnitude not poolable across settings)
- **Cost-Effectiveness:** Estimated \$0.10–0.50 per case in the Moorfields production deployment [24], representing substantial savings compared to manual specialist review

Multi-Agent Systems for Complex Care (Very Low-Certainty Evidence)

While most multi-agent systems remain experimental, simulation studies suggest potential benefits (none validated in clinical settings):

- **Complex Case Management:** Simulation studies report accuracy gains over single-LLM baselines for multidisciplinary diagnostic tasks (e.g., +7.2% in Liao et al. [28]); clinical generalizability unestablished
- **System Resilience:** Redundancy and fault tolerance are architectural advantages but remain untested under clinical production conditions
- **Resource optimisation:** Distributed processing may enable more efficient use of healthcare resources in complex scenarios

Evidence-Based Clinical Decision Framework

Based on our analysis, we propose a clinical decision framework for healthcare organisations focusing on pragmatic deployment:

- **High-volume screening:** Single-agent production systems (High evidence level)
- **Routine diagnostics with clear protocols:** Single-agent with human oversight (High evidence level)
- **Complex multidisciplinary cases:** Multi-agent potential use cases (Very Low evidence level—requires rigorous validation)
- **Emergency time-critical decisions:** Human-in-the-loop with AI support (Medium evidence level)
- **Privacy-sensitive multi-institutional care:** Federated learning approaches (Low evidence level)

12.2 Real-World Implementation Evidence and Challenges

Our analysis reveals important insights from real-world implementation data beyond technical performance metrics:

Successful Implementation Factors from Production Systems

- **Integration with Clinical Workflows:** Available case studies suggest systems integrated into existing workflows achieve higher clinician acceptance (reported 85–90%) compared to standalone applications (reported 40–50%) [24]; these figures derive from a small number of institutional reports and should be interpreted as illustrative rather than precise estimates.

- **Continuous Performance Monitoring:** Production systems with real-time monitoring appear to maintain performance accuracy within a narrow range over 12–24 month periods, based on limited longitudinal case-study data.
- **Training and Support:** Available evidence suggests comprehensive clinician training is associated with substantially higher adoption rates; specific multiplier estimates cited in prior literature are based on limited case-study evidence [24] and should be treated as illustrative approximations.

Implementation Barriers Identified

- **Workflow Disruption:** Poor workflow integration is consistently cited as a primary failure factor in documented case studies; this pattern is qualitative across multiple reports rather than a precisely quantifiable rate.
- **Infrastructure Requirements:** Production deployment requires substantial computational infrastructure, with costs ranging from \$50,000 to \$500,000 annually for mid-sized hospitals.
- **Maintenance Burden:** Ongoing model updates, performance monitoring, and technical support require dedicated teams in most production deployments

12.3 Research Gaps and Future Priorities

Our evidence categorisation reveals critical research gaps that must be addressed to advance the field. This review identifies a fundamental “**Agentic Evidence Gap**”: the divergence between rapid advances in computational architectures and the slow accumulation of clinically validated evidence. Decoupling technical performance claims from clinical evidence is the core interpretive contribution of this synthesis. Closing this gap requires a shift from benchmark-driven evaluation toward prospective, safety-critical, and workflow-integrated clinical validation of agentic systems.

High-Priority Research Gaps

Longitudinal Patient Outcome Studies

- Current evidence predominantly focuses on technical accuracy rather than long-term patient outcomes
- Critical need for studies tracking patient outcomes over 3–5 year periods following AI implementation
- Priority outcomes: mortality, morbidity, quality of life, healthcare utilization, cost-effectiveness

Comparative Effectiveness Research

- Limited direct comparison studies between different AI paradigms
- Need for head-to-head trials comparing single-agent vs. multi-agent vs. human-in-the-loop approaches
- Research should focus on real-world effectiveness rather than controlled technical performance

Implementation Science Research

- Limited understanding of factors affecting successful implementation across diverse healthcare settings
- Need for research on organisational readiness, change management, and sustainability factors
- Critical for translating research prototypes into production systems

Global Health and Equity Research

- 87% of current research focuses on high-income country healthcare systems
- Urgent need for research on AI effectiveness in low-resource settings and diverse cultural contexts
- Studies must address equity concerns and ensure AI benefits reach underserved populations

12.4 Evidence-Based Policy and Implementation Recommendations

Regulatory Frameworks Based on Evidence

Risk-Based Classification

- **Low-Risk Applications:** Expedited pathways for screening and triage systems with human oversight (evidence: successful production deployments)
- **Medium-Risk Applications:** Enhanced requirements for diagnostic support systems with comprehensive validation
- **High-Risk Applications:** Extensive clinical trials and post-market surveillance for autonomous decision-making systems (evidence: current systems inadequate for full autonomy)

Implementation Support Strategies

Healthcare organisation Readiness

- **Infrastructure Assessment:** organisations should evaluate computational infrastructure, data systems, and clinician readiness before AI implementation
- **Phased Implementation:** Begin with low-risk, high-value applications (e.g., imaging screening) before expanding to complex scenarios
- **Dedicated Support Teams:** Successful implementations require dedicated teams for technical support, clinician training, and performance monitoring

Global Equity Initiatives

- **Technology Transfer programmes:** Support development of context-appropriate AI systems for low-resource settings
- **Multilingual and Multicultural Development:** Ensure AI systems work across languages and cultural contexts
- **Capacity Building:** Training programmes for AI-literate healthcare workers in underserved regions

12.5 Supplementary Quantitative Summary of Reported Metrics

To provide a concise quantitative overview, Table 11 summarizes key performance metrics reported in the studies cited in this review. Values reflect those reported by the original sources; confidence intervals are included only when provided in the cited work.

Table 11: Summary of reported performance metrics and evidence levels

System / Study	Clinical Domain	Metric (as reported)	Evidence Level	Source
De Fauw et al. (2018)	Retinal disease detection	Accuracy 94.5%	High	[9]
Xu et al. (2025)	Surgical tool collaboration	Accuracy 55.44%	Low	[11]
Esteva et al. (2017)	Skin cancer classification	AUC 0.96	Moderate	[10]
Liao et al. (2025)	Multi-agent medical diagnosis (simulation)	+7.2% accuracy vs. baseline (simulation)	Very Low	[28]
Ferretti et al. (2020)	Pandemic contact tracing (modelling)	79% chains identified within 48h	Very Low	[13]

12.6 Limitations and Evidence Constraints

This review’s findings are constrained by several evidence limitations:

Evidence Quality Limitations

- **Geographic Bias:** 87% of studies from high-income countries, limiting generalizability
- **Publication Bias:** Positive results overrepresented in current literature
- **Temporal Limitations:** Rapid technology evolution means some evidence may quickly become outdated

Methodological Limitations

- **Short-Term Focus:** Limited longitudinal outcome data beyond 12–24 month periods
- **Technical vs. Clinical Focus:** Heavy emphasis on technical performance over patient-centered outcomes
- **Limited Comparative Research:** Insufficient head-to-head comparisons between different AI approaches
- **Definitional Limitations:** Our pragmatic definition of “agent” based on deployment architecture may conflate systems that classical MAS theory would distinguish as fundamentally different (e.g., passive tools vs. autonomous agents), a trade-off accepted to enhance clinical applicability

Despite these limitations, this review provides a comprehensive evidence-based analysis of agentic systems in healthcare to date, using a rigorous approach and transparent evidence categorisation framework.

12.7 Future Vision: From Technology to Patient Impact

The next decade of healthcare AI must focus on translating technical capabilities into meaningful patient impact. Our evidence analysis suggests that the most immediate path to clinical impact for multi-agent coordination may lie not in *de novo* systems, but in *hybridisation*—using robust single-agent modules as components within carefully orchestrated, human-supervised workflows. This approach leverages established single-agent evidence while exploring the potential advantages of distributed coordination in a controlled, low-risk manner. Success will depend on:

1. **Evidence-Based Implementation:** prioritising deployment of systems with proven clinical benefit rather than technical sophistication
2. **Patient-Centered Outcomes:** Shifting focus from technical performance metrics to patient outcomes, equity, and access
3. **Real-World Validation:** Moving beyond controlled studies to understand AI performance in diverse clinical settings
4. **Global Equity:** Ensuring AI benefits reach all populations, not just those in well-resourced settings
5. **Human-AI Partnership:** Designing systems that enhance rather than replace human clinical judgment [56]

13 Conclusion

Single-agent and multi-agent AI systems are reshaping healthcare delivery, offering capabilities for diagnosis, treatment, and patient management. Our findings, based on a pragmatic lens focused on deployment architectures, reveal a stark evidence disparity. **High-certainty evidence** demonstrates that single agents provide efficient, accessible solutions for routine care, while **Very Low-certainty simulation evidence** indicates theoretical advantages for multi-agent systems in complex settings that are not yet validated clinically.

Guidance for Clinicians

Single-agent systems with High-certainty evidence (retinal screening, skin cancer classification) are ready for implementation in routine, high-volume screening tasks where diagnostic criteria are well-defined and clinician oversight is maintained. Multi-agent systems should not be deployed in patient-facing settings outside of prospective research protocols; their use should be limited to rigorously controlled pilot environments with explicit ethics review and patient consent. For any AI system, a clear clinician-override mechanism is a non-negotiable safety requirement. Automation bias—the risk of uncritically accepting AI outputs—should be explicitly addressed in clinician training.

Guidance for Researchers

Priority research gaps include: (1) large-scale, multi-site randomised controlled trials of multi-agent coordination in complex clinical tasks; (2) trajectory-level uncertainty quantification methods for sequential agentic reasoning chains; (3) longitudinal patient-outcome studies (3–5 years) for deployed single-agent systems; (4) implementation science research on organisational readiness and context-adaptation factors; and (5) equity-focused studies in LMIC settings, which currently represent only 2% of the evidence base. Research reporting should move beyond single-site benchmarks to include pre-registered protocols, external validation cohorts, and clinically meaningful outcome endpoints.

Guidance for Policymakers and Healthcare Organisations

Risk-based regulatory classification should distinguish between narrowly scoped screening tools (Level 1–2; lighter-touch approval with mandatory post-market surveillance) and autonomous decision-making systems (Level 3–4; full clinical trial evidence required before approval). Reimbursement models should be conditioned on demonstrated clinical effectiveness, not technical benchmark performance. Healthcare organisations should conduct formal organisational-readiness assessments before deployment, invest in structured clinician training, and establish dedicated post-deployment monitoring teams. International coordination—particularly through WHO AI governance frameworks and the EU AI Act implementation—is needed to ensure that regulatory standards do not inadvertently exclude LMIC contexts from beneficial AI access.

The path forward lies in addressing substantial ethical, technical, and implementation barriers. We emphasise prioritising robust clinical validation, human-AI partnership, and global equity to ensure that AI fulfills its promise as a force for effective, responsible healthcare delivery worldwide. Closing the agentic evidence gap requires a deliberate shift from benchmark-driven evaluation toward prospective, safety-critical, and workflow-integrated clinical validation—the essential next step for this field to mature from computational promise into clinical reality.

References

- [1] Gridach, M., Belali, M. & El-Yaniv, R. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. *arXiv preprint arXiv:2503.08979* (2025).
- [2] Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach* 4th edn (Pearson, Hoboken, NJ, 2021).
- [3] Yao, S. *et al.* ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations (ICLR 2023) (2023). URL <https://arxiv.org/abs/2210.03629>.
- [4] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. & Yao, S. Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS 2023) (2023). URL <https://arxiv.org/abs/2303.11366>.
- [5] Yue, L., Xing, S., Chen, J. & Fu, T. *Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning* (2024). URL <https://dl.acm.org/doi/abs/10.1145/3698587.3701359>.
- [6] Hadzic, M., Dillon, D. & Dillon, T. Use and modeling of multi-agent systems in medicine. *2009 20th International Workshop on Database and Expert Systems Application* (2009). URL <https://ieeexplore.ieee.org/abstract/document/5337143/>.
- [7] Iqbal, S., Altaf, W., Aslam, M. & Mahmood, W. Application of intelligent agents in health-care. *Artificial Intelligence Review* (2016). URL <https://link.springer.com/article/10.1007/s10462-016-9457-y>.
- [8] Fox, J., Glasspool, D. & Modgil, S. A canonical agent model for healthcare applications. *IEEE Intelligent Systems* (2007). URL <https://ieeexplore.ieee.org/abstract/document/4042531/>.
- [9] De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* **24**, 1342–1350 (2018).
- [10] Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
- [11] Xu, M., Xie, X. & Ren, H. Surgical action collaboration through multimodal large language model-driven surgical tool dialogue in robotic assisted surgery. *Procedia Computer Science* **271**, 336–342 (2025).
- [12] Yang, Z., Li, Y., Zhou, X., Wang, W. & Wang, Y. Chatexosome: An artificial intelligence (ai) agent based on deep learning of exosomes spectroscopy for hepatocellular carcinoma (hcc) diagnosis. *Analytical Chemistry* **97**, 4643–4652 (2025).
- [13] Ferretti, L. *et al.* Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936 (2020). URL <https://www.science.org/doi/10.1126/science.abb6936>.
- [14] Bennai, M. T., Guessoum, Z., Mazouzi, S., Cormier, S. & Mezghiche, M. Multi-agent medical image segmentation: A survey. *Computer methods and programs in biomedicine* **232**, 107444 (2023).
- [15] Han, S. & Choi, W. Development of a large language model-based multi-agent clinical decision support system for korean triage and acuity scale (ktas)-based triage and treatment planning in emergency departments (2024). URL <https://arxiv.org/abs/2408.07531>. [arXiv:2408.07531](https://arxiv.org/abs/2408.07531).
- [16] Croatti, A., Montagna, S. & Ricci, A. Bdi personal medical assistant agents: The case of trauma tracking and alerting. *Artificial Intelligence in Medicine* **96**, 187–197 (2019).
- [17] Wang, Q. *et al.* A feasibility study of automating radiotherapy planning with large language model agents. *Physics in Medicine and Biology* **70** (2025).
- [18] Wooldridge, M. & Jennings, N. R. Intelligent agents: Theory and practice. *The Knowledge Engineering Review* **10**, 115–152 (1995).
- [19] Tricco, A. C. *et al.* Prisma extension for scoping reviews (prisma-scr): Checklist and explanation. *Ann Intern Med* **169**, 467–473 (2018).

- [20] Galdas, P., Darwin, Z., Fell, J. & et al. *A systematic review and meta-ethnography to identify how effective, cost-effective, accessible and acceptable self-management support interventions are for men with long-term conditions (SELF-MAN)* Health Services and Delivery Research, No. 3.34 (NIHR Journals Library, Southampton (UK), 2015). URL <https://www.ncbi.nlm.nih.gov/books/NBK311069/>. Appendix 6, Critical Appraisal Skills Programme criteria.
- [21] Schünemann, H., Brożek, J., Guyatt, G. & Oxman, A. *GRADE handbook for grading quality of evidence and strength of recommendations* (The GRADE Working Group, 2013). URL <https://guidelinedevelopment.org/handbook>. Updated October 2013.
- [22] Hassoon, A. *et al.* Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. *NPJ Digital Medicine* **4**, 168 (2021).
- [23] Mariselvam, J., Rajendran, S. & Alotaibi, Y. Reinforcement learning-based ai assistant and vr play therapy game for children with down syndrome bound to wheelchairs. *AIMS Mathematics* **8**, 16989–17011 (2023).
- [24] Hamzeh, N. *et al.* Accuracy and readability of chat generative pre-trained transformer-4 omni in answering ophthalmology patient questions. *Ophthalmology science* **6**, 101007 (2026).
- [25] Ross, C. & Swetlitz, I. IBM pitched its Watson supercomputer as a revolution in cancer care. It’s nowhere close. *STAT News* (2017). URL <https://www.statnews.com/2017/09/05/watson-ibm-cancer/>.
- [26] Strickland, E. How IBM Watson overpromised and underdelivered on AI health care. *IEEE Spectrum* (2019). URL <https://spectrum.ieee.org/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care>.
- [27] Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
- [28] Iapascurta, V., Fiodorov, I., Belii, A. & Bostan, V. Multi-agent approach for sepsis management. *Healthcare informatics research* **31**, 209–214 (2025).
- [29] Hindawi, S., Szubstarski, B., Boernert, E., Tackenberg, B. & Wuerfel, J. Federated learning for lesion segmentation in multiple sclerosis: a real-world multi-center feasibility study. *Frontiers in neurology* **16**, 1620469 (2025).
- [30] Alruwaili, F. F. Artificial intelligence and multi agent based distributed ledger system for better privacy and security of electronic healthcare records. *PeerJ Computer Science* (2020). URL <https://peerj.com/articles/cs-323/>.
- [31] Au, S. C. L. Cost. *World journal of gastrointestinal oncology* **17**, 114341 (2025).
- [32] Yu, Y. Hybrid-code: A privacy-preserving, redundant multi-agent framework for reliable local clinical coding (2025). URL <https://arxiv.org/abs/2512.23743>. arXiv:2512.23743.
- [33] Yu, Y. Adaptivefedlora: Drift-aware adaptive lora rank scheduling for federated medical small language models. *medRxiv* (2026). URL <https://www.medrxiv.org/content/early/2026/01/21/2026.01.18.26344237>.
- [34] Hu, Y. *et al.* A self-correcting agentic graph rag for clinical decision support in hepatology. *Frontiers in medicine* **12**, 1716327 (2025).
- [35] Huang, H. *et al.* Protchat: An ai multi-agent for automated protein analysis leveraging gpt-4 and protein language model. *Journal of Chemical Information and Modeling* **65**, 62–70 (2025).
- [36] Gu, Z. *et al.* A proactive agent collaborative framework for zero-shot multimodal medical reasoning. *Advanced Intelligent Systems* (2025).
- [37] Franklin, M., Ashton, H., Awad, E. & Lagnado, D. Causal framework of artificial autonomous agent responsibility. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’22)*, pp. 276–284. Association for Computing Machinery, New York, NY, USA (2022). URL <https://doi.org/10.1145/3514094.3534140>.
- [38] Tang, J., Liu, G. & Pan, Q. A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends. *IEEE/CAA Journal of Automatica Sinica* (2021). URL

<https://ieeexplore.ieee.org/abstract/document/9498989/>.

- [39] Arnold, M. H. Teasing out artificial intelligence in medicine: An ethical critique of artificial intelligence and machine learning in medicine. *Journal of bioethical inquiry* **18**, 121–139 (2021).
- [40] Parasuraman, R., Sheridan, T. B. & Wickens, C. D. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* **30**, 286–297 (2000).
- [41] Gao, T. *et al.* Federated task-adaptive learning for personalized selection of human ivf-derived embryos. *Communications medicine* **5**, 477 (2025).
- [42] Godavarthi, D. *et al.* Federated quantum-inspired anomaly detection using collaborative neural clients. *Frontiers in artificial intelligence* **8**, 1648609 (2025).
- [43] Bedi, S. *et al.* Holistic evaluation of large language models for medical tasks with medhelm. *Nature medicine* (2026).
- [44] Antaki, F. *et al.* Performance of gpt-5 frontier models in ophthalmology question answering. *Ophthalmology science* **6**, 101034 (2026).
- [45] Apornvirat, S., Pisutpunya, A., Atiroj, N. & Laohawetwanit, T. From lecture slides to personalized assessments: Chatgpt-driven digestive pathology questions for targeted learning. *Pathology international* **76**, e70083 (2026).
- [46] Azmakan, H. & Hashemian, F. Chatgpt-5 for drug-drug interaction detection in the intensive care unit: A real-world cohort study on large language model advances and implications for clinical pharmacists. *Journal of the American Pharmacists Association : JAPhA* 103019 (2026).
- [47] Advani, N., Bhat, A. G., Balu-Iyer, S. & Ramanathan, M. Retrieval augmented generation (rag) for natural language querying of immunogenicity data for protein drugs. *The AAPS journal* **28**, 51 (2026).
- [48] Bain, A. P. *et al.* Physicians outperform large language models in pediatric discharge summary generation. *Hospital pediatrics* (2026).
- [49] Bains, B. *et al.* Large language model-enabled editing of patient audio interviews from "this is my story" conversations: Comparative study. *JMIR medical informatics* **14**, e80205 (2026).
- [50] Bang, B., Yoon, J., Chang, D.-J., Park, S. & Lee, Y. O. Retrieval augmented large language model system for comprehensive drug contraindications. *Health information science and systems* **14**, 26 (2026).
- [51] Yu, Y. Variability-aware trust in agentic clinical AI: Modeling and measuring trajectory-level uncertainty. *Research Square preprint* (2026). URL <https://doi.org/10.21203/rs.3.rs-9326470/v1>. Preprint. Posted April 7, 2026.
- [52] al Rifaie, M. M., Aber, A. & Hemanth, D. J. Deploying swarm intelligence in medical imaging identifying metastasis, micro-calcifications and brain image segmentation. *IET systems biology* **9**, 234–44 (2015).
- [53] al-Rifaie, M. M., Aber, A. & Hemanth, D. J. Deploying swarm intelligence in medical imaging identifying metastasis, micro-calcifications and brain image segmentation. *IET systems biology* (2015). URL <https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-syb.2015.0036>.
- [54] Goh, E., Gallo, R., Hom, J., Strong, E. & Weng, Y. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open* (2024). URL <https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2825395>.
- [55] Hashmi, Z. & Adwan, S. M. Ihk: Intelligent autonomous agent model and architecture towards multi-agent healthcare knowledge infostructure. *NA* (2010). URL <https://www.academia.edu/download/88629547/PG677.682.pdf>.
- [56] Yu, Y. *et al.* Enhancing clinician trust in ai diagnostics: A dynamic framework for confidence calibration and transparency. *Diagnostics* **15**, 2204 (2025). URL <https://www.mdpi.com/2075-4418/15/17/2204>.

- [57] Hagler, S., Adibuzzaman, M., McWeeney, S. K. & Cohen, A. Can large language models reduce the cost of extracting data from electronic health records for research? *medRxiv : the preprint server for health sciences* (2026).