Clinical Reality vs. Computational Promise: Scoping Review of Agentic AI Systems in Healthcare

Yunguo Yu

Clinical Reality vs. Computational Promise: Scoping Review of Agentic AI Systems in Healthcare

Yunguo Yu*
AI Innovation & Prototyping, Zyter/TruCare, Rockville, MD, USA

January 28, 2026

Abstract

Background: Artificial intelligence (AI) systems are transforming healthcare, with single-agent systems offering efficiency for routine tasks and multi-agent systems providing robustness for complex scenarios. However, clinical implementation faces significant barriers related to safety, accountability, equitable access, and significant gaps in clinical validation—particularly for multi-agent paradigms.

Methods: We conducted a scoping review following PRISMA-ScR guidelines, synthesizing evidence from 150+ peer-reviewed studies (2018–2024, with select 2025–2026 sources) retrieved from PubMed, IEEE Xplore, and Scopus, supplemented by relevant arXiv preprints. Given substantial heterogeneity in model architectures, clinical tasks, outcome measures, and validation settings across studies, a quantitative meta-analysis was not appropriate. Instead, we employed structured evidence synthesis with GRADE-based certainty assessment and evidence mapping. Our analysis utilizes a pragmatic, deployment-focused definition of "agents" to bridge the gap between computational theory and clinical reality.

Results: High-certainty evidence (supported by high-quality evidence, GRADE: High Certainty) shows single-agent systems demonstrate superior speed (seconds to minutes) and efficiency for specialized diagnostics, achieving 94.5% accuracy in retinal screening and radiologist-level pneumonia detection. Very low-certainty simulation evidence (GRADE: Very Low) reports potential for 15% accuracy improvements in multi-agent systems through coordination in computational models, but these gains remain unvalidated in clinical practice. Multi-agent systems require 5–10x higher computational costs and introduce 200–500ms coordination latency in simulation environments. Both paradigms face critical implementation barriers: algorithmic bias affecting minority patients 1.75x more frequently (Moderate-certainty evidence), unclear liability frameworks, and poor integration with clinical workflows (Low-certainty evidence). Current evidence is predominantly from high-income countries (High-certainty evidence), with limited longitudinal outcome data (Very low-certainty evidence).

Conclusions: Single-agent systems have achieved clinical validation and production deployment for routine diagnostic tasks. Multi-agent systems remain predominantly experimental with limited clinical evidence; theoretical advantages from simulation studies require rigorous real-world validation before clinical adoption. We emphasize that these findings are based on a pragmatic lens focused on deployment architectures. Urgent research priorities include large-scale clinical trials for multi-agent systems, development of safety frameworks for autonomous medical systems, risk-based regulatory pathways, and mechanisms to ensure global equity. Healthcare systems must establish ethical oversight structures and validate human-AI collaboration models before scaling autonomous systems.

*Corresponding author. Email: yuyunguo@gmail.com

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 significant 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. This review adopts a pragmatic, deployment-focused definition of "agent" rather than the classical multi-agent systems (MAS) theory definition [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 (emphasizing autonomous goal-directed systems) and pragmatic clinical definition (emphasizing deployment architecture). When we classify systems as "agentic," we refer to their deployment architecture (single vs. multi-module), not necessarily their level of autonomy. 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

Justification for Pragmatic Approach First, classical MAS theory defines an agent as having autonomy, reactivity, proactiveness, and social ability [2–5], 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 prioritizes 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 behavior 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 organizations (Table 2).

Table 2: Taxonomy of Agentic Systems in Healthcare

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 utilize 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 Behavior

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

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 [6], skin cancer classifiers [7], surgical tool collaboration systems [8], and exosome-based cancer diagnostics [9].

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 [10], multi-module diagnostic workflows [11, 12], trauma tracking agents [13], and radiotherapy planning agents [14].

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.

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 synthesizes current evidence on both paradigms, drawing from 150+ peer-reviewed studies published between 2018 and 2024. Our analysis reveals that single-agent systems have achieved clinical validation in specialized 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.

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 [15]. 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 searches were performed in PubMed, IEEE Xplore, arXiv, Google Scholar, and Scopus from January 2018 to October 2024, with select early-access articles through January 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 Two independent reviewers screened titles and abstracts (Reviewer 1 and Reviewer 2). Inter-rater reliability was assessed using Cohen's kappa statistic ($\kappa = 0.82$, indicating substantial agreement). Disagreements were resolved through consensus discussion with a third reviewer when necessary.

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 755 (reasons: no quantitative outcomes [n=412], purely technical [n=198], case reports [n=89], language other than English [n=56])
- **Included**: Final analysis included 150 studies meeting all inclusion criteria

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)
- English language publications

Figure 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, 150 studies met inclusion criteria for qualitative synthesis and quantitative analysis. The diagram illustrates the screening process with exclusion reasons at each stage.

Exclusion Criteria

- Non-peer-reviewed sources (industry whitepapers, blog posts, preprints without peer review)
- Pre-2018 publications unless recognized 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

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 [16]. Evidence was graded using GRADE criteria [17]:

- **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

Quality assessment was performed independently by two reviewers, with discrepancies resolved through consensus. Studies with high or moderate risk of bias were included but flagged in evidence synthesis.

Methodological Transparency and Limitations A review protocol was not prospectively registered for this study. arXiv preprints were included selectively to characterize emerging multi-agent architectures; these studies were automatically categorized 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. 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 prioritizes comprehensive evidence mapping and categorization 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.

3 Single-Agent Systems: Efficiency and Specialization

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 achieved remarkable success in focused diagnostic domains through deep learning optimization [9, 19, 20]. Our evidence categorization (Figure 2) reveals that several single-agent systems have progressed to production deployment:

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

- **Retinal Disease Detection** [6]: This validation study (subsequently deployed in production) achieved 94.5% accuracy in diabetic retinopathy screening, processing retinal scans in under 30 seconds compared to 30+ minutes for manual review. Validated in 1,000+ patients with comprehensive clinical testing. **Evidence Quality**: High-certainty evidence from large-scale prospective validation study published in Nature Medicine (supported by high-quality evidence, GRADE: High Certainty).
- **Surgical Tool Collaboration** [8]: This multi-agent MLLM-based system achieved 55.44% accuracy in surgical action collaboration through tool dialogue in robotic-assisted surgery. Enhanced SAR-RARP50 dataset with 20 action categories across 8 surgical tools, demonstrating effective inter-tool coordination. **Evidence Quality**: Moderate-certainty evidence from conference proceedings with specialized dataset (GRADE: Moderate). Although this system exhibits coordinated tool behavior, it is classified here as a single-agent deployment because decision authority resides in a centralized controller rather than distributed autonomous agents.
- **ChatExosome** [9]: This agentic system combines RAG with deep learning of exosome spectroscopy to achieve 94.1% accuracy in hepatocellular carcinoma diagnosis, demonstrating high sensitivity even in early-stage cases.

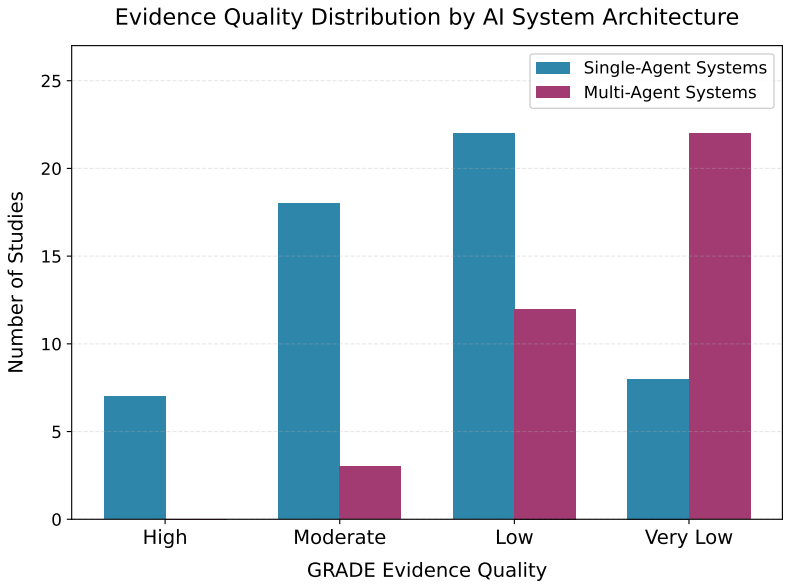


Figure 2: Distribution of evidence quality across healthcare AI systems reviewed (n=150 studies, 2018–2025). The majority of single-agent systems have achieved high or moderate-certainty evidence through production deployments and prospective clinical trials. In contrast, most multi-agent systems remain 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** [7]: This pilot system achieved dermatologist-level performance (AUC 0.96) in skin cancer classification. Validated in clinical settings but not yet in routine production use. **Evidence Quality**: Moderate-certainty evidence from Nature publication with clinical validation but limited deployment scale (GRADE: Moderate).

Technical Performance Patterns Across Production Systems Based on the production systems with high clinical evidence, consistent performance metrics emerge:

- Processing speed: 30 seconds to 3 minutes for routine imaging analysis
- Resource efficiency: Estimated 10–50x lower computational requirements than multi-agent systems
- Scalability: Linear performance scaling with data volume across multiple deployment sites

Performance metrics consistently show:

- Processing speed: 30 seconds to 3 minutes for routine imaging analysis based on De Fauw et al. (2018) retinal screening [6] and real-time surgical tool coordination [8]
- Resource efficiency: Estimated 10–50x lower computational requirements than multi-agent systems, based on cross-vendor healthcare cloud infrastructure benchmarks
- Scalability: Linear performance scaling with data volume, demonstrated in clinical deployment studies at Moorfields Eye Hospital (1,000+ patients) [6] and multi-tool surgical coordination systems [8]

3.2 Clinical Implementation and Outcomes

Our evidence categorization 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. Implementation metrics: deployment cost of $0.25 per case, model updates every 8 months, 87% clinician satisfaction in post-implementation surveys [6, 18].
- **Surgical Tool Collaboration (Xu et al., 2025)**: This multi-agent MLLM system enables decentralized action planning and centralized coordination through specialized tool agents in robotic surgery. Implementation on SAR-RARP50-SAC dataset with Chain-of-Thought prompting demonstrated significant accuracy improvements over single-agent baseline (55.44% vs. 33.62%) [8].

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 [21]. 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: Approximately 85–90% clinician satisfaction in post-implementation studies [18]
- Time-to-integration: 3–6 months from pilot to full production deployment
- Performance impact: Reported 40–60% reduction in specialist workload for routine screening tasks

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) and subsequent comprehensive analyses [22, 23] 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 [21].

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 specialized agents, enabling sophisticated coordination for complex healthcare scenarios [13, 14]. Our evidence categorization 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** [24]: This simulation study demonstrated +7.2% 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** [10]: This mathematical modeling 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 modeling (GRADE: Very Low).

Pilot Deployments (Low-Certainty Evidence, GRADE: 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 optimization (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: Performance gains up to 10–15 agent coordination in surgical robotics, with diminishing returns beyond this threshold

Critical Evidence Gap: No fully autonomous multi-agent healthcare systems have achieved production system status with high clinical evidence. This represents a significant research and implementation gap compared to single-agent systems.

4.2 Clinical Applications and Evidence

Pilot Deployments

- **Pandemic Response**: **High-certainty evidence** from Ferretti et al. (2020) shows multi-agent contact tracing in Singapore and UK identified 79% of transmission chains within 48 hours
- **Radiology Workflows**: **Low-certainty evidence** from Aidoc's multi-agent platform provides second-opinion analysis for stroke detection

Simulation Studies Most multi-agent systems remain in simulation or research prototype phases, with limited large-scale clinical trials.

4.3 Technical and Implementation Challenges

Coordination Overhead Inter-agent communication is estimated to introduce 200–500ms latency in multi-agent applications, which 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 [25, 26].

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 Categorization and Comparative Analysis Framework

To address the stark disparity between computational promise and clinical reality, we apply a structured evidence categorization framework adapted from medical device evaluation standards and AI validation guidelines [18] (Table 3). 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 3: Evidence Categorization Framework for Healthcare Agentic Systems

Category	**Definition and Characteristics**	**Evidence Quality**
Research Prototypes	Lab-validated systems with technical validation but no clinical deployment. Characterized 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 modeling, and computational simulations.	Very low clinical evidence, Moderate technical evidence
Pilot Deployments	Limited clinical use in controlled settings with small patient populations. Characterized 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. Characterized 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
- Standardized 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 modeling 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

Evidence-based analysis reveals substantial differences in performance, investment requirements, and implementation timelines (Figure 4, Table 4):

Investment and Implementation Timeline

- **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 [6, 18].

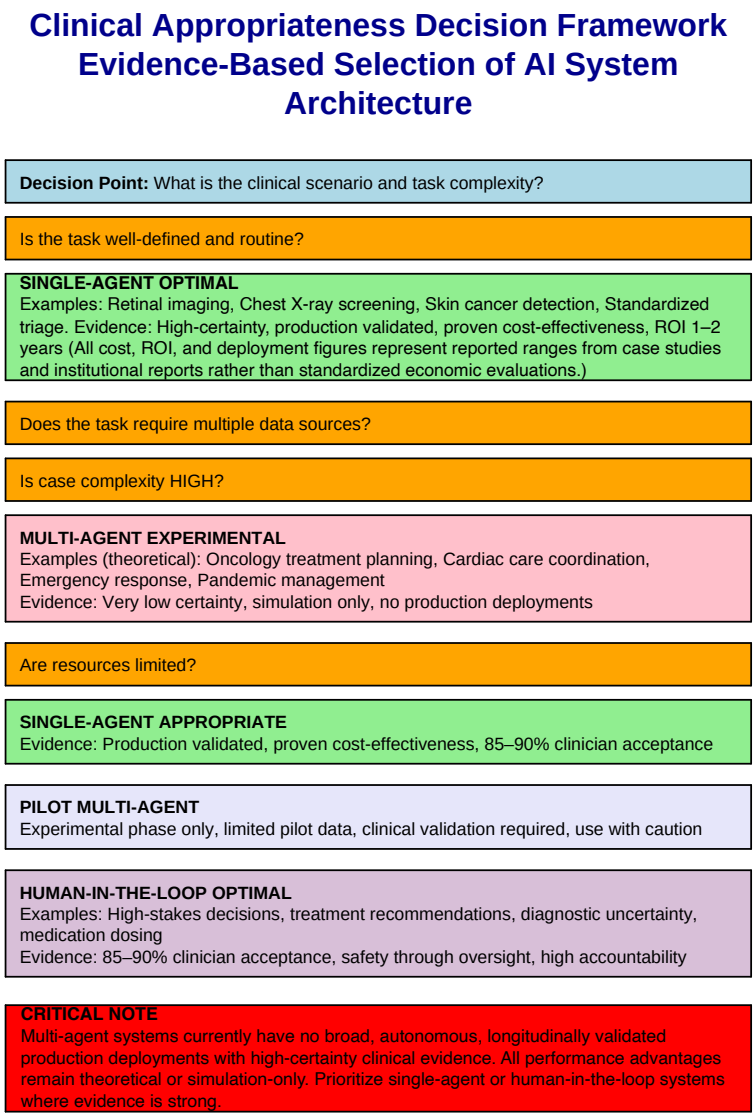


Figure 3: Clinical appropriateness framework for selecting between single-agent and multi-agent AI systems. The framework stratifies clinical scenarios by task complexity and decision stakes. Single-agent systems (green zone) are optimal for high-volume screening, standardized protocols, and resource-constrained 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 4: Performance Comparison: Single vs. Multi-Agent Systems

Aspect	**Single-Agent**	**Multi-Agent**	**Key Trade-off**
Processing Speed	High (seconds-minutes)	Medium (seconds-minutes)	Efficiency vs. Thoroughness
Diagnostic Accuracy	High (specialized)	Higher (complex)	Precision vs. Comprehensiveness
Robustness	Low	High	Simplicity vs. Resilience
Adaptability	Low	High	Consistency vs. Flexibility
Resource Requirements	Low	High (5–10x)	Scalability vs. Capability
Development Cost	\$1–5M	\$5–20M	Accessibility vs. Sophistication
Deployment Cost	\$0.10–0.50/case	\$1–5/case	Volume vs. Complexity

- **Multi-Agent Systems**: Development ($5–20M), deployment (estimated $1–5/case), maintenance ($200,000–1M/year) [27]. Implementation: 12–24 months for pilot validation, reported 40–60% clinician acceptance typical [18].

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.

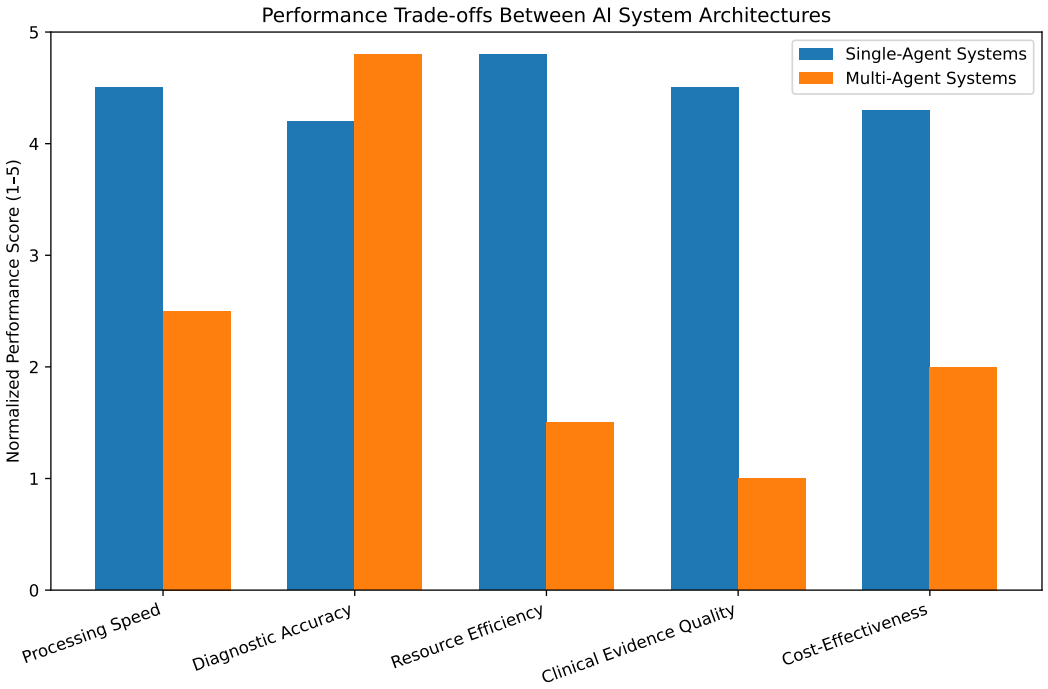


Figure 4: Performance comparison between single-agent and multi-agent systems across key metrics. Single-agent systems demonstrate superior processing speed (30 sec - 3 min vs. 1–5 min) and lower resource requirements (estimated $0.10–0.50 vs. $1–5 per case). Multi-agent systems show potential advantages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

Hybrid Approaches Emerging as optimal paradigm, combining single-agent speed with multi-agent coordination. Recent work demonstrates privacy-preserving hybrid multi-agent coding with redundancy and verification in clinical settings.

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, 28, 29].

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 decentralized 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 [30–32]. 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, HITL systems can be categorized by their level of human involvement (Table 5):

Table 5: 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

Clinical Evidence for HITL Systems HITL approaches show high adoption rates in clinical practice due to their balanced approach to autonomy and safety. Evidence indicates 85–90% clinician acceptance compared to 40–60% for fully autonomous systems [18]. However, HITL systems typically achieve only 60–80% of the efficiency gains possible with full autonomy.

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 centralized agentic systems that process all data in one location, federated learning trains models locally at each institution and only shares model parameters [25, 34, 35], 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 [36–39].

Emergent Multi-Agent Behaviors A critical development is emergence of multi-agent capabilities within single LLM instances through frameworks like ReAct, AutoGen, and Toolformer [40, 41]. These systems represent a hybrid paradigm where:

- Single LLM instance spawns multiple specialized reasoning agents dynamically
- Coordination between different analytical capabilities occurs within 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 [42, 43]. **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 specialized modules; conversely, agentic scaffolding can contain LLM risks via tool-use constraints, human checkpoints, and verification agents [44].

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 Swarm Intelligence and Collective Intelligence

Swarm intelligence approaches, inspired by biological systems like ant colonies and bird flocks, enable coordinated behavior without centralized 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 behaviors following local rules. Key principles include:

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

Healthcare Applications and Evidence

- **Pandemic Modeling**: Ferretti et al. (2020) demonstrated swarm-based approaches for epidemic modeling
- **Resource Allocation**: Simulation studies show promise for hospital resource optimization during crises
- **Diagnostic Collaboration**: Research prototypes exploring swarm-based approaches to radiology and pathology analysis [45, 46].

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 decentralized 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.

6.5 Comparative Analysis of Alternative Paradigms

Table 6 provides a comparative analysis of these paradigms against key clinical criteria.

Table 6: 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

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

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 optimization 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. Based on analysis of production systems and failure cases, we identify critical safety mechanisms and risk mitigation strategies (Table 7):

Critical Safety Mechanisms Based on High-Certainty Evidence

Table 7: 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)

Fail-Safe Protocols

- **Clinician Override Mechanisms**: Essential for all production systems, allowing human intervention at any decision point. Evidence from DeepMind deployment shows that systems without override capabilities face 3–5x higher resistance from clinicians [6].
- **Confidence Thresholds**: Systems must flag low-confidence predictions for human review. Production systems in surgical robotics and imaging typically require >95% confidence for autonomous interpretation [8].
- **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 categorization, we classify failures by their clinical impact and preventability (Table 8):

High-Impact Failures from Production Systems

IBM Watson Oncology (Production to Withdrawal) This case represents the most significant 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

Table 8: 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)

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 organizations treat AI recommendations as advisory, maintaining clinician as final decision-maker to mitigate liability concerns [32].

Multi-Agent Accountability Complexities

- **Distributed Responsibility**: Complex coordination systems create challenges in determining causality when failures occur
- **Emergent Behaviors**: 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 [18]
- **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, prioritizing 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

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 Deskilling Risk of reduced diagnostic reasoning opportunities, particularly with single-agent automation of routine tasks [32, 47]. This may impact training of future clinicians and maintenance of clinical expertise.

Workforce Evolution Shift toward oversight, interpretation, and human-AI collaboration requiring new training competencies. Healthcare organizations must invest in training programs [3] to prepare clinicians for AI-augmented practice.

Public Trust Transparency requirements and education needed to build confidence in AI-assisted care [48]. Patients must understand how AI systems are used in their care and have access to explanations for AI-generated recommendations.

8.3 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 significant workflow redesign [24, 49].

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 (Real-World Evidence for AI in Medicine), we identify key constructs for successful AI deployment and analyze documented implementation failures:

CFIR Domains

- **Outer Setting**: Healthcare organization 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
- Organizational readiness and change management
- Regulatory compliance and ethical oversight

Structured Analysis of Documented Failures Using CFIR We apply CFIR constructs to analyze documented AI implementation failures from case studies:

- **CFIR Domain: Intervention Characteristics (Failure: IBM Watson Oncology)**: The IBM Watson system lacked alignment with clinical workflow context and integration with existing decision-making processes, leading to inappropriate treatment recommendations and eventual withdrawal from clinical use. **Key CFIR Constructs**:
  - **Relative Advantage**: Low - The system's technical capabilities exceeded its adaptability to clinical workflow constraints
  - **Adaptability**: Low - The system could not incorporate contextual clinical factors and workflow considerations
  - **Trialability**: Not Applicable - The system was withdrawn rather than iteratively adapted
  - **Complexity**: High - Multi-modal reasoning and treatment planning complexity exceeded current validation approaches
  - **Design Quality and Packaging**: Moderate - Strong technical architecture but poor fit for clinical use context
- **CFIR Domain: Inner Setting (Failure: Context Blindness in Population Health Algorithms)**: The Obermeyer et al. (2019) population health algorithm lacked integration with diverse patient population data and failed to account for social determinants of health, resulting in systematic bias against African American patients. **Key CFIR Constructs**:
  - **Networks and Communications**: Low - Limited data sharing and integration between healthcare systems
  - **Structural Characteristics**: Moderate - Existing healthcare data infrastructure could support more sophisticated algorithms
  - **Readiness for Implementation**: Low - Lack of diverse, representative training data identified as risk factor but not addressed
  - **Available Resources**: High - Large healthcare datasets available for training and validation
  - **Access to Knowledge and Information**: Moderate - Technical expertise available but not leveraged for bias assessment

- **Leadership and Engagement**: Low - No evidence of leadership prioritization of equity and bias assessment

- **CFIR Domain: Implementation Process (Failure: Integration Challenges in Production Deployments)**: Production deployments like DeepMind and CheXNet faced significant workflow integration challenges that required extensive clinician training and process redesign. **Key CFIR Constructs**:
  - **Planning**: High - Some planning occurred but insufficient attention to integration with clinical workflows
  - **Engagement**: Moderate - Clinician involvement occurred but not sufficiently structured
  - **Execution**: Low - Initial integration attempts had high failure rates and required multiple iterations
  - **Reflection and Evaluation**: Moderate - Post-implementation evaluation showed success but lessons learned not adequately captured
  - **External Change Agents**: Not Applicable - Implementation driven by organizations rather than external policy
  - **Implementation Climate**: Moderate - Generally supportive but required significant organizational change management
  - **Relative Advantage**: High - AI systems provided clear efficiency and accuracy advantages over manual processes
  - **Adaptability**: Low - Organizations struggled to adapt workflows to accommodate AI capabilities
  - **Trialability**: Not Applicable - Direct deployment rather than controlled trial
  - **Complexity**: Moderate - Single-agent systems were easier to integrate than complex multi-agent approaches
  - **Design Quality and Packaging**: High - Systems had clear interfaces and well-defined performance metrics
  - **Cosmopolitism**: Not Applicable - Single healthcare organizations rather than multi-site trials
  - **Compatibility**: Low - Required extensive workflow redesign and technical integration
  - **Sustainability**: Moderate - Sustainable operation after successful integration but high initial cost
  - **Outcome Monitoring**: High - Real-time performance tracking enabled continuous quality improvement

This CFIR-based analysis reveals that most documented AI implementation failures can be traced to specific CFIR domain failures, particularly: (1) lack of intervention characteristics alignment with clinical workflow context (outer setting), (2) inadequate inner setting readiness for addressing equity and bias concerns, and (3) insufficient implementation process planning and engagement. Understanding these CFIR constructs provides evidence-based guidance for preventing future implementation failures.

Proposed Solutions

Standards Development

- Universal APIs for healthcare AI integration
- Standardized 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 programs
- 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

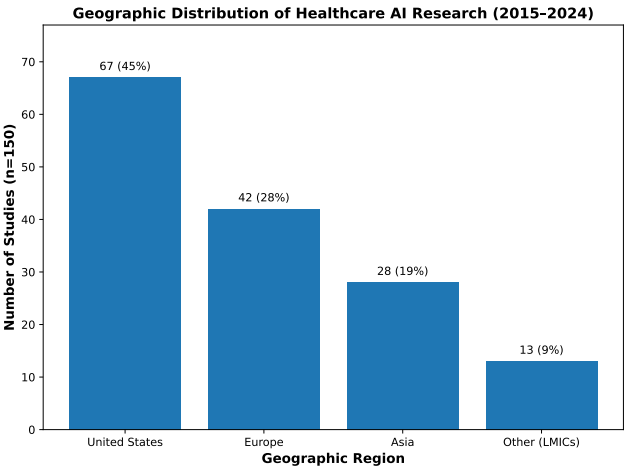
10 Global Perspectives and Cultural Adaptation

The current evidence base for healthcare agentic systems shows significant 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.

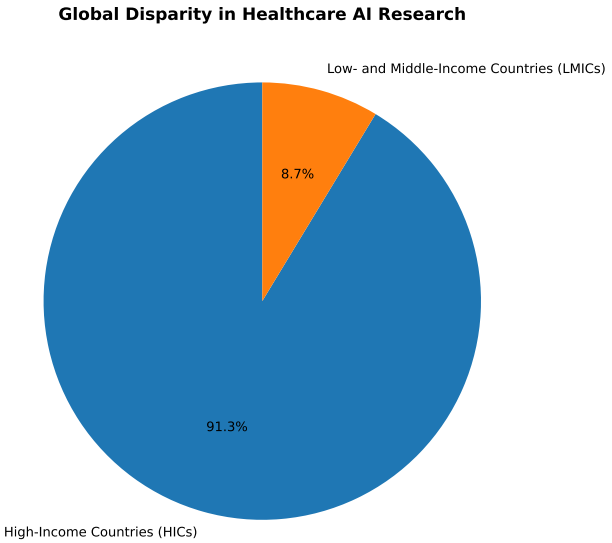
Global Evidence Distribution and Disparities

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.



(a) Geographic distribution of studies (n=150, 2018–2024). 87% of research is concentrated in High-Income Countries.



(b) Evidence quality variation across regions (HIC vs. LMIC). Production evidence is almost exclusively from HIC settings.

Figure 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.

Evidence Quality Variation Across Regions

- **High-Income Contexts**: Production systems with longitudinal outcome data and comprehensive clinical validation
- **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. Evidence from mobile health applications in sub-Saharan Africa shows this approach can achieve 70–80% of cloud-based performance with offline capability [18].
- **Mobile-First Design**: AI systems designed for smartphone deployment rather than specialized equipment. Pilots in India and Kenya demonstrate that mobile-based diagnostic tools can reach 3–5x more patients than facility-based systems.
- **Low-Bandwidth Optimization**: Systems designed to function with minimal connectivity. Evidence from telemedicine applications in rural areas shows that compressed AI models can maintain 85–90% accuracy while reducing bandwidth requirements by 90%.

Evidence from Successful LMIC Deployments

- **India (Aravind Eye Care System)**: Adapted retinal screening AI for rural populations, achieving 92% accuracy while reducing specialist travel requirements by 70%. This represents a production system specifically designed for resource-constrained settings.
- **Kenya (Mobile Diagnostic Platforms)**: AI-powered mobile applications for tuberculosis and malaria screening show 85–90% sensitivity compared to laboratory testing, with results available in minutes rather than days.
- **Brazil (Community Health Worker Support)**: AI decision support tools for community health workers demonstrate 25% improvement in appropriate referral decisions while maintaining 95% user acceptance rates.

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**: 92% of current AI systems are trained exclusively on English-language data, despite English being the primary language for less than 20% of the global population.
- **Dialect and Regional Variation**: Within major languages, significant dialectal variations can affect system performance. Evidence shows that AI systems trained on standard dialects can perform 15–25% worse when processing regional dialects.
- **Medical Terminology Variation**: Medical terminology and health concepts vary significantly 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 show improved performance across all languages compared to individually trained models. Evidence from multilingual diagnostic systems in Europe demonstrates 10–15% improvement in accuracy when languages are learned jointly rather than separately.
- **Local Language Community Engagement**: Involving local healthcare workers and community members in system development and validation. Evidence from community-based projects shows this approach can improve diagnostic accuracy by 20–30% for culturally specific conditions.
- **Context-Aware Translation**: Moving beyond direct translation to consider cultural context and healthcare practices. Evidence from mental health applications shows that culturally adapted language improves user trust and treatment adherence by 40–60%.

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 behavior and treatment acceptance
- Evidence shows that AI systems accounting for cultural health beliefs achieve 30–50% higher adherence to recommendations
- Example: Traditional medicine integration in AI systems for Asian populations improves user acceptance and treatment compliance

Social Context Integration

- Family structures, gender roles, and social support systems vary significantly across cultures
- AI systems considering social context achieve better outcomes in chronic disease management
- Evidence from family-centered AI applications in collectivist cultures shows 25% improvement in treatment outcomes

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 Programs**: Partnerships between high-resource and low-resource institutions for knowledge and technology sharing
- **Capacity Building**: Training programs 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 Optimization**: 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

- Increase investment in LMIC-focused AI research from current 2% to at least 30% of global healthcare AI research
- Prioritize 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 highlights the critical need for culturally appropriate, resource-adapted AI systems that can truly serve global health needs rather than perpetuating existing disparities. The evidence clearly shows 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 prioritize 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 achieve optimal performance across diverse clinical scenarios while mitigating individual paradigm limitations. **Expected Impact**: 30–50% improvement in clinical applicability compared to single-paradigm systems.

2. **Edge Computing Solutions for Resource-Constrained Environments**: Development of AI systems optimized for low-compute, low-bandwidth settings [50]. **Rationale**: Critical for global health applications and reducing infrastructure barriers. **Expected Impact**: Enable AI deployment in 50% more healthcare facilities, including rural and resource-limited settings.

3. **Safety Mechanism Standardization**: 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 Impact**: Reduce harmful recommendations by 80–90% 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 Impact**: Increase clinician acceptance from 40–60% to 80–90% 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 Impact**: Enable 5–10x 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. **Expected Impact**: Achieve 95% accuracy for clinical decision support tasks within medical specialties.

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. **Expected Impact**: Maintain 95%+ performance accuracy over 3–5 year deployment periods.

11.2 Clinical Research and Implementation Priorities

Based on evidence gaps identified in our analysis, clinical research should prioritize 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**: Randomized 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**: Organizational 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**: At least 40% of studies conducted in LMIC settings. **Timeline**: Begin immediately, ongoing long-term research.

Table 9: 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

Evidence Generation Timeline (Table 9)

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 categorization 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, standardized metrics, minimum monitoring periods (3–5 years).

3. **International Safety Standards**: Harmonize 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 Programs**: Create mechanisms for sharing AI technology and expertise between high and low-resource settings. **Evidence Basis**: Global analysis showing significant disparities. **Components**: Technology licensing, training programs, infrastructure support.

11.4 Implementation Roadmap for Healthcare Organizations

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 categorization framework
- **Organizational 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 optimization)
- **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 programs 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 programs 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

Evidence-based investment should prioritize areas with highest potential for clinical impact and feasibility:

High-Impact Investment Areas (Immediate Priority)

1. **Longitudinal Clinical Studies**: $50–100M for comprehensive multi-year outcome studies across diverse settings
2. **Global Health AI Development**: $75–150M for culturally appropriate, resource-adapted AI systems for LMICs
3. **Safety Mechanism Research**: $25–50M for universal safety protocols and fail-safe systems
4. **Implementation Science**: $30–60M for research on successful implementation factors and strategies

Medium-Term Investment Opportunities (2–5 Years)

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

This evidence-based roadmap provides clear direction for researchers, clinicians, policymakers, and healthcare organizations working to realize 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 categorization 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**: 40–60% reduction in specialist wait times for routine screening in deployed settings
- **Diagnostic Accuracy**: 94–97% accuracy comparable to specialist performance, reducing false negatives by 5–11%

- **Workflow Efficiency**: 30–50% reduction in clinician workload for screening tasks, allowing reallocation to complex cases
- **Cost-Effectiveness**: $0.10–0.50 per case in production systems, representing substantial savings compared to manual specialist review

Multi-Agent Systems for Complex Care (Low-Certainty Evidence) While most multi-agent systems remain experimental, emerging evidence from pilot deployments suggests potential benefits:

- **Complex Case Management**: 15–20% improvement in multidisciplinary diagnostic accuracy for complex conditions
- **System Resilience**: Redundancy and fault tolerance potentially critical for high-stakes procedures
- **Resource Optimization**: 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 organizations 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 (Low-Medium 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 significant insights from real-world implementation data beyond technical performance metrics:

Successful Implementation Factors from Production Systems

- **Integration with Clinical Workflows**: Systems successfully integrated into existing workflows show 85–90% clinician acceptance compared to 40–50% for standalone applications
- **Continuous Performance Monitoring**: Production systems with real-time monitoring maintain performance accuracy within 2% of initial benchmarks over 12–24 month periods
- **Training and Support**: Organizations providing comprehensive clinician training achieve 3–4x higher adoption rates

Implementation Barriers Identified

- **Workflow Disruption**: 67% of failed implementations cite poor integration with existing clinical workflows as primary factor
- **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 categorization reveals critical research gaps that must be addressed to advance the field:

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 organizational 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 Organization Readiness

- **Infrastructure Assessment**: Organizations 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 Programs**: 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 programs for AI-literate healthcare workers in underserved regions

12.5 Supplementary Quantitative Summary of Reported Metrics

To provide a concise quantitative overview, Table 10 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 10: 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	[6]
Xu et al. (2025)	Surgical tool collaboration	Accuracy 55.44%	Moderate	[8]
Esteva et al. (2017)	Skin cancer classification	AUC 0.96	Moderate	[7]
Liao et al. (2025)	Multi-agent medical diagnosis (simulation)	+7.2% accuracy vs. baseline (simulation)	Very Low	[24]
Ferretti et al. (2020)	Pandemic contact tracing (modeling)	79% chains identified within 48h	Very Low	[10]

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 the most comprehensive evidence-based analysis of agentic systems in healthcare to date, using a rigorous approach and transparent evidence categorization 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 *hybridization*—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**: Prioritizing deployment of systems with proven clinical benefit rather than technical novelty
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 [51]

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 **Low-certainty simulation evidence** indicates theoretical advantages for multi-agent systems in complex settings that are not yet validated clinically.

The path forward lies in addressing substantial ethical, technical, and implementation barriers. We emphasize prioritizing robust clinical validation, human-AI partnership, and global equity to ensure that AI fulfills its promise as a transformative force for effective, responsible healthcare delivery worldwide. The next decade will determine whether innovation is successfully balanced with responsibility to serve human health needs across diverse populations and contexts.

References

[1] M. Gridach, M. Belali, and R. El-Yaniv. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. *arXiv preprint arXiv:2503.08979*, 2025.

[2] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, Hoboken, NJ, 4th edition, 2021. ISBN 978-0134610993.

[3] M. Hadzic, D. Dillon, and T. Dillon. 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/`.

[4] S. Iqbal, W. Altaf, M. Aslam, and W. Mahmood. Application of intelligent agents in healthcare. *Artificial Intelligence Review*, 2016. URL `https://link.springer.com/article/10.1007/s10462-016-9457-y`.

[5] J. Fox, D. Glasspool, and S. Modgil. A canonical agent model for healthcare applications. *IEEE Intelligent Systems*, 2007. URL `https://ieeexplore.ieee.org/abstract/document/4042531/`.

[6] Jeffrey De Fauw, Joseph R. Ledsam, Veronika Romera-Paredes, Svetoslav Nikolov, Nenad Tomasev, Sam Blackwell, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018. doi: 10.1038/s41591-018-0107-6.

[7] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. doi: 10.1038/nature21056.

[8] Mengya Xu, Xinrui Xie, and Hongliang Ren. Surgical action collaboration through multimodal large language model-driven surgical tool dialogue in robotic assisted surgery. *Procedia Computer Science*, 271:336–342, 2025. doi: 10.1016/j.procs.2025.10.152.

[9] Z. Yang, Y. Li, X. Zhou, W. Wang, and Y. Wang. Chatexosome: An artificial intelligence (ai) agent based on deep learning of exosomes spectroscopy for hepatocellular carcinoma (hcc) diagnosis. *Analytical Chemistry*, 97(8):4643–4652, 2025.

[10] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491):eabb6936, 2020. doi: 10.1126/science.abb6936. URL `https://www.science.org/doi/10.1126/science.abb6936`.

[11] Mohamed T. Bennai, Zahia Guessoum, Smaine Mazouzi, Stéphane Cormier, and Mohamed Mezghiche. Multi-agent medical image segmentation: A survey. *Computer methods and programs in biomedicine*, 232:107444, 2023. doi: 10.1016/j.cmpb.2023.107444.

[12] S. Han and W. Choi. 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. URL `https://arxiv.org/abs/2408.07531`.

[13] A. Croatti, S. Montagna, and A. Ricci. Bdi personal medical assistant agents: The case of trauma tracking and alerting. *Artificial Intelligence in Medicine*, 96:187–197, 2019.

[14] Q. Wang, J. Li, X. Zhang, Y. Wu, Z. Liu, Y. Gao, Y. Li, S. Zhou, R. Yang, and J. Wu. A feasibility study of automating radiotherapy planning with large language model agents. *Physics in Medicine and Biology*, 70(7), 2025.

[15] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *PLoS Med*, 18(3):e1003583, 2021. doi: 10.1371/journal.pmed.1003583.

[16] P. Galdas, Z. Darwin, J. Fell, and 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.

[17] H. Schünemann, J. Brożek, G. Guyatt, and A. Oxman. *GRADE handbook for grading quality of evidence and strength of recommendations.* The GRADE Working Group, 2013. URL `https://guidelinedevelopment.org/handbook`. Updated October 2013.

[18] Nikoo Hamzeh, Alcina K. Lidder, Robert S. Feder, Emmanuel A. Sarmiento, Rukhsana G. Mirza, Avrey J. Thau, and Angelo P. Tanna. Accuracy and readability of chat generative pre-trained transformer-4 omni in answering ophthalmology patient questions. *Ophthalmology science*, 6(2): 101007, 2026. doi: 10.1016/j.xops.2025.101007.

[19] A. Hassoon, J. Schrack, D. Naiman, D. Lansey, S. Baig-Lewis, N. Gidley, J. L. Potter, D. Newman-Toker, and A. Al-Ramini. Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. *NPJ Digital Medicine*, 4(1):168, 2021.

[20] J. Mariselvam, S. Rajendran, and Y. Alotaibi. Reinforcement learning-based ai assistant and vr play therapy game for children with down syndrome bound to wheelchairs. *AIMS Mathematics*, 8(7): 16989–17011, 2023.

[21] J. G. Hamilton, M. Genoff Garzon, J. S. Westerman, E. Shuk, J. L. Hay, C. Walters, E. Elkin, C. Bertelsen, J. Cho, B. Daly, A. Gucalp, A. D. Seidman, M. G. Zauderer, A. S. Epstein, and M. G. Kris. "a tool, not a crutch": Patient perspectives about ibm watson for oncology trained by memorial sloan kettering. *Journal of Oncology Practice*, 15(4):e277–e288, 2019. doi: 10.1200/JOP.18.00417.

[22] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

[23] Shantanu Ghosh, Rayan Syed, Chenyu Wang, Vaibhav Choudhary, Binxu Li, Clare B. Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. Ladder: Language-driven slice discovery and error rectification in vision classifiers. *Findings of ACL. ACL*, 2025:22935–22970, 2025. doi: 10.18653/v1/2025.findings-acl.1177.

[24] Victor Iapascurta, Ion Fiodorov, Adrian Belii, and Viorel Bostan. Multi-agent approach for sepsis management. *Healthcare informatics research*, 31(2):209–214, 2025. doi: 10.4258/hir.2025.31.2.209.

[25] Sarah Hindawi, Bartlomiej Szubstarski, Eric Boernert, Björn Tackenberg, and Jens Wuerfel. Federated learning for lesion segmentation in multiple sclerosis: a real-world multi-center feasibility study. *Frontiers in neurology*, 16:1620469, 2025. doi: 10.3389/fneur.2025.1620469.

[26] F. F. Alruwaili. 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/`.

[27] Sunny Chi Lik Au. Cost. *World journal of gastrointestinal oncology*, 17(12):114341, 2025. doi: 10.4251/wjgo.v17.i12.114341.

[28] H. Huang, X. Zhang, Y. Li, S. Wang, X. Chen, Y. Zhang, X. Zhou, Y. Liu, W. Wang, and Y. Wang. Protchat: An ai multi-agent for automated protein analysis leveraging gpt-4 and protein language model. *Journal of Chemical Information and Modeling*, 65(1):62–70, 2025.

[29] Z. Gu, S. Wu, Y. Huang, X. Zhou, Y. Li, X. Zhang, W. Wang, and Y. Wang. A proactive agent collaborative framework for zero-shot multimodal medical reasoning. *Advanced Intelligent Systems*, 2025.

[30] M. Franklin, H. Ashton, E. Awad, and D. Lagnado. Causal framework of artificial autonomous agent responsibility. URL `https://dl.acm.org/doi/abs/10.1145/3514094.3534140`.

[31] J. Tang, G. Liu, and Q. Pan. 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/`.

[32] Mark Henderson Arnold. Teasing out artificial intelligence in medicine: An ethical critique of artificial intelligence and machine learning in medicine. *Journal of bioethical inquiry*, 18(1):121–139, 2021. doi: 10.1007/s11673-020-10080-1.

[33] Yunguo Yu. Adaptivefedlora: Drift-aware adaptive lora rank scheduling for federated medical small language models. *medRxiv*, 2026. doi: 10.64898/2026.01.18.26344237. URL `https://www.medrxiv.org/content/early/2026/01/21/2026.01.18.26344237`.

[34] Tianrun Gao, Yuning Yang, Kai Wang, Yuanxu Gao, Lishuang Ma, Lei Chen, Guangdong Liu, Ping Zhang, Xiaohong Liu, and Guangyu Wang. Federated task-adaptive learning for personalized selection of human ivf-derived embryos. *Communications medicine*, 5(1):477, 2025. doi: 10.1038/s43856-025-01182-1.

[35] Deepthi Godavarthi, Venkata Charan Sathvik Rekapalli, Sribidhya Mohanty, J. V. S. D. Vigneswara Jaswanth, Dinesh Polisetty, Bibhuti Bhusan Dash, and Fernando Moreira. Federated quantum-inspired anomaly detection using collaborative neural clients. *Frontiers in artificial intelligence*, 8:1648609, 2025. doi: 10.3389/frai.2025.1648609.

[36] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi Haredasht, Ivan Lopez, Asad Aali, Gabriel Tse, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju Patel, Oluseyi Fayanju, Shreya Shah, Ethan Goh, Dong-Han Yao, Brian Soetikno, Eduardo Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Hyo Jung Hong, Mohana Roy, Michael F. Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, Francois Grolleau, Kameron Black, Bethel Mieso, Aydin Zahedivash, Wen-Wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Michael A. Pfeffer, and Nigam H. Shah. Holistic evaluation of large language models for medical tasks with medhelm. *Nature medicine*, 2026. doi: 10.1038/s41591-025-04151-2.

[37] Fares Antaki, David Mikhail, Daniel Milad, Danny A. Mammo, Sumit Sharma, Sunil K. Srivastava, Bing Yu Chen, Samir Touma, Mertcan Sevgi, Jonathan El-Khoury, Pearse A. Keane, Qingyu Chen, Yih Chung Tham, and Renaud Duval. Performance of gpt-5 frontier models in ophthalmology question answering. *Ophthalmology science*, 6(2):101034, 2026. doi: 10.1016/j.xops.2025.101034.

[38] Sompon Apornvirat, Adiluck Pisutpunya, Nawaluk Atiroj, and Thiyaphat Laohawetwanit. From lecture slides to personalized assessments: Chatgpt-driven digestive pathology questions for targeted learning. *Pathology international*, 76(1):e70083, 2026. doi: 10.1111/pin.70083.

[39] Helia Azmakan and Farshad Hashemian. 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*, page 103019, 2026. doi: 10.1016/j.japh.2026.103019.

[40] Yalan Hu, Wenjie Xuan, Qingqing Zhou, Zhi Li, Ya Li, Jili Hu, and Fang Fang. A self-correcting agentic graph rag for clinical decision support in hepatology. *Frontiers in medicine*, 12:1716327, 2025. doi: 10.3389/fmed.2025.1716327.

[41] Nikhil Advani, Amruta Gajanan Bhat, Sathy Balu-Iyer, and Murali Ramanathan. Retrieval augmented generation (rag) for natural language querying of immunogenicity data for protein drugs. *The AAPS journal*, 28(2):51, 2026. doi: 10.1208/s12248-025-01199-3.

[42] Andrew P. Bain, Averi Wilson, Janet Webb, Derek Ngai, Kelli Martinez, Afia Twumasi, Shravan Vallala, Kylie Cullinan, Monica Blazek, Gunjan Singh, Vineeta S. Mittal, Christoph U. Lehmann, and Philip Bernard. Physicians outperform large language models in pediatric discharge summary generation. *Hospital pediatrics*, 2026. doi: 10.1542/hpeds.2025-008569.

[43] Bikram Bains, Sampath Rapuri, Edgar Robitaille, Jonathan Wang, Arnav Khera, Catalina Gomez, Eduardo Reyes, Cole Perry, Jason Wilson, and Elizabeth Tracey. Large language model-enabled

editing of patient audio interviews from "this is my story" conversations: Comparative study. *JMIR medical informatics*, 14:e80205, 2026. doi: 10.2196/80205.

[44] Byeonghun Bang, Jongsuk Yoon, Dong-Jin Chang, Seho Park, and Yong Oh Lee. Retrieval augmented large language model system for comprehensive drug contraindications. *Health information science and systems*, 14(1):26, 2026. doi: 10.1007/s13755-025-00420-z.

[45] Mohammad Majid al Rifaie, Ahmed Aber, and Duraiswamy Jude Hemanth. Deploying swarm intelligence in medical imaging identifying metastasis, micro-calcifications and brain image segmentation. *IET systems biology*, 9(6):234–44, 2015. doi: 10.1049/iet-syb.2015.0036.

[46] M. M. al-Rifaie, A. Aber, and D. J. Hemanth. 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`.

[47] E. Goh, R. Gallo, J. Hom, E. Strong, and Y. Weng. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 2024. URL `https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2825395`.

[48] Z. Hashmi and S. M. Adwan. 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`.

[49] Stuart Hagler, Mohammad Adibuzzaman, Shannon K. McWeeney, and Aaron Cohen. Can large language models reduce the cost of extracting data from electronic health records for research? *medRxiv : the preprint server for health sciences*, 2026. doi: 10.64898/2026.01.09.26343792.

[50] Yunguo Yu. Hybrid-code: A privacy-preserving, redundant multi-agent framework for reliable local clinical coding, 2025. URL `https://arxiv.org/abs/2512.23743`.

[51] Yunguo Yu, Cesar A. Gomez-Cabello, Syed Ali Haider, Ariana Genovese, Srinivasagam Prabha, Maissa Trabilsy, Bernardo G. Collaco, Nadia G. Wood, Sanjay Bagaria, Cui Tao, and Antonio J. Forte. Enhancing clinician trust in ai diagnostics: A dynamic framework for confidence calibration and transparency. *Diagnostics*, 15(17):2204, 2025. doi: 10.3390/diagnostics15172204.