REVIEW Open access

# Agentic Artificial Intelligence in Healthcare Systems: A Systematic Review of Autonomous Discharge Coordination, Task Routing, Human Oversight, Safety Guardrails, and Workflow Accountability

Natalie Brown[1*], Emma Taylor[1], James Wilson[2], Oliver Scott[1]

## Abstract

Agentic artificial intelligence systems—those capable of planning, executing, adapting, and learning across operational tasks—are beginning to influence healthcare administrative workflows. Their emergence raises important questions about autonomy, safety, oversight, and accountability in hospital systems. This systematic review examined the development and deployment of agentic AI in healthcare operations from 2017 to 2026. The review focused on autonomous discharge coordination, task routing, human oversight, safety guardrails, and workflow accountability. A PRISMA 2020–aligned search was conducted across PubMed, Scopus, IEEE Xplore, and Web of Science. Studies were screened by two reviewers, and eligible records were synthesised narratively according to task type, agent capability, oversight model, safety mechanism, and deployment maturity. The evidence base was nascent, heterogeneous, and dominated by predictive, prototype, implementation, and early deployment studies. Most systems supported discharge planning, workflow prediction, task prioritisation, or operational decision support rather than fully autonomous execution. Agentic AI has substantial potential to improve hospital operations, especially in discharge coordination and workflow routing. However, real-world deployment requires stronger safety engineering, prospective evaluation, explicit human oversight, and auditable accountability structures.

**Keywords** Healthcare operations, Agentic artificial intelligence, Autonomous agents, Discharge coordination, Task routing, Human oversight

*Correspondence:
Natalie Brown
natalie.brown@gmail.com
[1] Department of Digital Healthcare Engineering, Faculty of Medicine, University of Bristol, Bristol, United Kingdom
[2] Department of Clinical Informatics Systems, Faculty of Engineering, University of Nottingham, Nottingham, United Kingdom

## Introduction

Hospital operations have become increasingly complex as clinical teams manage rising patient volumes, constrained staffing, fragmented information systems, and time-sensitive administrative dependencies. These pressures have motivated interest in automation that extends beyond static rule-based alerts toward adaptive systems capable of anticipating workflow bottlenecks and coordinating operational tasks. Studies of discharge prediction, hospital throughput, and operational machine learning indicate that AI can support planning around patient flow, bed availability, and multidisciplinary coordination, although the evidence remains uneven across contexts [1-5]. Within this environment, agentic AI is being explored as a way to reduce cognitive burden while preserving clinical accountability.

Agentic artificial intelligence can be understood as systems that perceive operational state, form or update goals, plan sequences of action, execute tasks, and learn from feedback under varying degrees of autonomy. This differs from traditional clinical decision support, which typically recommends actions to clinicians, and from robotic process automation, which executes predefined scripts without flexible goal-directed reasoning. Multi-agent and autonomous-agent literature in healthcare has described architectures in which software agents coordinate across tasks, resources, and actors, while implementation studies of clinical AI emphasise that technical capability alone is insufficient without workflow integration and human-centred design [6-10]. In healthcare operations, agentic AI is therefore best treated not as a single technology but as a socio-technical configuration of autonomy, supervision, constraint, and accountability.

Delegating operational activities to AI agents may be particularly attractive in discharge coordination, task routing, transport scheduling, consult prioritisation, and escalation management. Discharge coordination is a multi-step process involving medication reconciliation, documentation, transport, bed management, post-acute placement, and communication across teams, making it suitable for systems that can sequence dependencies and track incomplete tasks [2, 11-14]. At the same time, delegation introduces risks when autonomous recommendations or actions are poorly calibrated, insufficiently monitored, or disconnected from clinical responsibility. Ethical and safety analyses of healthcare AI warn that autonomy must be constrained by oversight, fail-safe procedures, transparency, and careful monitoring for unintended harm [15-17].

This systematic review was designed to examine evidence on agentic AI for healthcare operations during 2017–2026, with emphasis on autonomous discharge coordination, task routing, human oversight, safety guardrails, and workflow accountability. The review followed principles by defining eligibility criteria, screening retrieved records, extracting standardised data elements, and synthesising findings according to prespecified operational domains. Because the literature spans predictive discharge models, multi-agent coordination systems, implementation studies, safety frameworks, and reporting guidelines, a narrative synthesis was selected rather than meta-analysis. The objective was to clarify what is known, what remains under-developed, and what conditions are required before agentic AI can be responsibly embedded in hospital workflows.

# Materials and Methods

## Search strategy

Searches were conducted in PubMed, Scopus, IEEE Xplore, and Web of Science for records published from January 1, 2017, through December 31, 2026, using combinations of terms related to “agentic artificial intelligence,” “autonomous agent,” “AI agent,” “multi-agent system,” “discharge coordination,” “task routing,” “workflow orchestration,” “human oversight,” “safety guardrails,” and “workflow accountability.” The search strategy intentionally combined emerging agentic-AI terminology with more established operational AI terminology because many relevant studies described predictive or coordination systems without using the term “agentic.” Targeted searches were also aligned with domains represented in discharge prediction, multi-agent workflow optimisation, machine-learning implementation, clinical AI safety, and reporting guidance. Reference chaining was used to identify additional eligible studies that addressed healthcare operations, autonomy, oversight, or accountability.

## Inclusion and exclusion criteria

Eligible records included peer-reviewed original research, implementation studies, systematic or scoping reviews, reporting guidance, or conceptual frameworks published in English between 2017 and 2026 that addressed autonomous or semi-autonomous AI systems in healthcare operations. Studies were included when they described systems relevant to administrative or operational tasks such as discharge planning, patient-flow forecasting, task routing, workflow prediction, resource allocation, or operational decision support. Records focused exclusively on diagnostic image classification, drug discovery, consumer wellness, or purely clinical risk prediction without workflow implications were excluded, unless they contributed directly to human oversight, safety, or accountability frameworks for autonomous

healthcare AI. This approach allowed inclusion of both operational AI systems and governance literature necessary to evaluate agentic deployment.

## Screening and selection

A total of 1,450 records were identified through database searches and reference chaining, of which 312 duplicates were removed before screening. Titles and abstracts of 1,138 records were reviewed independently by two reviewers, leading to 180 full-text articles assessed for eligibility and 51 studies included in the synthesis shown in **Figure 1.** Common reasons for exclusion at full text included no operational workflow component, no healthcare setting, absence of AI or autonomous decision logic, non-peer-reviewed publication type, or insufficient description of implementation context. Disagreements were resolved by discussion, with eligibility decisions guided by PRISMA 2020 principles and by prior reporting guidance for AI interventions and decision-support evaluation.

**Figure 1** shows the PRISMA 2020 study-selection process, including database identification, duplicate removal, title and abstract screening, full-text eligibility assessment, exclusion reasons, and final inclusion of 51 studies in the qualitative synthesis.

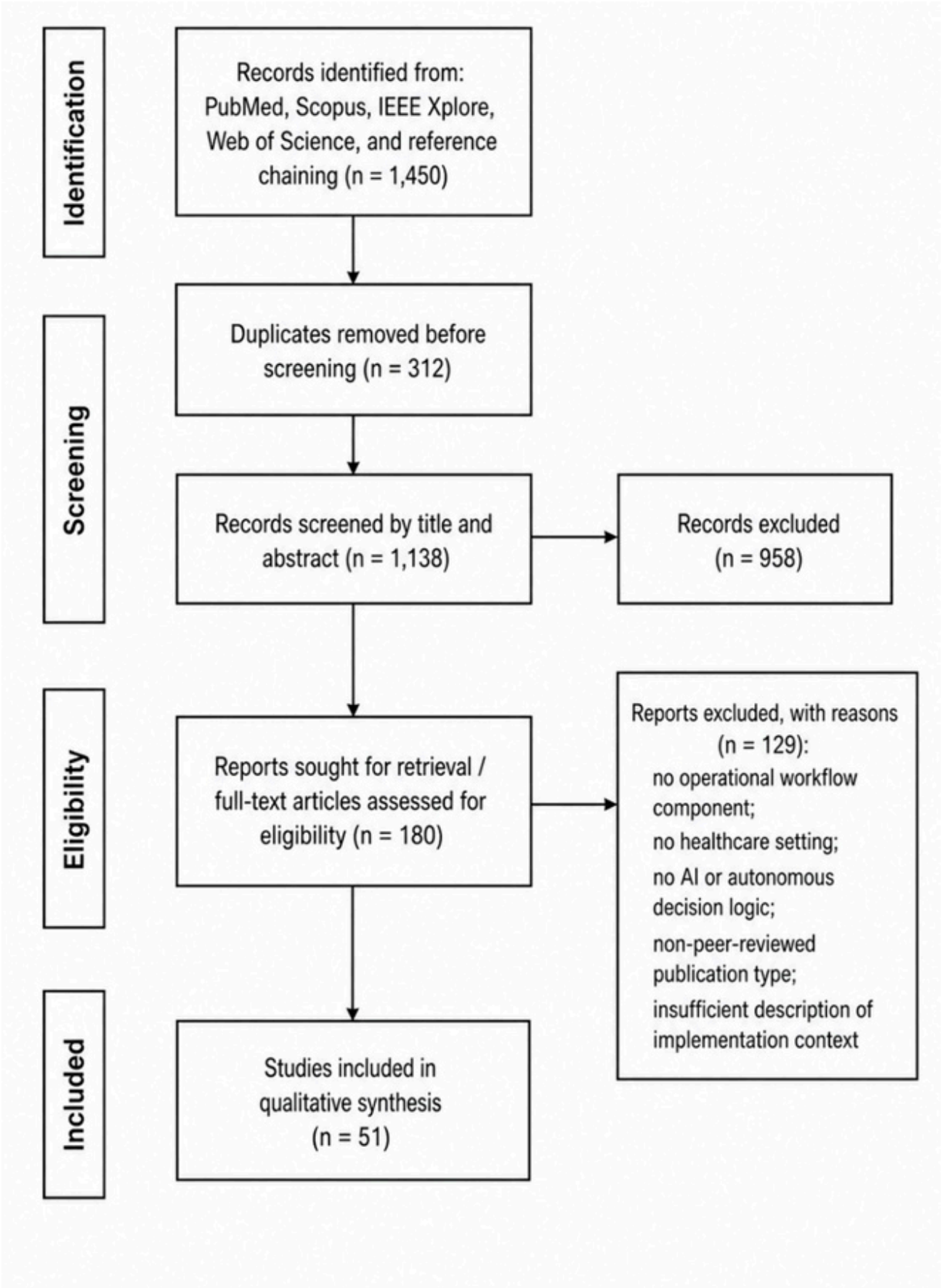

Identification
Records identified from: PubMed, Scopus, IEEE Xplore, Web of Science, and reference chaining (n = 1,450)
Screening
Duplicates removed before screening (n = 312)
Records screened by title and abstract (n = 1,138)
Records excluded (n = 958)
Eligibility
Reports sought for retrieval / full-text articles assessed for eligibility (n = 180)
Reports excluded, with reasons (n = 129):
no operational workflow component;
no healthcare setting;
no AI or autonomous decision logic;
non-peer-reviewed publication type;
insufficient description of implementation context
Included
Studies included in qualitative synthesis (n = 51)

**Figure 1**. PRISMA 2020 Flow Diagram for Study Selection

## Data extraction

For each included study, extracted fields included publication year, country or setting when available, healthcare environment, operational task, AI architecture, level of autonomy, human oversight model, safety guardrails, accountability mechanisms, and evaluation design. Discharge-focused studies were coded for whether the system supported prediction, coordination, sequencing, or execution of discharge-related actions, while task-routing studies were coded for triage, assignment, workload balancing, or re-routing capabilities. Safety and accountability fields were informed by responsible machine-learning frameworks, implementation studies, and early-stage clinical AI evaluation guidance. When studies reported limited operational detail, the missing information was recorded as not reported rather than inferred.

## Risk of bias assessment

Risk of bias was assessed using an adapted framework informed by PROBAST-style concerns for predictive modelling and qualitative appraisal of implementation evidence. Domains included participant or setting selection, data representativeness, outcome definition, confounding, model validation, integration into workflow, safety evaluation, and transparency of human oversight. Particular attention was given to whether systems were evaluated retrospectively, simulated prospectively, piloted in live settings, or embedded in routine care, because deployment maturity strongly affects claims about agentic reliability. Studies that lacked external validation, explicit monitoring, or assessment of unintended consequences were treated as having important evidence limitations even when their technical approach was well described.

## Synthesis methods

A narrative synthesis was used because the evidence base was heterogeneous in task type, study design, deployment maturity, and outcome reporting. Studies were grouped into domains covering autonomous discharge coordination, task routing and workflow orchestration, autonomy level, human oversight, safety guardrails, accountability, technology stack, implementation status, reported outcomes, and barriers. Maturity was assessed across a continuum from simulation and retrospective modelling to pilot testing and live operational deployment, reflecting distinctions made in implementation and AI evaluation literature. Quantitative pooling was not conducted because studies differed substantially in populations, workflows, comparators, and operational endpoints.

# Results and Discussion

## Study selection

The PRISMA process identified 51 eligible studies from 1,450 records, with **Figure 1** intended to depict identification, duplicate removal, abstract screening, full-text review, and inclusion. Excluded full-text studies most often lacked an operational workflow component, focused only on diagnostic prediction, or did not describe autonomy, oversight, or implementation details sufficient for review. The final evidence base included studies on discharge prediction, hospital operational forecasting, multi-agent optimisation, clinical AI implementation, responsible machine learning, and reporting frameworks [1, 5-7, 14, 18, 19]. Although the search targeted agentic AI explicitly, many included records used adjacent terminology such as autonomous agents, machine-learning workflow support, decision support, and implementation of AI into care delivery.

## Study characteristics

The included literature clustered around discharge planning, hospital-flow prediction, operating-room management, sepsis implementation, workflow integration, and AI governance rather than fully autonomous hospital agents. Several

studies were based in inpatient or hospital-wide settings, while others addressed specialty-specific workflows such as surgery, cardiovascular care, urology, or emergency operations [2, 4, 12, 20-22]. The publication years showed increasing attention to implementation, reporting standards, and governance after 2019, consistent with broader recognition that healthcare AI must be evaluated as part of clinical operations rather than as isolated algorithms [8, 9, 15, 23, 24]. Most studies came from technologically mature health systems, which may limit generalisability to lower-resource environments.

**Table 1** presents an operational taxonomy that distinguishes workflow domains, agent functions, autonomy boundaries, oversight arrangements, minimum safety guardrails, and deployment maturity across the included literature.

**Table 1.** Operational Taxonomy of Agentic AI Functions in Healthcare Systems

| Workflow domain | Typical operational goal | Representative agent functions | Predominant autonomy pattern in review | Human oversight model | Minimum safety guardrails | Deployment maturity observed |
|---|---|---|---|---|---|---|
| Discharge coordination | Accelerate safe discharge and reduce delays | Predict readiness; sequence outstanding tasks; track barriers; escalate unresolved issues | Mostly decision support or semi-autonomous coordination; full autonomous execution rare | Human-in-the-loop approval by care team or case manager | Hard stops for unresolved clinical tasks; override option; escalation log; audit trail | Moderate conceptual maturity; limited real-world autonomous deployment |
| Task routing and worklist prioritisation | Allocate consults, transport, documentation, and service tasks efficiently | Prioritise queues; route tasks; rebalance workload; reroute exceptions | Mostly recommendation or dashboard-based routing | Human-on-the-loop supervision by operational managers | Queue thresholds; fairness checks; conflict alerts; fallback to manual routing | Early implementation or pilot stage |
| Capacity and throughput orchestration | Optimise beds, transfers, and patient flow | Forecast bottlenecks; coordinate placement; suggest resource reallocation | Prediction-led support with constrained orchestration | Shared decision authority with bed managers | Capacity constraints; fail-safe defaults; continuous monitoring | Moderate evidence, mostly predictive |
| Resource coordination and scheduling | Align staffing, diagnostics, transport, and post-acute resources | Match availability; sequence dependent tasks; propose schedules | Limited agentic coordination; often simulation or optimisation | Human approval before execution | Rule-based boundaries; exception handling; accountability logs | Early-stage, largely simulation |
| Escalation and exception management | Detect stalled workflows or unsafe conditions | Trigger alerts; escalate unresolved barriers; | Low autonomy for action; mainly alerting | Mandatory human intervention at | Escalation thresholds; near-miss | Under-developed across literature |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | document actions | | escalation points | review; time-stamped logs | |

## Autonomous discharge coordination

Autonomous discharge coordination was most commonly represented by systems that predicted discharge readiness, anticipated discharge volume, or supported multidisciplinary planning rather than systems that independently executed all discharge tasks. Several studies reported models intended to assist discharge timing, next-day discharge prediction, patient-flow forecasting, or surgical discharge planning, suggesting a foundation for agentic coordination but not complete autonomy [1-3, 11-13, 20, 25]. These systems typically supported human teams by identifying likely discharges, prioritising attention, or enabling earlier planning around capacity and dependencies. Evidence for actual autonomous sequencing of discharge sub-tasks, such as automatically resolving documentation, medication, transport, and follow-up dependencies, remained limited.

## Task routing and workflow orchestration

Task routing and workflow orchestration were addressed through multi-agent optimisation, operational machine learning, and AI-supported management of resources and patient pathways. Agent-based modelling and hospital workflow optimisation studies described how distributed software agents could coordinate patient pathways and adapt to operational constraints, while broader operational AI literature emphasised workload prediction and resource allocation [5-7, 21]. In practice, most systems appeared to route information or recommendations to human teams rather than autonomously assigning tasks across departments. This pattern suggests that agentic task routing is technically plausible but remains constrained by integration complexity, liability concerns, and the need for explicit human supervision.

## Autonomy levels and decision authority

Across the evidence base, autonomy ranged from retrospective prediction models to live decision-support tools embedded in clinical workflows. Most discharge and operations studies produced predictions or alerts for human review, while fewer systems described autonomous initiation or completion of workflow actions [4, 11-13, 25]. Implementation and governance studies consistently framed decision authority as shared, with AI systems supporting prioritisation and coordination while clinicians or operational managers retained responsibility for final decisions [8, 9, 15, 16]. Fully autonomous action in high-stakes hospital operations was not supported by the reviewed evidence.

## Human oversight and supervisory models

Human oversight was universal in the reviewed literature, but oversight models varied substantially in timing, depth, and clarity. Some systems were effectively human-in-the-loop, requiring staff to interpret predictions before acting, while others approached human-on-the-loop supervision by embedding AI outputs into operational dashboards or routine clinical workflows [8, 10, 22]. Studies of explainability, clinician expectations, and human-centred deployment showed that users require contextual explanations, confidence calibration, and clear escalation pathways before trusting AI-supported workflow decisions [10, 26]. However, few studies formally compared oversight models or specified thresholds for mandatory human intervention.

## Safety guardrails implemented

Safety guardrails were commonly discussed as a requirement but inconsistently specified in operational studies. Responsible AI and ethical analyses emphasised constraints, monitoring, fail-safe defaults, bias assessment, and procedures for escalation when model outputs conflict with clinical judgement [15-17, 27]. Reporting and evaluation frameworks such as CONSORT-AI, SPIRIT-AI, and DECIDE-AI reinforced the need to describe intervention context, human interaction, safety monitoring, and failure handling [19, 23, 24]. In discharge and workflow studies, guardrails were more often implicit in human review than formalised as technical constraints on agent behaviour.

## Evaluation of safety and adverse events

Evaluation of safety outcomes was limited, especially for near-misses, erroneous workflow actions, or downstream patient harm caused by AI-supported operations. Many discharge studies focused on prediction or planning utility, while fewer examined whether AI recommendations could produce unsafe prioritisation, premature discharge pressure, or inequitable workflow decisions [1, 3, 11, 12, 28]. Broader AI evaluation literature cautioned that algorithmic performance does not guarantee safe clinical impact, particularly when deployed in complex socio-technical systems [15, 18, 27]. As a result, the evidence was insufficient to conclude that agentic operational AI can be safely deployed without rigorous prospective monitoring.

## Workflow accountability and audit trails

Workflow accountability was under-developed across the reviewed studies, even when systems generated logs, dashboards, or decision records. Implementation literature suggested that auditability is necessary for tracing AI recommendations, understanding human responses, and reviewing adverse events after deployment [8, 9, 22]. Ethical and governance analyses further indicated that responsibility cannot be delegated solely to software, particularly when AI systems influence clinical or administrative action [15-17]. However, few studies described how audit trails would be used in retrospective review, quality assurance, liability assessment, or professional accountability.

## Technology stack

The technology stack across included studies was heterogeneous, including machine-learning prediction models, statistical forecasting, agent-based simulation, workflow optimisation, rule-based decision support, and hybrid implementation architectures. Discharge-focused studies often used supervised learning from electronic health record data, access logs, or administrative variables, while multi-agent work more often used modelling and simulation to represent coordination between actors and resources [3, 4, 6, 7, 13]. The most recent literature on agentic AI and healthcare governance began to consider more flexible AI systems, including large language model–enabled agents, but robust hospital operations deployments remained sparse [29, 30]. Hybrid architectures combining predictive models, rules, human approval, and operational dashboards appeared more common than end-to-end autonomous agents.

## Implementation and deployment status

Most systems were evaluated retrospectively, in simulation, or as pilot implementations rather than as mature, multi-site, live autonomous systems. Some implementation studies described real-world integration of machine-learning decision support into routine care, showing that deployment requires governance, monitoring, workflow redesign, and user engagement beyond model development [8, 22]. Hospital operations work demonstrated practical interest in integrating AI into throughput and resource-management processes, but the evidence base rarely included long-term deployment across multiple institutions [4, 5, 21]. This pattern indicates an implementation gulf between technically promising agentic concepts and routine operational practice.

## Reported outcomes and metrics

Reported outcomes varied widely and included discharge prediction, discharge planning support, workflow efficiency, operational forecasting, model usability, staff perceptions, and implementation feasibility. Several studies reported that AI tools could support earlier identification of likely discharges or improve planning conversations, but the review did not treat these findings as proof of autonomous effectiveness because study designs and contexts differed substantially [2, 11, 12, 14, 25]. Implementation and human-centred evaluation studies highlighted acceptance, interpretability, and fit with existing routines as important outcomes alongside technical performance [8, 10, 26, 29]. Few studies used a comprehensive outcome framework that combined time, safety, accountability, staff burden, patient experience, and unintended consequences.

## Barriers and facilitators

Common barriers included fragmented electronic health record integration, limited interoperability, unclear accountability, clinician trust concerns, insufficient safety reporting, and difficulty aligning AI outputs with real-time operational priorities. Facilitators included human-centred design, transparent model outputs, embedded workflows, multidisciplinary governance, and incremental deployment in lower-risk tasks [8-10, 22, 29]. Governance literature also identified bias, automation complacency, unclear responsibility, and lack of external validation as recurring obstacles to safe AI deployment [15-17, 27]. These barriers were especially relevant for agentic AI because systems that can plan and act require stronger controls than systems that merely display predictions.

## The rise of agentic AI in operations

The reviewed literature suggests that agentic AI in healthcare operations is emerging from the convergence of predictive analytics, workflow automation, multi-agent systems, and more recent interest in autonomous AI agents. Earlier studies focused on prediction and optimisation, while newer work increasingly emphasises implementation, oversight, and governance as essential components of operational AI [5, 6, 8, 29, 30]. The technical feasibility of agentic systems is increasing as hospitals accumulate structured workflow data and integrate AI outputs into electronic health record environments. Nevertheless, the current evidence supports cautious interpretation because many systems remain decision-support tools rather than autonomous operational agents.

**Figure 2** synthesizes the review findings into an evidence-to-implementation map that connects the current evidence base, operational domains, agent capabilities, autonomy levels, human oversight models, safety guardrails, accountability mechanisms, reported outcomes, and remaining research needs.

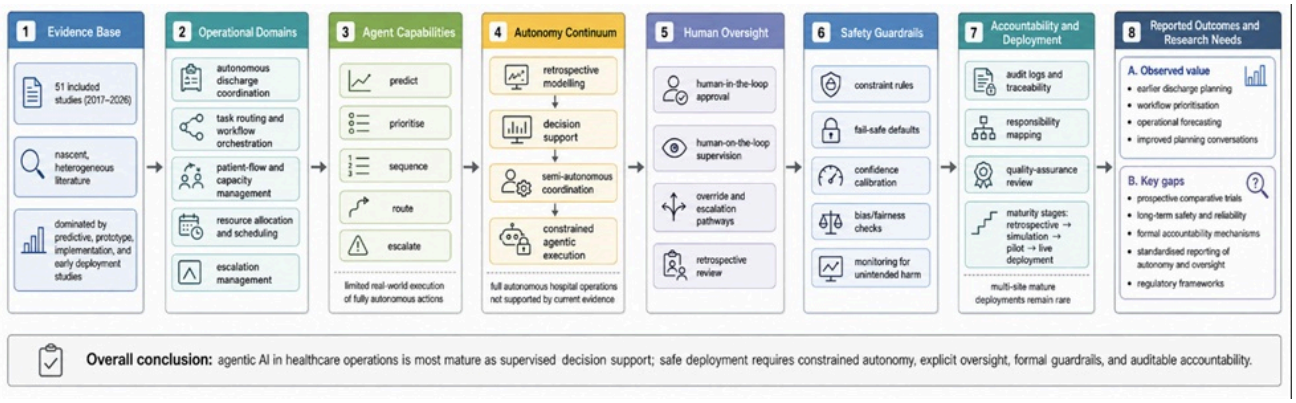



**Figure 2.** Evidence-to-Implementation Synthesis Map of Agentic Artificial Intelligence in Healthcare Systems

## Discharge coordination as the prototypical multi-step task

Discharge coordination appears to be the most developed operational domain for agentic AI because it involves predictable goals, recurring dependencies, measurable process states, and significant operational consequences. Studies of discharge prediction and planning show that AI can help identify likely discharges, support capacity planning, and focus multidisciplinary attention on patients who may be ready to leave hospital [1-3, 11-14, 25]. However, predicting discharge readiness is not the same as autonomously coordinating the full discharge process. A true agentic discharge system would need to sequence tasks, monitor dependencies, escalate unresolved barriers, and document actions in an auditable manner.

## Task routing shows promise for real-time resource optimization

Task routing and workflow orchestration show promise because hospital operations require continual reallocation of staff, beds, transport, diagnostics, and consult capacity. Multi-agent and operational AI studies illustrate how algorithmic systems could support dynamic coordination across patient pathways and hospital resources [5-7, 21]. Yet few reviewed systems operated with independent authority to assign or re-route tasks in real time. The evidence therefore supports the view that AI-enabled task routing is a plausible near-term application, but its safe deployment depends on constrained autonomy and clearly defined escalation rules.

## Human oversight is universal but heterogeneous

Human oversight was present across the evidence base, but the reviewed studies rarely used a shared taxonomy to distinguish human-in-the-loop approval, human-on-the-loop supervision, or retrospective human auditing. Implementation studies suggest that oversight works best when AI recommendations are embedded into familiar workflows, explanations are meaningful to clinicians, and staff can challenge or override model outputs [8, 10, 22, 26]. Ethical analyses similarly emphasise that human responsibility remains central when AI influences clinical or operational decisions [15, 16]. The absence of standardised oversight reporting makes it difficult to compare systems or determine which supervisory model is appropriate for different levels of autonomy.

## Safety guardrails are often implicit or incomplete

Safety guardrails were frequently acknowledged but often incompletely described, particularly in studies focused on model performance or operational feasibility. Formal reporting guidelines for AI interventions require clearer description of the intervention, human-AI interaction, monitoring procedures, and failure handling, but many operational studies predated or did not fully operationalise these expectations [19, 23, 24]. Responsible machine-learning literature warns that models can create harm through bias, poor calibration, automation dependence, and mismatch with clinical context [15, 17, 27]. For agentic AI, these concerns are amplified because unsafe outputs may lead not only to recommendations but also to action sequences.

## Accountability mechanisms are under-developed

Workflow accountability remains one of the least mature areas in the literature on agentic healthcare operations. Implementation studies show that logs, dashboards, and workflow records may be available, but few studies specify how these records support accountability after errors, near-misses, or contested decisions [8, 9, 22]. Ethical frameworks argue that responsibility must be distributed across designers, institutions, clinicians, vendors, and governance bodies rather than assigned vaguely to “the algorithm” [15-17]. Without explicit accountability models, autonomous workflow systems may obscure rather than clarify who is responsible for operational decisions.

**Table 2** outlines a deployment-oriented evaluation and accountability framework for future studies of agentic AI in healthcare operations, emphasizing autonomy specification, oversight, safety guardrails, auditability, deployment maturity, outcome measurement, and long-term reliability.

**Table 2.** Evaluation and Accountability Framework for Agentic AI Deployment in Healthcare Operations

| Evaluation domain | What future studies should explicitly report or measure | Example indicators or review questions | Frequent evidence gap identified in this review | Practical implication |
|---|---|---|---|---|
| Autonomy specification | Distinguish whether the system predicts, recommends, sequences, assigns, executes, or escalates tasks | Who can initiate action? What can be completed without approval? What authority is retained by humans? | Autonomy often implied rather than explicitly defined | Prevents overstatement of agentic capability |

| | | | | |
|---|---|---|---|---|
| Human oversight | Describe human-in-the-loop, human-on-the-loop, and retrospective audit arrangements, including override points | Approval checkpoints; override rate; escalation response time; staff role clarity | Oversight models are heterogeneous and rarely compared | Clarifies supervisory burden and safety expectations |
| Safety guardrails | Report constraints, fail-safe procedures, edge-case handling, and monitoring for harmful or biased outputs | Hard stops; confidence thresholds; near-miss capture; subgroup performance; conflict detection | Guardrails are often implicit in human review only | Supports safer deployment in complex workflows |
| Accountability and auditability | Specify event logs, traceability, responsibility mapping, and retrospective review procedures | Action traceability; audit-log completeness; responsibility matrix; adverse-event review | Auditability is rarely linked to quality assurance or liability processes | Enables post-event learning and professional accountability |
| Deployment maturity | Distinguish retrospective analysis, simulation, pilot deployment, live single-site use, and multi-site routine deployment | Maturity stage; duration of use; drift monitoring; workflow adaptation across sites | Evidence is dominated by early-stage studies | Tempers claims of real-world readiness |
| Outcome framework | Measure operational, safety, staff, patient, and equity outcomes together | Discharge delay; task completion time; adverse events; staff burden; patient experience; subgroup effects | Reported outcomes are fragmented and inconsistent | Supports meaningful cross-study comparison |
| Reliability over time | Evaluate drift, downtime, recovery procedures, and performance under changing hospital conditions | Recalibration frequency; outage handling; seasonal variation; edge-case performance | Long-term reliability is rarely studied | Essential for sustained trust and governance |

## From pilot to practice: the implementation gulf

The movement from prototype to routine practice remains difficult because operational AI must function within complex clinical environments shaped by staffing, culture, regulation, local workflow, and changing patient populations. Studies of healthcare AI implementation highlight that success requires stakeholder engagement, monitoring, maintenance, and alignment with everyday work rather than one-time technical deployment [8-10, 22, 29]. Discharge and hospital-flow studies show operational relevance, but few provide long-term evidence across multiple sites or describe how systems respond to drift, outages, or unexpected workflow disruptions [4, 13, 25]. This implementation gulf is central to the future of agentic AI because autonomy increases both potential benefit and potential risk.

## Limitations

### Review limitations

This review was limited by English-language eligibility criteria, heterogeneity of terminology, and the likelihood that some relevant systems were described as operational analytics or decision support rather than agentic AI. The review also included governance, reporting, and implementation literature because few studies directly evaluated autonomous agents in hospital operations, which may broaden the evidence base beyond strictly agentic systems [18, 19, 23, 24, 27].

Quantitative meta-analysis was not possible due to differences in tasks, settings, designs, outcomes, and deployment maturity. Publication bias is also possible because failed implementations, unsafe prototypes, and abandoned operational AI projects may be less likely to appear in peer-reviewed journals.

### Evidence base limitations

The evidence base was dominated by prediction studies, simulations, early implementation reports, and conceptual or reporting frameworks, with relatively few prospective evaluations of agentic workflow systems. Studies of discharge planning and hospital operations demonstrated practical relevance but rarely evaluated autonomous task execution, long-term safety, or accountability mechanisms under routine operating conditions [1-4, 11, 13, 21, 25]. Human oversight and safety guardrails were often described narratively rather than measured as primary outcomes, making it difficult to determine whether systems were robust to edge cases or workflow disruption [8, 15, 19, 22]. As a result, the current literature supports cautious development and structured evaluation rather than claims of readiness for broad autonomous deployment.

## Comparison with prior reviews

Prior reviews of healthcare AI have largely focused on clinical decision support, diagnostic model evaluation, predictive modelling quality, or reporting standards rather than goal-driven autonomous systems for hospital administrative workflows. For example, reviews and methodological analyses have examined the quality of deep-learning claims, the design of AI intervention trials, and the general translation of machine-learning tools into care delivery [18, 19, 23, 24, 27]. These contributions are essential, but they do not fully address the operational features that distinguish agentic AI from conventional prediction tools. In particular, they give limited attention to autonomous discharge coordination, real-time task routing, workflow accountability, and supervisory control in hospital operations.

This review extends prior work by examining the full stack of agentic AI for healthcare operations, including autonomy level, operational task type, oversight model, safety guardrails, technology stack, and deployment maturity. The included literature shows that hospital AI systems often begin as predictive models or workflow dashboards before becoming candidates for more autonomous coordination [1-5, 11, 13]. Multi-agent and agent-based studies provide conceptual and technical foundations for coordination, while implementation studies reveal the organisational requirements for embedding AI into real clinical settings [6-9, 22]. This broader perspective is necessary because agentic AI cannot be evaluated solely by model performance; it must also be evaluated by how safely and accountably it acts within a workflow.

A further distinction from prior reviews is the emphasis on safety and oversight as defining characteristics of operational AI agents. Ethical, responsible AI, and human-centred evaluation studies show that user trust, transparency, escalation, monitoring, and accountability are not secondary concerns but core requirements for systems that influence clinical work [10, 15-17, 26]. Reporting frameworks such as CONSORT-AI, SPIRIT-AI, and DECIDE-AI provide useful foundations, but agentic systems may require more explicit reporting of action authority, constraint enforcement, and retrospective review [19, 23, 24]. This review therefore positions operational autonomy as a governance challenge as much as a technical development.

## Recommendations

### For researchers

Researchers should standardise the reporting of autonomy levels, human oversight models, safety guardrails, escalation protocols, audit logs, and adverse workflow events in studies of healthcare AI agents. Existing AI reporting guidelines provide a foundation, but they should be expanded for operational agents to specify whether the system predicts, recommends, sequences, assigns, executes, or escalates tasks [19, 23, 24]. Studies should also distinguish between retrospective model performance, simulated workflow impact, pilot deployment, and routine operational use, because

these maturity levels imply different safety and accountability claims [5, 8, 22]. Without such standardisation, the field will continue to conflate predictive decision support with agentic workflow execution.

### For healthcare organisations

Healthcare organisations should introduce agentic AI incrementally, beginning with low-risk operational tasks that preserve explicit human approval and provide clear override mechanisms. Discharge prediction, capacity forecasting, and prioritisation dashboards may be appropriate early use cases because they support planning without requiring independent clinical authority [2, 4, 11, 12, 14, 25]. Organisations should also establish multidisciplinary governance involving clinicians, operational leaders, informaticians, safety officers, and patients where relevant, since implementation studies show that workflow fit and stakeholder trust strongly influence adoption [8-10, 29]. Audit trails should be treated as operational safety infrastructure rather than optional technical logs.

### For regulators

Regulators should develop frameworks for certification and post-deployment monitoring of autonomous healthcare operations agents, especially when systems initiate, sequence, or route tasks that affect patient flow or clinical workload. Current AI governance and evaluation literature indicates that safety cannot be inferred from internal validation alone, particularly when models interact with complex socio-technical environments [15-17, 27]. Regulatory approaches should require safety cases that describe intended use, boundaries of autonomy, human oversight, fail-safe behaviour, monitoring, and accountability for adverse events [19, 23, 24]. Such frameworks would help distinguish acceptable operational support from unsafe delegation of responsibility.

### For journal editors

Journal editors should require explicit description of safety and accountability measures in studies of agentic AI, autonomous agents, and workflow automation in healthcare. Manuscripts should report the system's operational authority, integration context, human review points, override options, failure modes, auditability, and procedures for detecting harmful or inequitable outcomes [15, 18, 19, 23, 24]. For discharge coordination and task-routing studies, authors should clarify whether AI merely predicts workflow states or directly influences task sequencing and assignment [1-3, 11,20]. These requirements would improve interpretability across studies and reduce the risk that autonomy is implied without adequate evidence.

## Research gaps

### Prospective clinical trials of agentic AI

A major research gap is the absence of rigorous prospective studies comparing agent-driven workflows with usual care on operational, safety, staff, and patient-centred outcomes. Existing discharge and workflow studies provide evidence that prediction and planning tools may support hospital operations, but most do not test autonomous coordination against a controlled comparator [1, 2, 11-14, 20, 25]. Implementation studies demonstrate that embedding AI into care delivery is possible, yet they also show that prospective evaluation must include workflow adaptation, human behaviour, and monitoring over time [8, 22]. Future studies should therefore evaluate agentic systems as socio-technical interventions rather than isolated algorithms.

### Long-term safety and reliability

Long-term safety and reliability remain under-studied, especially with respect to model drift, changing hospital conditions, edge cases, staffing variation, and unexpected operational disruptions. Hospital-flow and discharge models may perform differently across seasons, sites, patient populations, and electronic health record configurations, which limits confidence in generalisation [3, 4, 13, 25, 28]. Responsible AI literature warns that systems can degrade or behave unpredictably after deployment when data distributions and workflows change [15, 17, 27]. Agentic systems require continuous

monitoring because their risks may arise not only from incorrect predictions but also from action sequences that amplify small errors.

### Workflow accountability and legal responsibility

Legal and professional responsibility for autonomous workflow actions remains unresolved in the reviewed literature. Ethical analyses emphasise that clinicians, institutions, developers, and vendors may all contribute to AI-mediated decisions, but few operational studies specify how responsibility would be assigned after an erroneous discharge recommendation, unsafe task routing decision, or missed escalation [15-17]. Implementation studies show that audit logs and governance processes are essential, but they rarely test accountability mechanisms under real adverse-event conditions [8, 9, 22]. Future research should therefore examine accountability as an empirical and legal design problem, not merely as a documentation requirement.

## Conclusion

Agentic AI for healthcare operations is an emerging but rapidly advancing field, with autonomous discharge coordination and task routing representing some of the most promising early applications. The current evidence suggests that these systems are more mature as predictive and decision-support tools than as fully autonomous operational agents.

Human oversight remains universally required, but the form, depth, and timing of that oversight vary substantially across systems. Safety guardrails and accountability mechanisms are still under-developed, inconsistently reported, and rarely tested under real-world adverse conditions.

The evidence base is dominated by prototype systems, retrospective analyses, simulations, and early implementation studies. Real-world prospective evaluations of autonomous hospital workflow agents remain rare, and long-term evidence on reliability, safety, and organisational consequences is still lacking.

A coordinated research, governance, and regulatory effort is needed to define safe, accountable, and effective uses of agentic AI in hospital operations. Until such evidence matures, healthcare organisations should pursue constrained, auditable, human-supervised deployment rather than broad autonomous delegation.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 21 Mar 2026 Revised: 07 Apr 2026 Accepted: 30 May 2026

Published online: 20 July 2026

## Rights and permissions

**Open Access** The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

# References

McCoy TH Jr, Pellegrini AM, Perlis RH. Assessment of time-series machine learning methods for forecasting hospital discharge volume. JAMA Netw Open. 2018;1(7):e184087.

Safavi KC, Khaniyev T, Copenhaver M, Seelen M, Zenteno Langle AC, Zanger J, et al. Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. JAMA Netw Open. 2019;2(12):e1917221.

Zhang X, Yan C, Malin BA, Patel MB, Chen Y. Predicting next-day discharge via electronic health record access logs. J Am Med Inform Assoc. 2021;28(12):2670-80.

Gao R, Cheng WX, Suganthan PN, Yuen KF. Inpatient discharges forecasting for Singapore hospitals by machine learning. IEEE J Biomed Health Inform. 2022;26(10):4966-75.

Pianykh OS, Guitron S, Parke D, Zhang C, Pandharipande P, Brink J, et al. Improving healthcare operations management with machine learning. Nat Mach Intell. 2020;2(5):266-73.

Montagna S, Castro Silva D, Henriques Abreu P, Ito M, Schumacher MI, Vargiu E. Autonomous agents and multi-agent systems applied in healthcare. Artif Intell Med. 2019;96:142-4.

Ajmi F, Zgaya H, Othman SB, Hammadi S. Agent-based dynamic optimization for managing the workflow of the patient's pathway. Simul Model Pract Theory. 2019;96:101935.

Sendak MP, D'Arcy J, Kashyap S, Gao M, Nichols M, Corey K, et al. A path for translation of machine learning products into healthcare delivery. EMJ Innov. 2020;4(1):19-00172.

Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. BMJ Innov. 2020;6(2):45-7.

Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020. p. 1-12.

Levin S, Barnes S, Toerper M, Debraine A, DeAngelo A, Hamrock E, et al. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. BMJ Innov. 2021;7(2):388-95.

Ahn I, Gwon H, Kang H, Kim Y, Seo H, Choi H, et al. Machine learning-based hospital discharge prediction for patients with cardiovascular diseases: development and usability study. JMIR Med Inform. 2021;9(11):e32662.

Wei J, Zhou J, Zhang Z, Yuan K, Gu Q, Luk A, et al. Predicting individual patient and hospital-level discharge using machine learning. Commun Med. 2024;4(1):236.

Pahlevani M, Taghavi M, Vanberkel P. A systematic literature review of predicting patient discharges using statistical methods and machine learning. Health Care Manag Sci. 2024;27(3):458-77.

Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40.

Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981-3.

Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. NPJ Digit Med. 2019;2(1):77.

Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. https://doi.org/10.1136/bmj.m689.

Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904.

Zhao CC, Bjurlin MA, Wysock JS, Taneja SS, Huang WC, Fenyo D, et al. Machine learning decision support model for radical cystectomy discharge planning. Urol Oncol. 2022;40(10):453.e9-453.e15.

Bellini V, Russo M, Domenichetti T, Panizzi M, Allai S, Bignami EG. Artificial intelligence in operating room management. J Med Syst. 2024;48(1):19.

Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med Inform. 2020;8(7):e15182.

Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Ashrafian H, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020;2(10):e537-48.

Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, Ashrafian H, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020;2(10):e549-60.

Acosta-Perez F, Boutilier J, Zayas-Caban G, Adelaine S, Liao F, Patterson B. Toward real-time discharge volume predictions in multisite health care systems: longitudinal observational study. J Med Internet Res. 2025;27:e63765.

Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Machine Learning for Healthcare Conference. 2019. p. 359-80.

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.

Ghazalbash S, Zargoush M, Mowbray F, Papaioannou A. Examining the predictability and prognostication of multimorbidity among older delayed-discharge patients: a machine learning analytics study. Int J Med Inform. 2021;156:104597.

Finkelstein J, Gabriel A, Schmer S, Truong TT, Dunn A. Identifying facilitators and barriers to implementation of AI-assisted clinical decision support in an electronic health record system. J Med Syst. 2024;48(1):89.

Collaco BG, Haider SA, Prabha S, Gomez-Cabello CA, Genovese A, Wood NG, et al. The role of agentic artificial intelligence in healthcare: a scoping review. NPJ Digit Med. 2026;9(1).