ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir

Evidence-Gated Autonomy in Healthcare AI Agents: A Systematic Review of Action Scope, Human-Oversight Effectiveness, and Multi-Step Safety Failures

H. Ghayoumi Zadeh [1,*], Kh. Rezaee[2], A. Fayazi[1], O. Rahmani Seryasat[3]

[1] Associate Professor, Department of Electrical Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.
[2] Department of Biomedical Engineering, Meybod University, Meybod, Iran.
[3] Department of Electrical Engineering, Shams Gonbad Higher Education Institute, Gorgan, Iran.

ARTICLE INFO

Article History:
Received 3 December 2025
Received in revised form 14 January 2026
Accepted 3 February 2026
Available online 5 March 2026

Keywords:
Artificial Intelligence; AI Agents; Agentic AI; Healthcare Workflow; Patient Safety; Operational Autonomy; Action Scope; Human Oversight; Oversight Effectiveness; Multi-Step Failure; Evidence-Gated Autonomy; Systematic Review

ABSTRACT

Background: Healthcare AI agents increasingly plan, invoke tools, and execute multi-step workflows. Existing reviews have mapped applications and architectures, but have not jointly classified consequential action authority, multi-step failure expression, and the empirically measured effectiveness of human oversight. This review evaluates whether greater operational autonomy is supported by evidence on safety and human control.

Methods: PubMed/MEDLINE, Scopus, and Web of Science Core Collection were searched from 1 January 2022 through 27 July 2026. The searches retrieved 4,096 records, of which 2,557 unique report-level records remained after deduplication. For the current synthesis, 50 priority reports underwent full-text retrieval; 34 were assessed, 28 met the operational agent criteria, six were excluded, and one was unavailable. One reviewer performed AI-assisted structured extraction of workflow, architecture, action scope, autonomy, safety, oversight, evaluation, and governance data. Heterogeneous evidence was synthesized narratively following SWiM principles.

Results: The 28 studies were published from 2023 to 2026; 18 (64.3%) appeared in 2026. Diagnostic and clinical decision-support workflows accounted for 13 studies, EHR/information or administrative workflows for nine, treatment planning or prescribing for four, and patient-facing or cross-domain workflows for two. Twelve systems were single-agent, 12 multi-agent, three hybrid, and one incompletely specified. Operational autonomy was L0 in five studies, L1 in 13, L2 in three, L3 in one, and L4 in six; none reached L5. Five of the six L4 systems operated only in sandboxes or retrospective research environments, whereas one prospective clinical trial implemented live conditional prescribing. Safety-evidence grades were high in one study, moderate in 15, low in 11, and very low in one. Only three studies empirically characterized an operating property of oversight, and none measured approval-gate catch rate under sustained clinical workload.

* Corresponding Author: h.ghayoumizadeh@vru.ac.ir
Associate Professor, Department of Electrical Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.

http://dx.doi.org/10.22034/TMI.2026.248202

© 2026 by the authors. Licensee T.M.I, Tehran, Iran. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).

Conclusions: The current evidence supports tool-using and workflow-capable agents, but not unrestricted clinical autonomy. Operational capability frequently exceeded the maturity of safety evaluation, and most oversight mechanisms were asserted by design rather than validated in use. Autonomy should therefore be promoted only within a bounded operational design domain and only after enacted and propagated failures, escalation sensitivity, override behavior, and reviewer workload have been measured.

1. INTRODUCTION

1.1. From predictive AI to agentic AI in healthcare

Over the past decade, artificial intelligence (AI) in healthcare has been dominated by predictive and classificatory paradigms. Deep learning systems trained on medical images, physiological signals, and structured electronic health record (EHR) data achieved expert-level performance on narrowly defined tasks such as diabetic retinopathy screening, skin-lesion classification, and sepsis risk prediction [1,2]. These systems, however, shared a defining architectural property: they were passive. A model received a fixed input, produced a probability or label, and stopped. Any subsequent action - ordering a test, adjusting a medication, documenting an encounter - remained entirely in human hands. The clinical AI literature, its evaluation frameworks, and its regulatory apparatus were all built around this single-input, single-output paradigm [3,4].

The arrival of large language models (LLMs) initiated a first shift in this landscape. Foundation models such as GPT-4 and Med-PaLM demonstrated that a single general-purpose model could encode broad clinical knowledge, pass medical licensing examinations, and generate differential diagnoses and clinical documentation in free text [5,6]. Moor and colleagues anticipated that such generalist medical AI would eventually "carry out diverse tasks with minimal task-specific supervision" [7]. Yet even these early LLM applications remained fundamentally conversational: they answered questions and drafted text, but they did not act.

The current transition - the subject of this review - is qualitatively different. Since approximately 2023, LLMs have been embedded within agentic architectures: systems that decompose a goal into steps, maintain memory across those steps, invoke external tools and application programming interfaces (APIs), and execute multi-step plans with varying degrees of independence [8,9]. In an agentic configuration, the model no longer merely suggests that a laboratory test may be informative; it can query the EHR to check whether the test was already performed, draft the order, route it for approval, and schedule follow-up. This shift from language models as knowledge repositories to agents as workflow participants has been described as the "agentic turn" in medical AI, and recent surveys document its rapid expansion across biomedicine [9,10].

Concrete implementations now span the full breadth of healthcare workflows. In clinical documentation, ambient AI scribes - systems that record the clinical encounter, generate a structured draft note, and in newer versions propose billing codes and pending orders - have moved from pilots to large-scale deployment, with one integrated delivery system reporting more than 2.5 million uses within a single year [11]. Randomized trials have begun to quantify their effect on documentation time and clinician burnout [12,13]. In EHR interaction, agents such as EHRAgent translate clinical questions into executable code over patient databases [14], while Almanac Copilot navigates the record autonomously to retrieve and act on patient information [15]. In administrative workflows, multi-agent frameworks have been applied to prior authorization, medical-necessity justification, and scheduling [16,17]. In diagnostics and treatment planning, conversational diagnostic agents such as AMIE have matched or exceeded primary-care physicians on selected consultation metrics [18], tool-calling agents have been developed for autonomous clinical decision support in oncology [19], and orchestrated multi-agent "teams" of LLMs have been proposed to emulate multidisciplinary clinical collaboration [20,21]. Entire simulated hospitals populated by interacting agents are now used to train and evaluate such systems before real-world exposure [22].

Three features distinguish these agentic systems from their predictive predecessors and motivate the present review. First, action scope: agents write into live clinical and administrative systems, so their errors are no longer confined to an incorrect suggestion that a clinician may ignore, but can propagate into orders, messages, and records.

Second, compositionality: agents chain many model calls and tool invocations, allowing small per-step error rates to compound across a workflow in ways that single-shot evaluation cannot capture [8,23]. Third, variable autonomy: identical underlying models can be deployed anywhere on a spectrum from fully supervised drafting to unsupervised end-to-end execution, and the level of autonomy is a design decision - often an implicit one - rather than an intrinsic model property [24]. As agentic systems diffuse into routine care faster than the evidence base that should govern them, the questions of how much autonomy these agents are actually granted, how safely they behave, and how humans oversee them become central to patient safety and to emerging regulatory regimes such as the EU Artificial Intelligence Act and evolving FDA guidance on AI-enabled clinical software [25,26]. These questions define the scope of this systematic review.

1.2. Why agents raise new safety and governance questions

The defining property of an agent - the capacity to act - fundamentally alters the risk profile of clinical AI. In the predictive paradigm, an erroneous output was an incorrect recommendation: consequential only if a clinician accepted it, and by design always mediated by human judgment. In an agentic configuration, an erroneous output can itself be an event in the world. An agent that hallucinates a laboratory value does not merely mislead a reader; it may write that value into a note that becomes the substrate for downstream decisions, message a patient with incorrect instructions, or draft an order that enters the execution pathway. Error severity thus becomes a function not only of model accuracy but of action scope: what systems the agent can touch, whether its actions are reversible, and how quickly a human can detect and retract them [8,27]. This coupling between output and action has no counterpart in the static models around which clinical AI evaluation was built [3].

A second source of novel risk is compositionality. Agentic workflows chain together many model invocations, tool calls, and intermediate decisions, so that per-step error rates that appear acceptable in isolation can compound multiplicatively across a task: a system that is 95% reliable at each of ten sequential steps completes the full workflow correctly only about 60% of the time. Empirical analyses of multi-agent LLM systems confirm that failures frequently arise not within a single model call but between steps - through mis-specified hand-offs, premature termination, faulty tool use, and inter-agent misalignment - and have produced dedicated failure taxonomies for such systems [28]. Early benchmarks that embed agents in realistic virtual EHR environments similarly find that current models complete only a fraction of routine clinical tasks reliably, with failure modes that single-turn question-answering benchmarks are structurally unable to detect [29]. Hallucination, extensively documented in medical LLM outputs [30], is thereby transformed from a textual artifact into a potential hallucinated action.

Third, agents reshape the human side of the sociotechnical system. Most deployed configurations interpose a human approval step between agent proposal and execution, on the assumption that clinician review neutralizes agent error. Decades of human-factors research caution against this assumption: automation bias - the tendency to accept automated outputs with insufficient scrutiny - and automation complacency are robust, replicated phenomena that intensify as automation becomes more reliable and workload increases [31,32]. In high-volume workflows, approval gates risk degenerating into click-through rituals, an effect already familiar from alert fatigue in clinical decision support [33]. Studies of ambient documentation agents show that clinicians edit AI drafts variably and sometimes minimally, raising the question of whether nominal human oversight corresponds to effective human oversight [34]. Oversight, in other words, is not a binary property that a deployment either has or lacks; it is a designed mechanism whose real-world efficacy is an empirical question - one that, as this review will show, is rarely measured.

Fourth, agency complicates accountability. When a supervised classifier errs, responsibility questions - though contested - at least involve a short causal chain from model output to human decision [35,36]. Agentic systems distribute causation across model developers, orchestration-layer designers, tool providers, deploying institutions, and supervising clinicians, and across many steps in time. Existing legal and professional frameworks presuppose an identifiable human decision-maker at the point of care; the more an agent plans and executes independently, the harder it becomes to locate that decision-maker, and the greater the risk of both accountability gaps and defensive over-attribution of blame to frontline clinicians [36,37].

Finally, governance frameworks are only beginning to catch up. The EU Artificial Intelligence Act classifies most health-related AI as high-risk and mandates "effective human oversight," but does not specify what effective

oversight means for systems that act in multi-step loops [25]. FDA guidance on AI-enabled device software addresses adaptive model updates through predetermined change control, yet its evaluative machinery remains oriented toward fixed input-output functions rather than open-ended tool-using agents [26]. The World Health Organization's guidance on large multi-modal models likewise flags autonomy as an emerging concern without providing an operational grading of it [38], and commentators have noted that LLM-based systems sit awkwardly within existing medical-device categories altogether [39]. The resulting governance problem is not simply whether a human is nominally present, but whether the authority granted to an agent is matched by evidence that its actions are safe and that the oversight layer can detect, interrupt, or contain failures. The purpose of this review is therefore to determine whether the operational autonomy granted to healthcare AI agents is supported by empirical evidence on multi-step safety and human-oversight effectiveness.

1.3. Gap statement and contributions of this review

The existing secondary literature has substantially advanced understanding of medical AI agents but leaves a distinct gap relative to this objective. Reviews of medical LLMs synthesize diagnostic, documentation, and question-answering performance but largely treat models as conversational or predictive systems [40]. Healthcare-focused reviews have mapped applications, architectures, external tools, evaluation settings, and broad translational gaps [8,9,41,42]. A systematic review of clinical-agent studies has additionally compared agent systems with baseline LLMs and catalogued tool use and architecture-performance relationships [43]. A subsequent clinical-practice evidence map has highlighted workflow integration and deployment maturity [44]. None of these reviews jointly classifies consequential action scope and reversibility, distinguishes contained failures from enacted and propagated multi-step failures, and evaluates whether approval, escalation, monitoring, override, or audit mechanisms have demonstrated operating characteristics in realistic use.

Three linked evidence gaps therefore remain. First, autonomy terminology is not operationalized consistently: labels such as "autonomous," "copilot," and "assistant" are used for systems with radically different action scopes, from passive information support to preparation or execution of consequential actions. Second, safety evidence is usually reported as output accuracy or contained textual error, whereas the clinically distinctive risks of agents arise when errors are enacted through tools or silently propagated across dependent steps. Third, human oversight is frequently described as a design feature but rarely treated as an intervention whose effectiveness must be measured under realistic workload, including gate catch-rate, escalation sensitivity, override success, reviewer vigilance, and audit detection probability. These gaps prevent capability results from establishing whether a system is ready to receive greater action authority.

Accordingly, this systematic review has one overarching objective: to determine whether increasing operational autonomy in healthcare AI agents is supported by empirical evidence on multi-step safety and the effectiveness of human oversight. It makes three contributions. First, it characterizes evaluated systems by healthcare workflow, architecture, deployment stage, consequential action scope, reversibility, and operational design domain. Second, it applies a six-level L0-L5 taxonomy to distinguish rhetorical descriptions of autonomy from operational authority. Third, it cross-tabulates autonomy, task risk, failure expression, and measured oversight effectiveness to construct an autonomy-oversight-risk matrix and a framework for evidence-gated autonomy promotion.

1.4. Research questions

Guided by the gaps identified above, this review addresses one primary question: among AI agents evaluated in healthcare workflows, how do operational autonomy, consequential action scope, reversibility, and task risk relate to the design and empirically measured effectiveness of human-oversight mechanisms and to reported multi-step safety failures? This question is operationalized through six research questions (RQs), each mapped to a dedicated results subsection:

RQ1. In which clinical, administrative, and patient-facing healthcare workflows have AI agents been evaluated, and with what architectures, tools, and deployment stages?

RQ2. What operational autonomy do these agents exercise when classified by consequential action scope, reversibility, human authority, and operational design domain, and how do systems distribute across the L0-L5 taxonomy?

RQ3. What safety outcomes and failure modes are reported, including content, tool-use, planning, and coordination errors, and are failures contained, enacted, or propagated across multiple workflow steps?

RQ4. What human-oversight mechanisms are implemented - including approval gates, escalation rules, monitoring, audit logging, override, and kill-switch functions - and what empirical evidence demonstrates their effectiveness?

RQ5. How do reported autonomy levels, task risks, and oversight practices align with emerging regulatory and ethical requirements, including the EU Artificial Intelligence Act, FDA guidance on AI-enabled device software, and WHO guidance on large multi-modal models?

RQ6. What evaluation methodologies, benchmarks, comparators, and study designs are used, and what methodological and evidentiary gaps limit evidence-gated progression from human-gated to increasingly autonomous operation?

RQ1 and RQ2 map where agents operate and the authority they exercise; RQ3 and RQ4 evaluate multi-step safety and the oversight layer intended to control it; RQ5 compares observed practice with governance requirements; and RQ6 appraises whether current evaluation methods can support evidence-gated autonomy claims. The remainder of this article is organized as follows: Section 2 establishes working definitions of agency, autonomy, and oversight; Section 3 reports the review methods; Section 4 presents results by research question; and Section 5 discusses implications for developers, health systems, and regulators, followed by limitations and a future research agenda.

2. BACKGROUND AND KEY DEFINITIONS

2.1. What counts as an "AI agent"

The term "agent" long predates large language models. In classical artificial-intelligence theory, an agent perceives an environment and acts upon it, with intelligent agency associated with goal-directed, proactive, and adaptive behavior [46,47]. Contemporary LLM-based agents place a language model within a control structure that can maintain task state, plan or coordinate intermediate steps, invoke tools and external systems, and produce effects outside a single conversational response [48,49]. For this review, agentic status is determined by the system's operational control structure and externalized workflow effect, not by the base model, the sophistication of its language, or labels such as "assistant," "copilot," or "autonomous."

A system was considered an AI agent only when all four elements of the prespecified action-agency test were satisfied. First, healthcare-workflow anchoring: the system was evaluated within clinical care, healthcare operations, or a patient-facing care process. Second, goal-directed multi-step state: it maintained a goal, plan, or task state across at least two dependent steps rather than producing an isolated single-shot output. Third, autonomous intermediate control: it selected, sequenced, delegated, or revised at least one intermediate step without turn-by-turn human specification. Fourth, externalized workflow effect: it invoked an external tool, database, application programming interface, software system, or device, or it created, prepared, or committed a structured workflow artifact or action beyond a free-form chat reply. Read-only tool use could satisfy this fourth criterion; consequential clinical action was not required for agentic eligibility and was graded separately in the autonomy taxonomy [48,49].

This operational test deliberately separates agentic status from autonomy level. A read-only EHR agent that autonomously formulates and executes database queries may qualify as an agent while remaining L0 if it produces information only; a system that recommends a course of action is L1; and an ambient scribe that creates a structured note or pended order for explicit approval is L2. Human approval therefore constrains autonomy but does not negate agency. Conversely, purely conversational chatbots, fixed single-turn retrieval-augmented generation pipelines, static predictive or classificatory models, and systems whose intermediate steps are fully specified by the user were

excluded. Multi-agent systems were included only when the collective architecture satisfied the same operational test; merely eliciting several independent model opinions was insufficient [20,21,28]. The resulting boundary is summarized in Figure 1 and calibrated examples are provided in Box 1.

2.2. Autonomy: lessons from other safety-critical domains

Healthcare is not the first safety-critical field to confront machine autonomy, and the fields that preceded it converged on the same insight: autonomy is not a binary property but a graded, task-specific allocation of authority between human and machine. The earliest formalization is Sheridan and Verplank's ten-level scale of automation, ranging from a computer that offers no assistance to one that acts entirely on its own and ignores the human [50]. Parasuraman, Sheridan, and Wickens later refined this into a two-dimensional framework in which the level of automation is specified separately for information acquisition, information analysis, decision selection, and action implementation - a decomposition that maps naturally onto agentic pipelines, where a system may be highly autonomous in retrieving and analyzing data yet tightly constrained in executing actions [51]. Human-factors research built on these frameworks has repeatedly shown that intermediate autonomy levels create the most delicate supervision problems, because the human retains nominal responsibility while losing moment-to-moment engagement [32,52].

Two domain-specific taxonomies are especially instructive precedents. In road transport, SAE J3016 defines six levels of driving automation (L0-L5), anchored not in algorithmic sophistication but in the division of the dynamic driving task between human and system and in the operational design domain within which automation is permitted [45]. Its success lies in giving regulators, manufacturers, and users a shared, testable vocabulary - precisely what healthcare agents lack. In surgical robotics, Yang and colleagues proposed six levels running from no autonomy through task autonomy and conditional autonomy to full autonomy, and argued that each increment shifts the regulatory burden and the locus of legal responsibility [24]. Three lessons carry over to healthcare agents. First, levels must be defined by action scope and reversibility, not by underlying model capability: the same LLM can operate at different levels in different deployments. Second, each level presupposes an operational design domain - the bounded set of tasks, patients, and conditions within which the stated autonomy holds. Third, the taxonomy must be legible to non-engineers, because its principal consumers are clinicians, procurement committees, and regulators. The L0-L5 taxonomy introduced in Section 4.3 of this review is constructed on these principles.

2.3. Typology of human oversight

If autonomy describes what the machine may do, oversight describes how humans remain in command of it. Following established usage in the automation and autonomous-systems literature, we distinguish three supervisory configurations [51,53]. In human-in-the-loop (HITL) arrangements, a human is interposed in the causal chain of every consequential action: the agent proposes, and nothing executes without affirmative human approval. In human-on-the-loop (HOTL) arrangements, the agent executes actions autonomously while a human monitors its operation and can intervene, veto, or halt it; authority is exercised by exception rather than by transaction. In human-out-of-the-loop (HOOTL) arrangements, actions execute without real-time human involvement, and oversight - if any - is retrospective, through audit and quality assurance. These configurations correspond only loosely to autonomy levels: a nominally HITL deployment supervising thousands of low-salience approvals per day may afford weaker effective control than a well-instrumented HOTL deployment with crisp escalation criteria, which is why this review records oversight mechanisms separately from autonomy levels rather than inferring one from the other.

Within these configurations, concrete oversight mechanisms recur across deployments and form the extraction categories used in this review. Approval gates require explicit human sign-off before a specified action class executes (for example, note filing, order entry, or patient messaging). Escalation rules oblige the agent to hand control to a human when defined triggers fire - uncertainty above threshold, detection of a high-risk entity, patient safety keywords, or task states outside the operational design domain. Audit logging preserves a tamper-evident record of agent perceptions, plans, tool calls, and actions, enabling retrospective review and incident reconstruction. Override and kill-switch functions guarantee that a human can interrupt, reverse, or terminate agent activity at any time. The philosophical literature on meaningful human control adds two demanding conditions that mere mechanism does not secure: the system must track the reasons of the humans responsible for it, and those humans must have the

knowledge and capacity to actually exercise the control they nominally hold [53]. Regulatory texts increasingly echo this distinction - Article 14 of the EU AI Act requires oversight measures that are "commensurate with the risks" and effective in practice, not merely present in design [25]. Throughout this review, we therefore treat oversight as a set of designed, inspectable mechanisms whose real-world effectiveness is an empirical property to be evidenced - the property interrogated by RQ4.

3. METHODS

3.1. Protocol and registration

The review methods were designed around PRISMA 2020 and PRISMA-S [54,55]. A protocol amendment dated 27 July 2026 fixed the electronic source set as PubMed/MEDLINE, Scopus, and Web of Science Core Collection after pilot testing. The executed queries, exports, and deduplication audit were archived. The current analysis is based on a priority-corpus synthesis: protocol registration, complete full-corpus screening, and independent Reviewer 2 verification remain outstanding and must be completed before journal submission.

3.2. Eligibility criteria

Eligibility was defined using an adapted PICOS framework. Population/setting: any healthcare workflow, including diagnosis, triage, treatment planning, monitoring, nursing, pharmacy, documentation, coding, scheduling, prior authorization, telehealth, and patient-facing care services. Studies in non-health domains and biological or drug-discovery pipelines without a direct connection to a care-delivery workflow were excluded. Intervention: single-agent, tool-augmented, hybrid, or multi-agent AI systems satisfying all four elements of the operational action-agency test in Section 2.1. Agentic eligibility required healthcare-workflow anchoring, goal-directed state across dependent steps, autonomous control of at least one intermediate step, and an externalized workflow effect. A system was not excluded merely because its external interaction was read-only or because every consequential action required human approval; those properties informed autonomy classification rather than agentic eligibility.

Comparator: any or none, including clinicians, standard care, non-agentic AI, alternative agent architectures, or uncontrolled designs. Outcomes: at least one extractable outcome relevant to workflow performance, action scope, autonomy, safety or failure behavior, human oversight, human-agent interaction, or deployment had to be reported. Studies limited to decontextualized examination scores, free-form question answering, or benchmark accuracy without an interactive workflow or agent-specific evaluation were excluded. Study designs: original empirical research of any design, including randomized trials, prospective and retrospective evaluations, simulations, usability and human-factors studies, and technical evaluations in clinically meaningful environments. Reviews, editorials, commentaries, protocols without results, and conceptual frameworks were excluded from synthesis but retained for citation chasing. The search window extended from 1 January 2022 to the final search date ([DD Month 2026]); English-language records and eligible arXiv or medRxiv preprints were included. Preprints were flagged, examined separately in sensitivity analyses, and replaced by peer-reviewed versions when available before final extraction. At title/abstract screening, uncertainty on any action-agency criterion advanced the record to full-text review; exclusion as non-agentic required an explicit failed criterion documented by the reviewers.

Box 1. Prespecified boundary cases used to calibrate reviewer decisions.

System configuration	Decision	Operational rationale
General medical chatbot answering questions without tools or persistent task control	**Exclude**	Free-form conversational output only; no externalized workflow effect or autonomous intermediate control.
Fixed single-turn RAG pipeline that always performs one predefined retrieval before answering	**Exclude**	Tool use alone is insufficient when the control sequence is fixed and no goal-directed multi-step state is maintained.

System configuration	Decision	Operational rationale
Read-only EHR agent that autonomously formulates, runs, checks, and revises database queries	**Include**	Meets multi-step control and external-tool criteria; typically L0 when it only returns information.
Ambient scribe that produces a structured note, billing code, or pended order for clinician approval	**Include**	Creates a persistent workflow artifact; explicit per-action approval generally maps the system to L2.
Multi-agent diagnostic team that autonomously delegates roles and produces a structured recommendation	**Include**	The collective meets the action-agency test; typically L1 unless it prepares or executes a consequential action.
Scheduling agent that reserves low-risk appointments and escalates exceptions	**Include**	Executes bounded workflow actions with exception-based human monitoring; typically L3, subject to actual configuration.

Fig. 1. Operational eligibility decision tree for identifying healthcare AI agents. All five screening questions must be satisfied for inclusion; ambiguous title/abstract records proceed to full-text assessment, and non-agentic exclusion requires documentation of the failed criterion.

3.3. Information sources and search strategy

Three complementary bibliographic databases were searched: PubMed/MEDLINE, Scopus, and Web of Science Core Collection. PubMed provided biomedical indexing and controlled vocabulary, whereas Scopus and Web of Science broadened coverage to interdisciplinary health informatics, computer science, engineering, and conference literature. Backward and forward citation chasing of included studies and closely related reviews will be used as an additional identification method, but no further bibliographic databases will be searched. The final database set was fixed before formal screening and is documented as a protocol amendment.

The PubMed strategy was developed first and then translated to Scopus and Web of Science. Searches combined two complementary retrieval routes: (1) explicit terminology for AI, LLM, agentic, autonomous, tool-using, copilot, scribe, and healthcare-specific agent systems; and (2) AI or large-language-model terminology combined with concrete behaviors such as function calling, tool calling, external-tool use, API calls, database querying, EHR navigation, or named orchestration frameworks. A healthcare-context block was mandatory. Generic planning terminology and standalone multi-agent terminology were excluded after pilot searches showed unacceptable retrieval of treatment-planning and non-healthcare control-systems literature. Safety, oversight, governance, and error terms were intentionally not required because absence of such reporting is itself an outcome of interest.

Search sensitivity was tested against a prespecified sentinel set spanning tool-using EHR agents, multimodal oncology agents, interactive EHR benchmarks, multi-agent clinical systems, simulated hospitals, and ambient documentation systems. Searches were executed on 27 July 2026. PubMed retrieved 1,309 records; Scopus retrieved 1,294 records in explicit-agent Search A and 510 records in tool-behavior Search B; Web of Science retrieved 769 records in Search A and 214 records in Search B. Exact executed strategies and export specifications are provided in Supplementary Material S1.

3.4. 3.4 Study selection

Records were exported separately from each source with database, query pass, execution date, and original accession identifiers retained. The five exports contained 4,096 raw records. Cross-database deduplication used DOI, PMID, Scopus EID, Web of Science UT, and normalized title/first-author/year comparisons, with manual adjudication of ambiguous matches. A total of 1,539 duplicates were removed, leaving 2,557 unique report-level records. For the present synthesis, a rule-based priority set of 50 records was advanced for immediate full-text retrieval. Thirty-four full texts were retrieved and assessed: 28 met all operational agent criteria and six were excluded. One additional report could not be retrieved, and 15 priority records remained unassessed. This priority-corpus workflow is not equivalent to completed screening of all 2,557 records; the selection flow is therefore reported as interim.

3.5. Data extraction

For the current synthesis, Reviewer 1 performed structured AI-assisted extraction into a piloted workbook. The extracted domains were bibliographic characteristics; agent eligibility and architecture; workflow context; action scope, reversibility, and L0-L5 autonomy; safety and failure expression; oversight configuration and mechanisms; evaluation design and outcomes; and governance. Every full-text decision included an explicit rationale. Independent Reviewer 2 screening, extraction checking, consensus, and inter-rater reliability are pending; no kappa statistic is reported in this version.

3.6. Quality appraisal

Because the current corpus included benchmarks, simulations, retrospective studies, prospective human evaluation, and one live clinical trial, a four-level safety-evidence grade was applied for this synthesis. High evidence required prospective or live evaluation with clinical outcomes or direct oversight behavior; moderate evidence required realistic tool-integrated or real-data evaluation with quantitative failure analysis; low evidence denoted benchmark, retrospective, or small-sample evaluation without direct oversight-effectiveness measurement; and very low evidence denoted limited technical evidence or poorly characterized clinical validity. Formal design-matched risk-of-bias appraisal using RoB 2, QUADAS-AI, PROBAST, or the planned Agentic Safety Evidence Checklist remains to be independently completed.

3.7. Synthesis methods

Structured tabulation and narrative synthesis followed SWiM principles [65]. Counts and percentages summarized workflow domain, architecture, deployment stage, autonomy, oversight, failure type, and safety-evidence grade. No meta-analysis was attempted because tasks, denominators, comparators, and outcome definitions were incompatible. The analytical core was a cross-tabulation of autonomy level and safety-evidence grade, supplemented by qualitative comparison of approval, escalation, audit, override, and workload evidence.

4. RESULTS

4.1. Study selection and characteristics

The electronic searches yielded 4,096 records: 1,309 from PubMed/MEDLINE, 1,804 from Scopus, and 983 from Web of Science. After removal of 1,539 duplicates, 2,557 unique reports remained. The current priority subset contained 50 reports; 34 full texts were assessed, 28 were included, six were excluded, one could not be retrieved, and 15 remained unassessed (Figure 2). The six exclusions comprised three systems without an eligible operational agent, one conceptual architecture without empirical evaluation, one static classification workflow, and one fixed OCR-to-medication matching pipeline without demonstrated autonomous control.

Included studies were highly recent: 18 of 28 (64.3%) were published in 2026, eight (28.6%) in 2025, and one each (3.6%) in 2024 and 2023. Sixteen studies (57.1%) were limited to benchmarks, simulations, or sandboxes; 10 (35.7%) used retrospective or offline real-world data; and two (7.1%) included prospective human or live evaluation. Only daGOAT executed a prospective treatment action in routine hospital infrastructure [92]. Safety-evidence grades were high in one study (3.6%), moderate in 15 (53.6%), low in 11 (39.3%), and very low in one (3.6%). Characteristics of all 28 studies are summarized in Table 1.

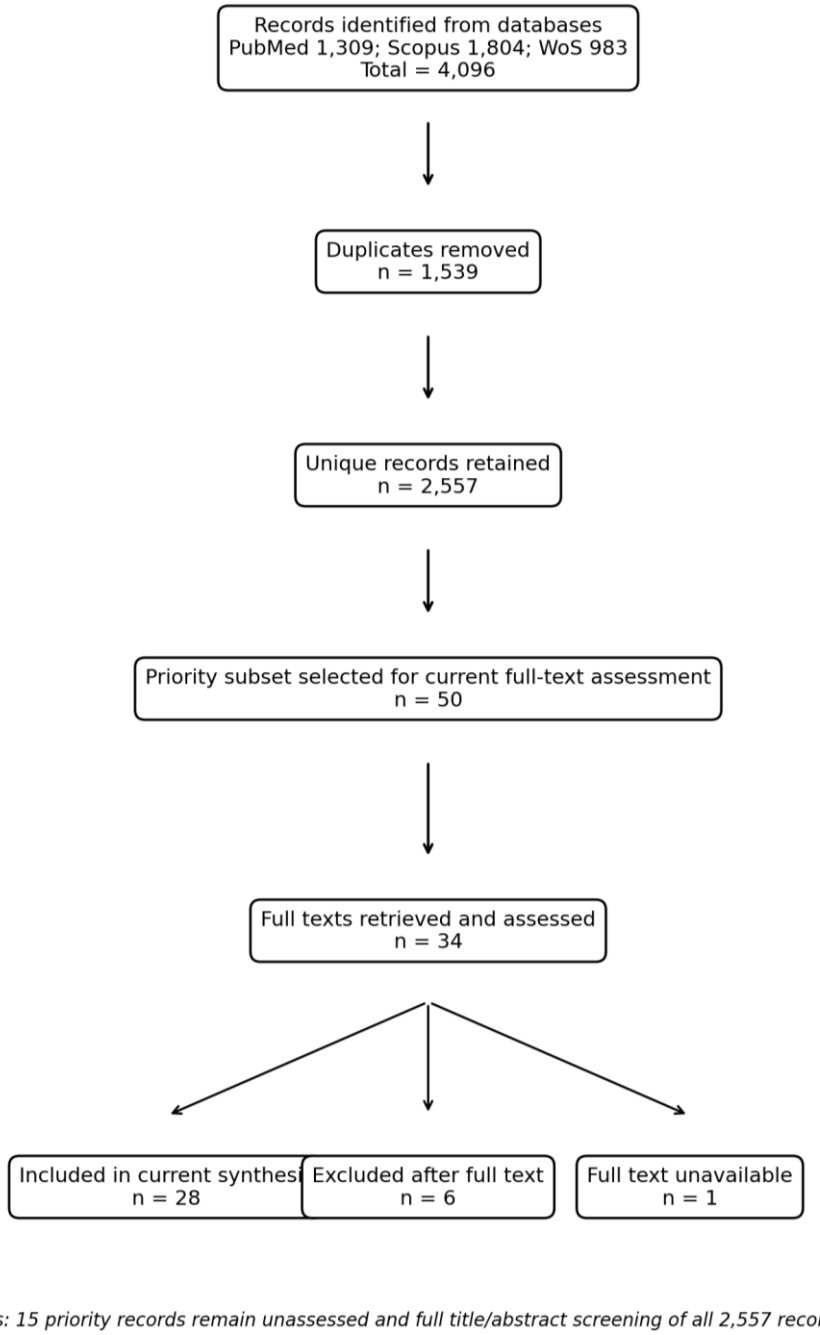

Fig. 2. Study-selection flow for the current 28-study synthesis.

4.2. Healthcare workflows and agent architectures (RQ1)

4.2.1. Clinical workflows

Thirteen studies (46.4%) primarily evaluated diagnosis, triage, or clinical decision support. Examples ranged from interactive diagnostic simulation and full-process diagnosis to oncology evidence synthesis, emergency triage, neuro-symbolic guideline reasoning, and multimodal PET/CT interpretation [70,72,74-83,87]. Performance was

usually reported as diagnostic accuracy, guideline agreement, or clinician-rated appropriateness. Only DxDirector incorporated a quantitatively evaluated risk supervisor with explicit halt-and-handoff behavior [75].

4.2.2. Administrative workflows

Nine studies (32.1%) targeted EHR, information-management, or administrative workflows. These included EHR navigation and code execution, FHIR read/write sandboxes, rare-disease screening, note-to-medication extraction, oncology guideline retrieval, prior authorization, and event-driven FHIR orchestration [14,15,23,71,73,78,84,88,89,93]. Most were read-only or sandboxed, but Hard to Halt demonstrated that browser agents could reliably complete forms while failing to withhold deficient submissions [23].

4.2.3. Patient-facing workflows

Four studies (14.3%) evaluated treatment planning or prescribing: two radiotherapy-planning agents, EcoRxAgent, and the daGOAT clinical trial [85,86,91,92]. One study (3.6%) was primarily patient-facing, delivering personalized nutrition coaching [90], and one (3.6%) was a cross-domain scalability experiment [78]. Patient-facing and prescribing systems carried the most direct action risk, yet only daGOAT prospectively monitored clinical outcomes and clinician deviations.

4.2.4. Agent architectures

Architectural patterns were evenly divided: 12 studies (42.9%) used single-agent or single-controller designs and 12 (42.9%) used multi-agent architectures. Three studies (10.7%) were hybrid systems coupling an LLM controller to deterministic models or clinical tools, and one proprietary agentic workflow did not fully disclose its internal agent structure. Multi-agent systems were concentrated in diagnostic support and treatment planning, whereas single agents predominated in EHR navigation, prescribing optimization, and browser automation.

Table 1. Characteristics of the 28 studies included in the current synthesis.

Ref / system	Year	Workflow	Architecture	Evaluation setting	Sample	Autonomy / oversight
[70] AgentClinic	2026	Diagnosis / CDS	Multi-agent	Benchmark / simulation	AgentClinic-MedQA, MIMIC-IV and NEJM scenarios; 20 complete dialogues rated by 3 clinic...	L1 / HOOTL
[15] Almanac Copilot	2025	EHR / information	Single	Benchmark / simulation	300 synthetic clinician-derived EHR queries based on real-patient data	L2 / HITL
[71] MedAgentBench v2	2026	EHR / information	Single	Benchmark / simulation	Original 300 tasks plus 300 new physician-designed multi-step tasks	L4 / HOOTL
[72] Agent-system benchmark	2026	Diagnosis / CDS	Hybrid	Benchmark / simulation	1,676 benchmark items/scenarios across AgentClinic, MedAgentsBench and HLE subsets	L1 / HOOTL
[14] EHRAgent	2024	EHR / information	Single	Retrospective / offline	Three datasets (MIMIC-III, eICU and TREQS), averaging 718.7 examples per dataset	L0 / HOOTL
[73] FHIR-AgentEval	2026	EHR / information	Single	Benchmark / simulation	43 modular tasks; 23 tasks create or modify server state; four experimental settings	L4 / HOOTL
[19] Oncology agent	2025	Diagnosis / CDS	Single	Benchmark / simulation	20 base cases with 15 demographic variations each (300 combinations)	L1 / Mixed
[74] CDAFlow	2026	Diagnosis / CDS	Single	Benchmark / simulation	1,500 ClinicalBench cases across 24 departments; 100 manually categorized error cases	L1 / HITL

Ref / system	Year	Workflow	Architecture	Evaluation setting	Sample	Autonomy / oversight
[75] DxDirector	2026	Diagnosis / CDS	Hybrid	Retrospective / offline	26,018 total cases, including 22,901 RareArena, 344 NEJM, 1,500 ClinicalBench, 1,273 US...	L3 / HOTL
[76] KTAS multi-agent CDSS	2025	Diagnosis / CDS	Multi-agent	Benchmark / simulation	43 randomly selected simulated emergency cases	L1 / HITL
[77] Headache orchestrator	2026	Diagnosis / CDS	Multi-agent	Benchmark / simulation	90 expert-validated secondary headache cases	L1 / HITL
[78] Clinical-scale orchestration	2026	Technical benchmark	Multi-agent	Benchmark / simulation	Batch sizes 5, 10, 20, 40 and 80; ten batches for each of four models and two topologie...	L0 / HOOTL
[79] Cholangitis NS-LLM	2026	Diagnosis / CDS	Multi-agent	Benchmark / simulation	30 ABIM-style clinical questions; 14 physician comparators from four tertiary centers	L1 / HITL
[80] Cholecystitis NS-LLM	2026	Diagnosis / CDS	Multi-agent	Benchmark / simulation	30 clinical questions and 30 physician comparators from four tertiary centers	L1 / HITL
[81] HopeAI	2025	Diagnosis / CDS	Not fully reported	Benchmark / simulation	50 clinical scenarios; 250 model responses; three blinded hematologist-oncologist evalu...	L1 / HITL
[82] Virtual oncology tumor board	2025	Diagnosis / CDS	Multi-agent	Benchmark / simulation	34 ASCO guidelines and 100 oncologist-authored question-answer pairs	L0 / HOOTL
[83] ICH agentic planning	2026	Diagnosis / CDS	Hybrid	Retrospective / offline	50 consecutive patients with spontaneous ICH	L2 / HITL
[84] RESCUE	2026	EHR / information	Multi-agent	Retrospective / offline	494,577 active patients; 175,842 screening-eligible; 12,591 holdout patients; 30-chart ...	L1 / HITL
[85] Autonomous RT platform	2026	Treatment / prescribing	Multi-agent	Retrospective / offline	60 evaluation cases: 20 brain, 20 lung and 20 prostate; separate 20-case calibration co...	L4 / HOOTL
[86] GPT-Plan	2025	Treatment / prescribing	Multi-agent	Retrospective / offline	12 non-small-cell lung IMRT cases and 5 cervical VMAT cases	L4 / HOOTL
[87] PET/CT agent	2026	Diagnosis / CDS	Single	Retrospective / offline	170 adult baseline lung-cancer PET/CT examinations	L1 / HOOTL
[88] Agent-MIRA	2026	EHR / information	Single	Retrospective / offline	Public fine-tuning cohort n=463; curated PET retrieval cohort n=179 (171 database cases...	L0 / HITL
[89] GLLaucoMed	2026	EHR / information	Single	Retrospective / offline	1,250 subjects; held-out test set included 1,984 topical-medication labels and 992 oral...	L0 / HOOTL
[90] Nutrition coaching agents	2025	Patient-facing	Multi-agent	Prospective / live	User research N=16; direct workflow validation n=6; 187 simulated vignettes, with 153 h...	L1 / HOOTL
[23] Hard to Halt	2026	EHR / information	Single	Benchmark / simulation	836 synthetic patient records spanning compliant, error-injected, irrelevant and homony...	L4 / HOOTL
[91] EcoRxAgent	2026	Treatment / prescribing	Single	Retrospective / offline	Guangzhou: 1,698 cases, 799 evaluable substitution sets; Shenshan: 1,600 screened, 760 ...	L2 / HITL
[92] daGOAT	2025	Treatment / prescribing	Single	Prospective / live	152 eligible, 114 enrolled, 110 HLA-haploidentical focus-group participants; 57 receive...	L4 / HOTL
[93] CAPABLE Case Manager	2023	EHR / information	Multi-agent	Benchmark / simulation	Approximately 20 clinical scenarios, each containing multiple longitudinal time segments	L1 / Mixed

4.3. Autonomy levels of deployed agents (RQ2)

Author terminology was inconsistent: systems labeled autonomous ranged from read-only information retrieval to live prescribing. We therefore assigned levels from observed action scope rather than labels. Five systems (17.9%) were L0, 13 (46.4%) L1, three (10.7%) L2, one (3.6%) L3, six (21.4%) L4, and none L5 (Figure 3).

L0 systems retrieved or structured information without initiating a care action. L1 systems produced a diagnosis, treatment suggestion, coaching recommendation, or referral candidate that a human had to act on. L2 systems prepared a consequential artifact or executable plan for explicit approval, as in Almanac Copilot, the intracerebral-hemorrhage planning framework, and EcoRxAgent [15,83,91]. DxDirector was the sole L3 system because it controlled bounded diagnostic steps while escalating high-risk cases [75]. The six L4 systems completed end-to-end workflows in a defined environment: MedAgentBench v2, FHIR-AgentEval, two radiotherapy planners, Hard to Halt, and daGOAT [23,71,73,85,86,92].

The L4 count should not be interpreted as evidence of clinical readiness. Five of the six L4 systems operated in a sandbox, synthetic browser, virtual FHIR server, or research treatment-planning system, where incorrect actions were reversible and could not directly harm a patient. Only daGOAT implemented L4 conditional autonomy in live care, and even there physicians could delay, modify, or discontinue the prescribed intervention [92]. Thus, capability-level autonomy substantially exceeded deployment-level evidence.

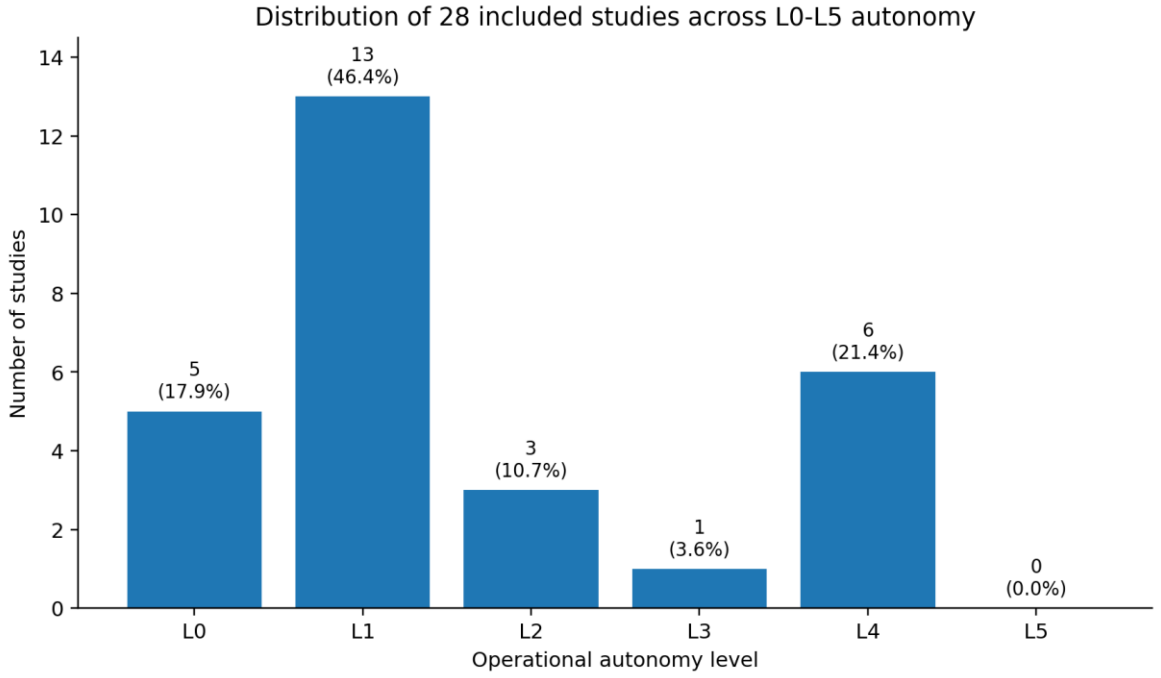

Fig. 3. Distribution of the 28 included studies across operational autonomy levels L0-L5.

4.4. 4.4 Safety outcomes and failure modes (RQ3)

4.4.1. Reported error types

Content or hallucination-related errors were reported or explicitly evaluated in 27 studies (96.4%), planning or control failures in 27 (96.4%), tool or interface failures in 19 (67.9%), and coordination or hand-off failures in 13 (46.4%). Examples included fabricated tools and invalid FHIR parameters, premature diagnosis, omitted specialist agents, inappropriate treatment objectives, browser state loss, and missing laboratory values that did not halt a live

prescribing model [15,23,71,73,77,85,86,92]. Only a minority of studies reported a formal error denominator, so these counts reflect presence of evidence rather than comparable incidence rates.

4.4.2. Adverse and near-miss events

Ten studies (35.7%) documented an enacted error or direct action in a sandbox, research system, patient-facing interaction, or live pathway. Examples included incorrect FHIR writes, invalid radiotherapy-plan modifications, deficient prior-authorization submissions, direct behavioral advice, and live ruxolitinib prescriptions [23,71,73,85,86,90,92]. Fifteen studies (53.6%) contained identifiable near-miss examples, but none used a standardized prospective near-miss definition. Only daGOAT prospectively monitored patient outcomes after an agent-generated treatment action; adverse events and two deaths occurred in the trial population, but were not established as agent-caused [92]. The corpus therefore supports absence of demonstrated agent-attributable harm, not evidence of safety.

4.4.3. Failure-mode taxonomy

Failures were organized by locus - perception/context, reasoning/planning, action interface, and coordination - and by expression - contained, enacted, and propagated. Contained errors dominated because most evaluations ended at a recommendation or sandbox output. Four studies (14.3%) explicitly quantified or experimentally isolated propagation across dependent steps: CDAFlow attributed 15% of analyzed failures to partial propagation, FHIR-AgentEval linked early tool choices to later state errors, Hard to Halt demonstrated that early retrieval and browser errors contaminated submission, and the general agent benchmark evaluated residual propagation through planner-executor-verifier workflows [23,72-74]. Many additional studies described propagation as plausible but did not measure it.

4.4.4. Mitigation strategies

Mitigation strategies included structured schemas, rule checks, constrained tool interfaces, sandboxes, independent critics, retrieval grounding, confidence thresholds, audit logs, and human review. However, most mitigations were evaluated only as part of a whole system. The strongest component-level evidence came from memory and specification ablations in FHIR-AgentEval, bounded reflection in CDAFlow, risk-supervisor testing in DxDirector, rule-plus-mirror-model checking in GPT-Plan, and physician modification of daGOAT prescriptions [73-75,86,92]. No study established that a mitigation remained effective under long-term clinical workload.

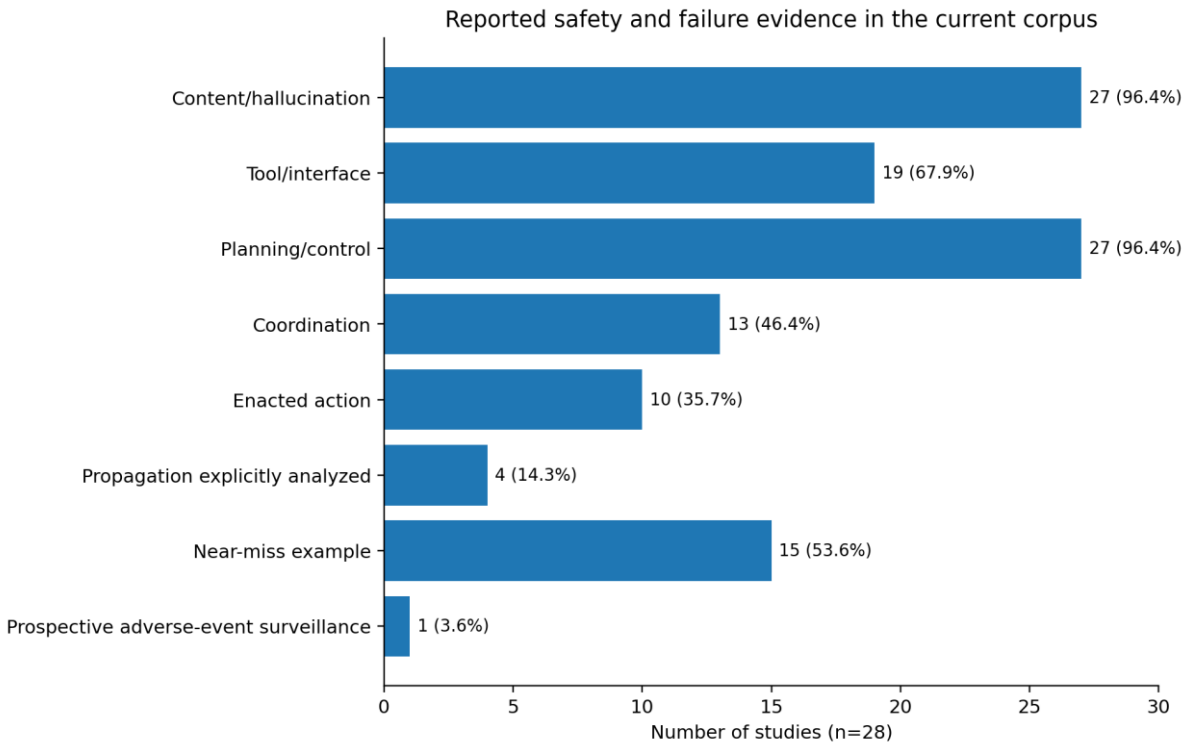

Fig. 5. Frequency of reported safety and failure evidence in the current 28-study corpus.

4.5. Human-oversight mechanisms (RQ4)

Oversight was human-in-the-loop in 11 studies (39.3%), human-on-the-loop in two (7.1%), mixed in two (7.1%), and effectively human-out-of-the-loop during evaluation in 13 (46.4%). Ten systems specified a full per-action approval gate and two a partial gate. Escalation was fully specified in three and partly specified in 10; an explicit or partial override was described in 18; and complete or partial audit logging in 26. These high mechanism counts overstate empirical maturity, because many controls existed only as architecture descriptions or sandbox capabilities.

Only three studies (10.7%) empirically characterized an operating property relevant to oversight. DxDirector measured high-risk escalation precision, recall, and F1 and quantified requested physician operations [75]. HopeAI reported that only 25.3% of outputs were ready for use without editing, providing indirect evidence of continuing review burden [81]. daGOAT reported 98% initial compliance, eight clinically motivated delays or modifications, and staff attitudes in a live pathway [92]. Oversight workload was explicitly measured in only two studies (7.1%), and no study measured the proportion of erroneous actions caught by an approval gate under sustained clinical volume. Accordingly, the central safety mechanism supporting L1-L2 deployment remains largely unvalidated.

Table 2. Safety, failure expression, and oversight evidence by study.

Ref / system	Key failure evidence	Enacted or propagated?	Safety grade	Approval	Escalation / override	Measured oversight evidence
[70] AgentClinic	Diagnostic errors; biased dialogues; tool-dependent performance degradation	Yes	Low	No	No / Not reported	Not measured; clinician ratings were retrospective evaluation, not runtime oversight.
[15] Almanac Copilot	Hallucinated facts, fabricated tools and invalid parameters/sequences	Yes	Low	Yes	Not reported / Not reported	Approval is described conceptually but its effectiveness was not experimentally measured.

Ref / system	Key failure evidence	Enacted or propagated?	Safety grade	Approval	Escalation / override	Measured oversight evidence
[71] MedAgentBench v2	Malformed requests, hallucinated actions, output-schema errors and misunderstood conditions	Yes	Moderate	No	No / Not reported	Not measured; the paper recommends physician oversight for memory and consequential actions.
[72] Agent-system benchmark	Hallucinations, low diagnostic accuracy and workflow inefficiency	No / contained	Moderate	No	No / Not reported	No human oversight mechanism was evaluated.
[14] EHRAgent	Runtime/code errors and incorrect query answers	No / contained	Low	No	No / Not reported	Not measured.
[73] FHIR-AgentEval	Incorrect tool selection/order, wrong resource type, prohibited tool and tool errors	Yes	Moderate	No	No / Partly	No runtime human oversight; deterministic validation occurs after execution.
[19] Oncology agent	Incorrect, incomplete or potentially harmful recommendations; citation errors	No / contained	Moderate	No	No / Partly	Four clinicians retrospectively evaluated outputs; manual prompting of tool use was possible but not a formal oversight me...
[74] CDAFlow	Incomplete disease diagnosis (30%), overextended differential diagnosis (30%) and redundant treatment planning (...	No / contained	Moderate	Yes	Partly / Yes	Final clinician review is required by design, but oversight effectiveness and workload were not measured.
[75] DxDirector	Misdiagnosis, inappropriate diagnostic strategies and false-positive/false-negative risk flags	No / contained	Moderate	Partly	Yes / Yes	Physician workload was quantified by requested clinical operations; high-risk detection was measured with precision 92.47%...
[76] KTAS multi-agent CDSS	Over-triage, mid-acuity misclassification and contextually unsuitable medication recommendations	No / contained	Low	Yes	Not reported / Not reported	One emergency physician retrospectively reviewed outputs; runtime oversight was not evaluated.
[77] Headache orchestrator	Missed red flags, false positives, malformed JSON, formatting errors, routing and tool-call failures	No / contained	Low	Yes	Not reported / Not reported	Two specialists established the gold standard; runtime clinician oversight was not evaluated.
[78] Clinical-scale orchestration	Accuracy collapse, invalid/no JSON, excessive token consumption and latency under context overload	No / contained	Moderate	No	No / Not reported	No clinician oversight was tested; every call could be logged and replayed.
[79] Cholangitis NS-LLM	Risk of overfitting to board-style questions, knowledge gaps outside TG18 and uncertain performance on incomplet...	No / contained	Low	Yes	Partly / Yes	Human experts served as comparators; clinical workflow oversight, usability and end-user acceptance were not tested.
[80] Cholecystitis NS-LLM	One incorrect response, partial inter-agent agreement in three cases and uncertain generalization to incomplete ...	No / contained	Low	Yes	Partly / Yes	Confidence was associated with correctness, but no runtime clinician-oversight or usability study was conducted.
[81] HopeAI	Incomplete clinical-trial interpretation, inadequate mechanism consideration, insufficient supportive-care discu...	No / contained	Moderate	Yes	No / Yes	Three blinded clinicians evaluated outputs; only 25.3% were rated ready for use without editing, demonstrating a substanti...
[82] Virtual oncology tumor board	Incorrect guideline selection in 7% and incorrect final answers in 12% of questions	No / contained	Low	No	No / Not reported	Board-certified oncologists created the reference answers; runtime clinical oversight was not studied.

Ref / system	Key failure evidence	Enacted or propagated?	Safety grade	Approval	Escalation / override	Measured oversight evidence
[83] ICH agentic planning	Nine treatment-recommendation disagreements, primarily in borderline surgical indications; possible segmentation...	No / contained	Moderate	Yes	Partly / Yes	Three blinded experts defined the reference standard; prospective effects on surgeon workload and outcomes were not measured.
[84] RESCUE	False positives/negatives, schema transfer limitations and missed evidence of prior testing	No / contained	Moderate	Yes	Partly / Yes	A blinded genetic counselor reviewed 30 charts; no prospective measurement of provider workload or clinical outcomes.
[85] Autonomous RT platform	Contradictory or clinically inappropriate objectives, cascading pipeline errors, fixed-slice image limitations a...	Yes	Moderate	No	Partly / Yes	Real-time human steering and monitoring were available but were not used in the 60 autonomous evaluation runs.
[86] GPT-Plan	Illogical/prompt-inconsistent optimization parameters and inaccurate vision-based interpretation of dose images	Yes	Moderate	No	Partly / Yes	The Human-proxy could guide optimization or select the final plan, but it was not used for all reported test patients.
[87] PET/CT agent	One DICOM/metadata-dependent quantitative failure, false-positive nodal/metastatic calls and false-negative smal...	Yes	Moderate	No	Partly / Yes	Expert clinical reports were the retrospective reference standard; the effect of live expert review on safety, reading tim...
[88] Agent-MIRA	Retrieval mismatches, incorrect majority-label predictions and uncertainty estimates affected by heterogeneous o...	No / contained	Very low	Not applicable	Partly / Yes	Two experienced nuclear-medicine clinicians rated eight test cases; no prospective measure of decision accuracy, workload ...
[89] GLLaucoMed	Invalid responses, extraction mismatches, note ambiguity and post-processing errors	No / contained	Low	No	No / Yes	Specialists created and adjudicated the reference labels; no runtime oversight or workload reduction was evaluated.
[90] Nutrition coaching agents	Barrier misclassification, incomplete or generic tactics, premature termination and unrecognized unsafe or unsui...	Yes	Low	No	No / Yes	Participants and behavioral-science experts assessed outputs after interaction; no clinician monitoring or safety-escalati...
[23] Hard to Halt	Technical aborts and undesigned withholding based on hallucinated or unsupported issues	Yes	Moderate	No	Yes / Partly	No human gate was present during execution. Manual review characterized failures, and the authors propose intermediate spe...
[91] EcoRxAgent	For 4.13% of evaluable patients all generated alternatives were clinically inferior; non-critical duplication co...	No / contained	Moderate	Yes	Not reported / Yes	Six senior physicians independently reviewed prescriptions, but actual selection behavior, workflow burden and protection ...
[92] daGOAT	Incorrect or unstable risk classification, incomplete extraction of values recorded as dashes and uncertainty in...	Yes	High	No	Yes / Yes	Initial compliance was 98%; eight participants had delayed, modified or discontinued prescriptions, and staff surveys quan...

Ref / system	Key failure evidence	Enacted or propagated?	Safety grade	Approval	Escalation / override	Measured oversight evidence
[93] CAPABLE Case Manager	Potential missing-data activation errors, rule-language limitations and responsibility transfer from infrastruct...	No / contained	Low	Partly	Partly / Not reported	No runtime oversight or human-factor effectiveness was measured.

4.6. Alignment with regulatory and ethical frameworks (RQ5)

Governance reporting lagged behind technical reporting. One study was a registered prospective phase-2 clinical trial with live conditional prescribing (daGOAT), and one platform described an EU Medical Device Regulation class IIa pathway (CAPABLE) [92,93]. Several retrospective studies reported institutional review board approval, HIPAA-compatible infrastructure, local execution, or bounded research environments, but none demonstrated regulatory authorization for an LLM agent to perform unrestricted autonomous clinical action. Across the corpus, complete audit logs were described in 13 studies, partial logs in 13, explicit overrides in 15, and workload measurement in only two. These findings align with the review's central distinction: presence of an oversight mechanism is not evidence that the mechanism is effective.

Table 3. Aggregate governance and oversight indicators in the current corpus.

Governance / oversight indicator	Studies, n (%)	Interpretation
Full per-action approval gate	10 (35.7%)	Common at L1-L2, but gate catch-rate was not measured.
Full or partial escalation rule	13 (46.4%)	Only three studies fully specified triggers.
Full or partial override function	18 (64.3%)	Usually described technically rather than evaluated in use.
Full or partial audit logging	26 (92.9%)	Complete traceability was documented in only 13 studies.
Oversight workload measured	2 (7.1%)	Reviewer burden and vigilance were largely absent from evaluation.
Direct oversight operating measure	3 (10.7%)	No study measured approval-gate catch-rate under sustained workload.
Prospective/live clinical action	1 (3.6%)	daGOAT was the sole live interventional example.

4.7. Evaluation methodologies and benchmarks (RQ6)

The evidence base remained laboratory-heavy. Sixteen studies (57.1%) were benchmark, simulation, or sandbox evaluations; 10 (35.7%) were retrospective or offline real-data studies; and two (7.1%) included prospective human or live evaluation. Only one study (3.6%) tested an agent-generated treatment action in live care. Outcome metrics were dominated by task accuracy, completion, or clinician-rated output quality; workflow cost and latency were occasionally measured, but patient outcomes, near-miss surveillance, reviewer workload, and post-deployment drift were rare. The autonomy-by-evidence map showed one high-grade study, 15 moderate, 11 low, and one very low, with no L5 evidence and only one high-grade L4 example (Figure 4).

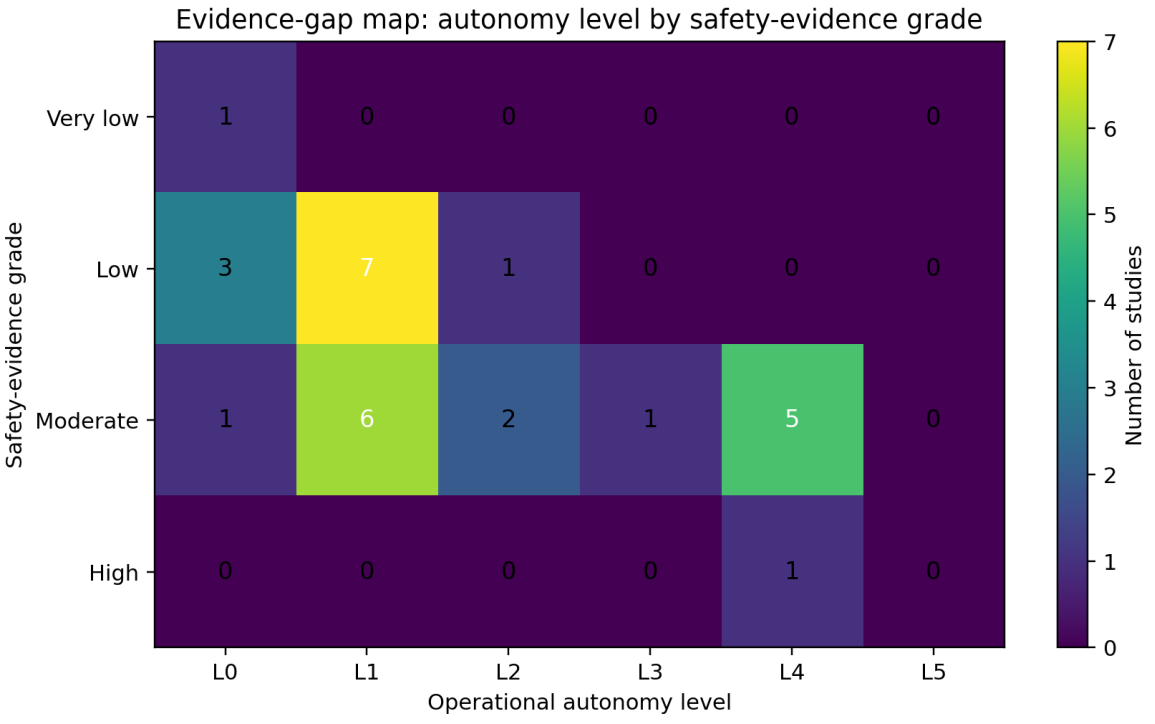

Fig. 4. Evidence-gap map crossing operational autonomy level with safety-evidence grade.

5. DISCUSSION

5.1. Principal findings

This 28-study synthesis yields four principal findings. First, operational agency is now technically credible across diagnosis, EHR interaction, treatment planning, prescribing, and administrative workflows, but most evidence remains in benchmarks, sandboxes, or retrospective environments. Second, observed authority was concentrated at L1, while six systems demonstrated L4-like end-to-end behavior only within tightly bounded domains. Third, content, planning, and tool failures were common, yet explicit measurement of propagation across dependent steps was rare. Fourth, oversight was extensively described but minimally tested: only three studies quantified an operating property of oversight, and no study measured approval-gate error-catching under realistic sustained workload. The field therefore has more evidence that agents can act than that humans can reliably supervise their actions.

5.2. The autonomy-oversight-risk matrix

The autonomy-oversight-risk matrix supports evidence-gated promotion rather than fixed categorical permission. For low-risk, reversible information tasks, L3-L4 operation may be reasonable once failure rates and audit detection are characterized. For intermediate-risk actions such as patient messaging, administrative submission, or plan modification, autonomy above L2 requires validated escalation and override performance. For high-risk actions such as medication prescribing or treatment delivery, the current corpus supports only bounded conditional autonomy with active clinical monitoring. The contrast between Hard to Halt and daGOAT is instructive: both reached L4 operationally, but one exposed completion bias in a synthetic prior-authorization portal [23], whereas the other embedded physician override and outcome monitoring in a live trial [92].

Mapping the six L4 systems onto this matrix revealed a recurrent evidence gap. Five operated in reversible research environments, yet their headline descriptions could be read as clinical autonomy. The one live L4 study retained physician authority and generated several clinically motivated deviations, showing that a well-designed

human-on-the-loop configuration is not passive: it must detect context the model missed and modify action promptly. Autonomy should therefore be attributed to a specific workflow and operating domain, not to a model or product in the abstract.

5.3. Sociotechnical considerations

Designed oversight can fail sociotechnically even when it is technically present. Per-action approval may degenerate into routine acceptance when volume is high, while human-on-the-loop monitoring may fail if escalation criteria are opaque or alarms are too frequent. The corpus offered almost no evidence on vigilance decay, edit burden, skill retention, or responsibility allocation. HopeAI's finding that only one quarter of outputs were ready without editing [81], EcoRxAgent's residual inferior alternatives [91], and daGOAT's physician modifications [92] show that human review adds real clinical value; they do not establish that the same value would persist at scale.

5.4. Comparison with other safety-critical domains

The comparison with aviation, autonomous driving, and surgical robotics remains useful, but the present findings sharpen the analogy. Healthcare agents need declared operational design domains, action-class permissions, event logs, and required disengagement or override reporting. The current evidence base rarely reports the equivalent of a disengagement rate. A standardized report containing number of agent actions, escalations, overrides, caught errors, enacted errors, and near misses would directly address the most important gap identified here.

5.5. Implications

5.5.1. For developers

Developers should declare the operational autonomy level, action scope, reversibility, and operational design domain as first-class specifications. Agent evaluations should include perturbations that expose missing data, conflicting records, tool failure, and conditional abstention. Each safety mechanism should have an operating characteristic: approval-gate catch rate, escalation sensitivity, override success, and audit detection probability. Systems should not be described as autonomous solely because they plan or call tools.

5.5.2. For hospitals and clinicians

Hospitals should deploy new agents initially behind explicit gates and instrument the review process rather than assuming that nominal clinician approval is protective. Local evaluation should measure review time, edit rate, caught-error rate, override rate, and vigilance across shifts. Promotion to supervised or conditional autonomy should require locally observed evidence, and incident reporting should include agent-related near misses from the first day of use.

5.5.3. For regulators and policymakers

Regulators and policymakers should tie oversight obligations to both autonomy level and worst-case task risk. Movement from L2 gated execution to L3 or L4 operation should be treated as an evidence event requiring performance data from the oversight layer, not merely from the base model. Action-taking administrative systems also require scrutiny because erroneous submissions, messages, or records may propagate into clinical care even when the product is marketed as non-clinical.

5.6. Limitations

This synthesis has important limitations. Most importantly, it is based on a priority subset rather than completed title/abstract screening of all 2,557 unique records, and one AI-assisted reviewer performed the current full-text decisions and extraction. The 28 included studies may therefore over-represent prominent, accessible, or high-signal agent studies, and the interim counts cannot be treated as definitive prevalence estimates. Independent Reviewer 2 screening, consensus, inter-rater reliability, full-corpus screening, citation chasing, and formal design-matched risk-of-bias appraisal remain pending. The electronic search was limited to English-language records in PubMed, Scopus,

and Web of Science, and omission of Embase and engineering-specific databases may miss studies. The field is evolving rapidly, and several 2026 studies were preprints. Finally, heterogeneous tasks and outcomes precluded meta-analysis. Accordingly, this manuscript is scientifically complete as a working synthesis of the current corpus but should not be submitted as a final systematic review until the remaining review procedures are completed.

6. FUTURE RESEARCH AGENDA

The gaps documented in this review translate into eight concrete research priorities, ordered roughly by the dependency structure among them: measurement precedes evidence, and evidence precedes governance.

1. Standardized safety benchmarks for multi-step clinical agents. Current benchmarks reward task completion; the field needs shared evaluation environments that score enacted and propagated failures - injecting realistic perturbations (missing data, ambiguous orders, conflicting records) and measuring error containment across steps, with validation against real deployment performance.

2. Prospective trials with pre-registered safety endpoints. Agentic systems should graduate to trials in which near-misses, escalations, overrides, and harm events are primary or co-primary pre-registered endpoints - not post-hoc observations - with reporting aligned to CONSORT-AI and DECIDE-AI extensions adapted for action-taking systems [3,64].

3. Empirical validation of oversight effectiveness. The approval gate must be studied as an intervention in its own right: what proportion of erroneous agent actions do reviewers actually catch, how does catch-rate decay with volume, reliability, and shift length, and which interface designs sustain vigilance? Analogous operating characteristics are needed for escalation triggers (sensitivity/specificity) and retrospective audit (detection probability).

4. Certification methods and graded-promotion criteria. Research should define auditable evidence packages for autonomy promotion - what a system must demonstrate at L2 to be permitted L3 operation within a stated design domain - including statistical criteria for sufficient exposure, error bounds, and post-promotion monitoring, translating the evidence-gated promotion principle of Section 5.2 into certifiable procedure.

5. Longitudinal studies of automation bias and deskilling. Cross-sectional edit-rate snapshots cannot reveal how reviewer scrutiny and clinician competence evolve over months of agent exposure; cohort designs tracking vigilance, error-detection ability, and independent task performance are required to know whether oversight capacity erodes precisely as reliance grows.

6. Incident surveillance infrastructure. A confidential, blame-protected reporting system for agent-related near-misses and adverse events - the healthcare analogue of aviation safety reporting and autonomous-vehicle disengagement reports [68,69] - should be piloted, with a shared taxonomy (Section 4.4.3) enabling aggregation across institutions and vendors.

7. Governance models for multi-agent and distributed-responsibility systems. When a decision emerges from several interacting agents and an orchestrator, existing accountability constructs strain; research spanning law, ethics, and system engineering should develop attribution methods (which component contributed what to an outcome) and corresponding liability and audit frameworks.

8. Equity and generalizability of agentic deployment. Agent performance, escalation behavior, and oversight burden should be characterized across languages, health-system resource levels, and patient populations, before deployment patterns harden around the well-resourced settings that dominate the current evidence base.

7. CONCLUSION

Healthcare AI agents can now retrieve data, coordinate tools, prepare executable actions, and complete bounded workflows, but evidence supporting their safe supervision remains much thinner than evidence of capability. In the current 28-study corpus, L1 recommendation systems predominated, six systems reached L4 within constrained environments, and only one implemented live conditional prescribing with prospective outcome monitoring.

Content, planning, tool, and coordination failures were widely reported, while propagated failures and oversight effectiveness were rarely measured. The practical standard should therefore be evidence-gated autonomy: each increase in action authority must be earned through demonstrated containment of multi-step failure, validated escalation and override behavior, and measurable human control within a declared operational design domain.

Declaration

We acknowledge that we used ChatGPT to enhance the academic writing of our manuscript while ensuring the originality and integrity of our work.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[2] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine, 28*(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

[3] Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

[4] van de Sande, D., et al. (2022). Developing, implementing and governing artificial intelligence in medicine: A step-by-step approach to prevent an artificial intelligence winter. *BMJ Health Care Informatics, 29*(1), e100495. https://doi.org/10.1136/bmjhci-2021-100495

[5] Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[6] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on medical challenge problems* [Preprint]. arXiv. https://arxiv.org/abs/2303.13375

[7] Moor, M., Banerjee, O., Abad, Z. S. H., et al. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616*(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

[8] Zhao, L., Liu, S., Xin, T., et al. (2026). AI agent in healthcare: Applications, evaluations, and future directions. *npj Artificial Intelligence, 2*, 31. https://doi.org/10.1038/s44387-026-00076-4

[9] Xu, X., & Sankar, R. (2025). Large language model agents for biomedicine: A comprehensive review of methods, evaluations, challenges, and future directions. *Information, 16*(10), 894. https://doi.org/10.3390/info16100894

[10] Masterman, T., Besen, S., Sawtell, M., & Chao, A. (2024). *The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey* [Preprint]. arXiv. https://arxiv.org/abs/2404.11584

[11] Tierney, A. A., Gayre, G., Hoberman, B., et al. (2025). Ambient artificial intelligence scribes: Learnings after 1 year and over 2.5 million uses. *NEJM Catalyst Innovations in Care Delivery, 6*(5), CAT.25.0040. https://doi.org/10.1056/CAT.25.0040

[12] Lukac, P. J., Turner, W., Vangala, S., et al. (2025). Ambient AI scribes in clinical practice: A randomized trial. *NEJM AI, 2*(12), AIoa2501000. https://doi.org/10.1056/AIoa2501000

[13] Olson, K. D., Meeker, D., Troup, M., et al. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open, 8*(10), e2534976. https://doi.org/10.1001/jamanetworkopen.2025.34976

[14] Shi, W., et al. (2024). EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of EMNLP 2024* (pp. 22315–22339). https://doi.org/10.18653/v1/2024.emnlp-main.1245

[15] Zakka, C., Cho, J., Fahed, G., et al. (2025). *Almanac Copilot: Towards autonomous electronic health record navigation* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-6102516/v1

[16] Gebreab, S. A., et al. (2024). LLM-based framework for administrative task automation in healthcare. In *Proceedings of the 12th International Symposium on Digital Forensics and Security (ISDFS)* (pp. 1–7). IEEE. https://doi.org/10.1109/ISDFS60797.2024.10527275

[17] Pandey, H. G., Amod, A., & Kumar, S. (2024). Advancing healthcare automation: Multi-agent system for medical necessity justification. *Proceedings of the BioNLP Workshop*. https://doi.org/10.18653/v1/2024.bionlp-1.4

[18] Tu, T., Palepu, A., Schaekermann, M., et al. (2025). Towards conversational diagnostic artificial intelligence. *Nature*. https://doi.org/10.1038/s41586-025-08866-7

[19] Ferber, D., El Nahhas, O. S. M., Wölflein, G., et al. (2025). Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer, 6*(8), 1337–1349. https://doi.org/10.1038/s43018-025-00991-6

[20] Kim, Y., et al. (2024). MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.52202/079017-2522

[21] Wang, Z., et al. (2025). ColaCare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM Web Conference 2025* (pp. 2250–2261). https://doi.org/10.1145/3696410.3714877

[22] Li, J., et al. (2024). *Agent Hospital: A simulacrum of hospital with evolvable medical agents* [Preprint]. arXiv. https://arxiv.org/abs/2405.02957

[23] Nie, M., Chung, W., Waxler, J., et al. (2026). Hard to halt: Automation bias in agent-driven sequencing prior authorization workflows [Preprint]. *medRxiv*. https://doi.org/10.64898/2026.06.16.26355782

[24] Yang, G. Z., Cambias, J., Cleary, K., et al. (2017). Medical robotics—Regulatory, ethical, and legal

considerations for increasing levels of autonomy. *Science Robotics, 2*(4), eaam8638. https://doi.org/10.1126/scirobotics.aam8638

[25] European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union*.

[26] U.S. Food and Drug Administration. (2025). *Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations—Draft guidance*. U.S. Food and Drug Administration.

[27] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). *Concrete problems in AI safety* [Preprint]. arXiv. https://arxiv.org/abs/1606.06565

[28] Cemri, M., et al. (2025). *Why do multi-agent LLM systems fail?* [Preprint]. arXiv. https://doi.org/10.52202/085713-4082

[29] Jiang, Y., Black, K. C., Geng, G., et al. (2025). MedAgentBench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI, 2*(9), AIdbp2500144. https://doi.org/10.1056/AIdbp2500144

[30] Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), e35179. https://doi.org/10.7759/cureus.35179

[31] Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

[32] Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

[33] Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., & Kaushal, R. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making, 17*(1), 36. https://doi.org/10.1186/s12911-017-0430-8

[34] Guo, Y., Hu, D., Zhou, Y., et al. (2026). From conversation to chart: An analysis of clinician edits to ambient AI draft notes [Preprint]. *medRxiv*. https://doi.org/10.64898/2026.01.05.26343471

[35] Price, W. N., II, Gerke, S., & Cohen, I. G. (2019). Potential liability for physicians using artificial intelligence. *JAMA, 322*(18), 1765–1766. https://doi.org/10.1001/jama.2019.15064

[36] Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: Accountability and safety. *Bulletin of the World Health Organization, 98*(4), 251–256. https://doi.org/10.2471/BLT.19.237487

[37] Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295–336). Academic Press. https://doi.org/10.1016/B978-0-12-818438-7.00012-5

[38] World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. World Health Organization.

[39] Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E., & Wicks, P. (2023). Large language model AI chatbots require approval as medical devices. *Nature Medicine, 29*(10), 2396–2398. https://doi.org/10.1038/s41591-023-02412-6

[40] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large

language models in medicine. *Nature Medicine, 29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

[41] Collaco, B. G., Haider, S. A., Prabha, S., et al. (2026). The role of agentic artificial intelligence in healthcare: A scoping review. *npj Digital Medicine, 9*, 345. https://doi.org/10.1038/s41746-026-02517-5

[42] Njei, B., Al-Ajlouni, Y. A., Kanmounye, U. S., Boateng, S., Nguefang, G. L., Njei, N., et al. (2026). Artificial intelligence agents in healthcare research: A scoping review. *PLOS ONE, 21*(2), e0342182. https://doi.org/10.1371/journal.pone.0342182

[43] Gorenshtein, A., Omar, M., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2025). AI agents in clinical medicine: A systematic review [Preprint]. *medRxiv*. https://doi.org/10.1101/2025.08.22.25334232

[44] Xu, J., Ko, J. M., & Kvedar, J. C. (2026). AI agents in clinical practice: An evidence map. *npj Digital Medicine*. https://doi.org/10.1038/s41746-026-02960-4

[45] SAE International. (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (SAE Standard J3016_202104).

[46] Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115–152. https://doi.org/10.1017/S0269888900008122

[47] Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

[48] Xi, Z., Chen, W., Guo, X., et al. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences, 68*(2), 121101. https://doi.org/10.1007/s11432-024-4222-0

[49] Wang, L., Ma, C., Feng, X., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

[50] Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. MIT Man-Machine Systems Laboratory. https://doi.org/10.21236/ADA057655

[51] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30*(3), 286–297. https://doi.org/10.1109/3468.844354

[52] Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors, 59*(1), 5–27. https://doi.org/10.1177/0018720816681350

[53] Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 15. https://doi.org/10.3389/frobt.2018.00015

[54] Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

[55] Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., et al., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA Statement for reporting literature searches in systematic reviews. *Systematic Reviews, 10*(1), 39. https://doi.org/10.1186/s13643-020-01542-z

[56] McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS peer review of electronic search strategies: 2015 guideline statement. *Journal of Clinical Epidemiology, 75*, 40–46. https://doi.org/10.1016/j.jclinepi.2016.01.021

[57] Bramer, W. M., Giustini, D., de Jonge, G. B., Holland, L., & Bekhuis, T. (2016). De-duplication of database search results for systematic reviews in EndNote. *Journal of the Medical Library Association, 104*(3), 240–243. https://doi.org/10.3163/1536-5050.104.3.014

[58] Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews, 5*, 210. https://doi.org/10.1186/s13643-016-0384-4

[59] McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

[60] Sterne, J. A. C., Savovic, J., Page, M. J., et al. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ, 366*, l4898. https://doi.org/10.1136/bmj.l4898

[61] Sounderajah, V., Ashrafian, H., Rose, S., et al. (2021). A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nature Medicine, 27*(10), 1663–1665. https://doi.org/10.1038/s41591-021-01517-0

[62] Wolff, R. F., Moons, K. G. M., Riley, R. D., et al. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51–58. https://doi.org/10.7326/M18-1376

[63] Gallifant, J., Afshar, M., Ameen, S., et al. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine, 31*(1), 60–69. https://doi.org/10.1038/s41591-024-03425-5

[64] Vasey, B., Nagendran, M., Campbell, B., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine, 28*(5), 924–933. https://doi.org/10.1038/s41591-022-01772-9

[65] Campbell, M., McKenzie, J. E., Sowden, A., et al. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ, 368*, l6890. https://doi.org/10.1136/bmj.l6890

[66] Miake-Lye, I. M., Hempel, S., Shanman, R., & Shekelle, P. G. (2016). What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. *Systematic Reviews, 5*, 28. https://doi.org/10.1186/s13643-016-0204-x

[67] Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA, 318*(6), 517–518. https://doi.org/10.1001/jama.2017.7797

[68] Helmreich, R. L. (2000). On error management: Lessons from aviation. *BMJ, 320*(7237), 781–785. https://doi.org/10.1136/bmj.320.7237.781

[69] Barach, P., & Small, S. D. (2000). Reporting and preventing medical mishaps: Lessons from non-medical near miss reporting systems. *BMJ, 320*(7237), 759–763. https://doi.org/10.1136/bmj.320.7237.759

[70] Schmidgall, S., Ziaei, R., Harris, C., et al. (2026). AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*. https://doi.org/10.1038/s41746-026-02674-7

[71] Chen, E., Postelnik, S., Black, K., et al. (2026). MedAgentBench v2: Improving medical LLM agent design. In *Pacific Symposium on Biocomputing*. https://doi.org/10.1142/9789819824755_0025

[72] Liu, Y., Carrero, Z. I., Jiang, X., et al. (2026). Benchmarking large language model-based agent systems for clinical decision tasks. *npj Digital Medicine*. https://doi.org/10.1038/s41746-026-02443-6

[73] Mokssit, Y., Ravi, K., Nie, M., et al. (2026). FHIR-AgentEval: A modular sandbox for benchmarking clinical

LLM agents with an evaluation of memory-augmented configurations [Preprint]. *Research Square*. https://doi.org/10.21203/rs.3.rs-8746188/v1

[74] Hou, R., Xue, D., Sun, H., et al. (2026). CDAFlow: Enhancing LLM clinical decision-making through agentic workflow. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2026.131806

[75] Xu, S., Huang, X., Wei, Z., et al. (2026). DxDirector: An agentic large language model driving the full-process clinical diagnosis. *Nature Communications*. https://doi.org/10.1038/s41467-026-71928-5

[76] Han, S., & Choi, W. (2025). 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. *Advances in Artificial Intelligence and Machine Learning*. https://doi.org/10.54364/AAIML.2025.51187

[77] Wu, X., Zhang, H., Garduno-Rapp, N. E., et al. (2026). Orchestrator multi-agent clinical decision support system for secondary headache diagnosis in primary care. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocag111

[78] Klang, E., Omar, M., Raut, G., et al. (2026). Orchestrated multi agents sustain accuracy under clinical-scale workloads compared to a single agent. *npj Health Systems*. https://doi.org/10.1038/s44401-026-00077-0

[79] Ucdal, M., & Ekingen, E. (2026). Performance comparison of a neuro-symbolic large language model system versus conventional AI models and human experts in cholangitis management. *BMC Medical Informatics and Decision Making*. https://doi.org/10.1186/s12911-026-03593-z

[80] Ekingen, E., & Ucdal, M. (2026). Performance comparison of a neuro-symbolic large language model system versus human experts in acute cholecystitis management. *Journal of Clinical Medicine, 15*(5), 1730. https://doi.org/10.3390/jcm15051730

[81] Zhai, G., Bar, M., Cowan, A. J., et al. (2025). AI for evidence-based treatment recommendation in oncology: A blinded evaluation of large language models and agentic workflows. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2025.1683322

[82] Wang, J., Mullick Chowdhury, S., & Nazha, A. (2025). Virtual oncology collaborative tumor board using multiple artificial intelligence agents. *Journal of Clinical Oncology, 43*(16_suppl), 1563. https://doi.org/10.1200/JCO.2025.43.16_suppl.1563

[83] Kochuiev, E., Kaliuzhka, V., Markevych, M., et al. (2026). An agentic AI framework for integrated decision support and surgical planning in intracerebral hemorrhage. *Acta Neurochirurgica*. https://doi.org/10.1007/s00701-026-06954-9

[84] Liu, C., Geltzeiler, A., Afyouni, A., et al. (2026). RESCUE: An end-to-end multi-agent LLM system for proactive rare-disease patient screening in the EHR [Preprint]. *medRxiv*. https://doi.org/10.64898/2026.06.24.26356357

[85] Maniscalco, M. A., Park, Y. K., Domal, S. J., et al. (2026). Autonomous radiotherapy planning via agentic orchestration using a multimodal TPS-integrated compound AI platform. *Machine Learning: Health*. https://doi.org/10.1088/3049-477X/ae7978

[86] Wang, Q., Wang, Z., Li, M., et al. (2025). A feasibility study of automating radiotherapy planning with large language model agents. *Physics in Medicine & Biology*. https://doi.org/10.1088/1361-6560/adbff1

[87] Choi, H., Bae, S., & Na, K. J. (2026). End-to-end PET/CT interpretation and quantification with an LLM-orchestrated AI agent: A real-world pilot study. *Journal of Nuclear Medicine*.

https://doi.org/10.2967/jnumed.126.272362

[88] Vashistha, R., Brosda, S., Aoude, L. G., et al. (2026). Agent-MIRA: AI-orchestrated medical imaging agent for PET image retrieval and assistance. *Computerized Medical Imaging and Graphics*. https://doi.org/10.1016/j.compmedimag.2026.102725

[89] Solages, N., Scherer, R., Samico, G. A., et al. (2026). GLLaucoMed: A secure LLM-powered agentic workflow for automated medication extraction from free-text glaucoma clinical notes [Preprint]. *medRxiv*. https://doi.org/10.64898/2026.06.12.26355525

[90] Yang, E., Garcia, T., Williams, H. G., et al. (2025). A behavioral science-informed agentic workflow for personalized nutrition coaching: Development and validation study. *JMIR Formative Research*. https://doi.org/10.2196/75421

[91] Li, C., Lai, P., Zhang, N., et al. (2026). EcoRxAgent: An AI agent for generating economically substitutable prescriptions. *npj Digital Medicine*. https://doi.org/10.1038/s41746-026-02612-7

[92] Chen, J., Cao, Y., Feng, Y., et al. (2025). Autonomous artificial intelligence prescribing a drug to prevent severe acute graft-versus-host disease in HLA-haploidentical transplants. *Nature Communications*. https://doi.org/10.1038/s41467-025-62926-0

[93] Lanzola, G., Polce, F., & Parimbelli, E. (2023). The Case Manager: An agent controlling the activation of knowledge sources in a FHIR-based distributed reasoning environment. *Applied Clinical Informatics*. https://doi.org/10.1055/a-2113-4443