Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# AI Agents and Agentic AI in Medicine: An Autonomy-Centred Scoping Review and Research Agenda

**Valentina De Nicolò**

valentina.denicolo@unimi.it

University of Milan

**Davide La Torre**

Université Côte d'Azur

**Ignazio Iacone**

LUM University "Giuseppe Degennaro"

**Research Article**

**Keywords:** AI agents, agentic AI, large language models, multi-agent systems, clinical agent autonomy, clinical decision support

**Posted Date:** November 11th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10652827/v1

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

**Additional Declarations:** No competing interests reported.

# AI Agents and Agentic AI in Medicine:

# An Autonomy-Centred Scoping Review and Research Agenda

**Valentina De Nicolò[1] Davide La Torre[2,3] Ignazio Iacone[4]**

[1]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy

[2]SKEMA Business School, Université Côte d'Azur, Sophia Antipolis, France

[3]Department of Mathematics, University of Bologna, Bologna, Italy

[4]Department of Medicine and Surgery, LUM University "Giuseppe Degennaro", Casamassima (BA), Italy

**Corresponding author:** Valentina De Nicolò, valentina.denicolo@unimi.it

ORCID: Valentina De Nicolò https://orcid.org/0000-0002-2599-4430

## Abstract

Artificial intelligence agents—software systems that perceive, reason, plan and act over multiple steps with limited human intervention—are moving from research prototypes towards clinical deployment. Recent reviews map what has been built; this one asks what would have to hold for such systems to be deployed safely.

We conducted a scoping review reported according to the PRISMA extension for scoping reviews, searching eight bibliographic sources from January 2017 to June 2026, admitting foundational computer science and regulatory literature alongside clinical applications.

Five architectural paradigms recur—tool-augmented large language model agents, multi-agent systems, offline reinforcement learning agents, embodied robotic agents and autonomous monitoring agents—nearly all sharing a common four-layer stack. We introduce a six-level scale of clinical agent autonomy that grades systems by where human judgement enters relative to the reasoning, and by whether the resulting action remains reversible. The scale exposes an inversion in the evidence base: demonstrated patient benefit concentrates at high autonomy in narrow, bounded domains, whereas research effort concentrates at intermediate autonomy, where prospective outcome evidence is absent and automation bias is the dominant risk. We further propose seven evaluation dimensions specific to agency, a mapping of evaluation requirements to autonomy level, and a ten-item failure mode taxonomy assigned to the architectural layer at which each is addressable.

Deploying agentic systems safely requires prospective multi-site validation, reporting standards adapted to agency, equity-aware design, and regulatory obligations indexed to autonomy rather than algorithm class. We conclude with an eight-point research agenda.

**Keywords:** AI agents; agentic AI; large language models; multi-agent systems; clinical agent autonomy; clinical decision support

# 1 Introduction

The practice of medicine has long benefited from computational assistance, from the statistical decision-support tools of the 1970s to the machine-learning diagnostic aids of the 2010s. Bibliometric evidence documents sustained growth in artificial intelligence (AI) and machine learning (ML) research applied to medicine over the past decade (De Nicolò and La Torre,

2024a), and successive overviews of the field describe a transition from narrow, task-specific models towards general-purpose systems (Topol, 2019; Rajpurkar et al., 2022). Among the most consequential recent developments is the emergence of *agentic AI*: systems that do not merely respond to a single query but perceive, reason, plan, act and iterate over multiple steps in pursuit of a goal.

An *AI agent* is a software entity that maintains an internal state, selects actions from a repertoire—which may include invoking external tools, writing code, querying databases or instructing robotic actuators—and executes those actions to achieve one or more objectives (Russell and Norvig, 2020). When such agents are built on foundation models and coordinated into networks, they become capable of multi-step clinical reasoning, longitudinal patient management and autonomous hypothesis generation in biomedical research. Gao et al. (2024) articulate this vision for biomedical discovery, describing agents that integrate machine learning tools, biological knowledge and experimental platforms within structured workflows.

The clinical stakes are correspondingly high. Reviews of ML for neurological and psychiatric disease report considerable promise in early diagnosis and risk stratification alongside persisting difficulties with data quality, privacy and the ethical implications of algorithmic decisions (Ricetti et al., 2025). Extending such systems from passive prediction to autonomous, iterative action amplifies both the potential benefit and the potential harm: an agent that retrieves data, calls a model, applies a rule and issues a recommendation compresses several points of human review into a single automated pathway.

## 1.1 Relation to existing reviews

The review literature on this topic expanded rapidly during 2025 and 2026, and we begin by stating plainly what it now covers, because the contribution of the present work depends on the distinction.

General surveys of agents built on large language models (LLMs) (Xi et al., 2023; Wang et al., 2024) give thorough treatments of architecture, memory and planning but are domain-agnostic and do not engage with clinical evidence, regulation or patient safety. Within medicine, coverage is no longer scarce. Xu et al. (2026) synthesise more than two hundred studies into a taxonomy of healthcare agents organised by clinical function. Three scoping reviews have appeared in close succession. Collaco et al. (2026a) applied strict eligibility criteria and identified only seven qualifying systems, across emergency medicine, oncology, radiology and rehabilitation; most were exploratory, and only one involved patients. Njei et al. (2026) used broader criteria over a longer window and included forty-three studies, the large majority published in a single year. The same group scoped the narrower question of explainability in agentic healthcare systems (Collaco et al., 2026b). Karunanayake (2026) argues from the deployment side that commercial agentic systems entered clinical settings during 2026 faster than the evidence base could follow.

The variation among these reviews is itself informative. Counts of the eligible literature range from seven to over two hundred depending on how *agent* is defined, which is a symptom of the conceptual imprecision that Collaco et al. (2026a) identify explicitly. It also means that a further inventory of the same literature would add little.

Accordingly, this review does not claim to extend coverage. Its contribution is analytical. None of the works above proposes a scale by which the autonomy of a clinical reasoning agent can be

graded, none indexes evaluation requirements to such a scale, and none identifies the systematic relationship between autonomy, scope and strength of evidence that Section 11 sets out. The one established autonomy framework in medicine (Yang et al., 2017) was developed for robotics and grades the division of physical labour between surgeon and machine; Section 4.2 explains why reasoning agents require a distinct scale and how the two relate. Where the existing reviews describe what has been built, this one asks what would have to be true for it to be deployed safely, and derives an agenda from the answer.

Table 1 positions this review against these works.

**Table 1** Positioning of this review relative to existing reviews and frameworks in the medical agentic AI literature. Y = addressed substantively; ~ = partial; N = not addressed. Sources ordered by publication date. n denotes the number of sources included where the work reports one

| | Yang et al. (2017) | Xi et al. (2023) | Mehandru et al. (2024) | Xu et al. (2026) | Collaco et al. (2026a) | Njei et al. (2026) |
|---|---|---|---|---|---|---|
| | | | | $n > 200$ | $n = 7$ | $n = 43$ |
| Medical domain focus | Y | N | Y | Y | Y | Y |
| Systematic, reported method | N | N | N | N | Y | Y |
| Agent theory and planning | N | Y | N | Y | ~ | ~ |
| Taxonomy of agent types | N | Y | N | Y | ~ | ~ |
| Autonomy scale, physical execution | Y | N | N | N | N | N |
| Autonomy scale, reasoning agents | N | N | N | N | N | N |
| Evaluation requirements by autonomy | N | N | N | N | N | N |
| Agent-specific evaluation criteria | N | N | ~ | N | N | N |
| Failure mode taxonomy by stack layer | N | N | N | N | N | N |
| Safety, ethics and regulation | Y | ~ | ~ | ~ | ~ | ~ |
| Structured research agenda | ~ | Y | N | Y | ~ | ~ |

The present review addresses every row. Its distinctive contribution lies in the four rows for which no existing work is marked: the autonomy scale for reasoning agents, the indexing of evaluation requirements to that scale, the agent-specific evaluation criteria, and the failure mode

taxonomy assigned to architectural layer. Coverage of the primary literature is deliberately not claimed as a contribution, since Xu et al. (2026) and Njei et al. (2026) already provide it.

## 1.2 Contributions

This review makes four analytical contributions and one methodological one.

**C1. An autonomy scale for clinical reasoning agents.** We introduce a six-level scale (Section 4.2) that distinguishes systems by the point at which human judgement is exercised, and by whether the resulting action is still reversible, rather than by the underlying algorithm. It complements rather than duplicates the established autonomy framework for medical robotics (Yang et al., 2017), which grades the division of physical labour; the two are shown to be orthogonal. The scale organises the applications reviewed here and aligns with the regulatory distinctions of Section 9.

**C2. Seven evaluation dimensions specific to agency** (Section 6), motivated by the finding that the prevailing benchmark landscape measures knowledge recall rather than the capacity to act.

**C3. A ten-item failure mode taxonomy** with severity ratings and mitigations mapped to the layer of the agentic stack at which each can be addressed (Section 7).

**C4. An eight-point research agenda** ordered by urgency and tractability, including a concrete specification for the agent-specific reporting extension we argue the field most needs (Section 10).

**C5. A transparently reported evidence base.** The search strategy, eligibility criteria, screening procedure and charting variables are reported according to PRISMA-ScR, with the

completed checklist and the complete database-specific search strings provided as Online Resources, so that the method can be reproduced. We emphasise that this is a condition of the argument rather than a contribution in itself: inventory of the primary literature is already available from Xu et al. (2026) and Njei et al. (2026), and we build on it rather than duplicating it.

The organising claim, developed in Section 11, is that autonomy and evidence stand in an inverse relation across the field: systems with demonstrated patient benefit are those granted the most autonomy in the narrowest domains, while research effort concentrates where autonomy is intermediate, scope is broad, and prospective outcome evidence is absent. The existing reviews record the constituent facts; the relationship between them, and its implications for how these systems should be evaluated and regulated, has not previously been articulated.

## 1.3 Article type and intended readership

This article is a critical survey rather than a descriptive one: it evaluates the techniques, architectures and evaluation practices of the field against the demands of the deployment setting, and reaches a position on where the field's effort is well and badly directed. Sections 3 and 4 additionally serve a tutorial function, giving researchers entering this area from artificial intelligence a compact account of the algorithmic components—planning, memory, tool use, multi-agent coordination, offline reinforcement learning—together with the clinical constraints that make each of them harder here than in the benchmark settings where they were developed.

We have written for two audiences within the AI community. For researchers building medical agents, the operational content is the autonomy scale and its coding rules (Section 4.2), the evaluation dimensions (Section 6) and the failure mode taxonomy with its mapping to

architectural layers (Section 7); each is designed to be applied to a system under development rather than merely read. For those evaluating such systems, the reporting specification of Section 10.2 sets out what a study of an agentic system should state in order to be interpretable.

Two threads connect the material to cognitive science. The first is the relationship between an agent's reasoning trace and a human supervisor's capacity to evaluate it, which we argue in Section 7 is the binding constraint on intermediate autonomy and is a problem of attention and workload before it is a problem of interpretability. The second is trust calibration: the conditions under which practitioners come to rely on, or disregard, an automated recommendation, a question with an established literature in human factors that the agentic AI literature has largely proceeded without.

## 1.4 Research questions

**RQ1. Theory.** What are the theoretical and algorithmic foundations of AI agents and agentic AI systems?

**RQ2. Architecture.** What architectural paradigms dominate medical agentic AI, and how do they build on general-purpose agent frameworks?

**RQ3. Applications.** Across which clinical and biomedical domains have agentic systems been evaluated or deployed?

**RQ4. Evaluation.** How are these systems currently evaluated, and what does the evidence base support?

**RQ5. Safety and governance.** What failure modes, ethical concerns and regulatory requirements characterise the field?

**RQ6. Future agenda.** What open problems and research priorities should guide the next decade?

We adopt a broad definition of agentic AI encompassing single-agent LLM systems with tool use, multi-agent pipelines, reinforcement-learning-based treatment optimisation, embodied surgical robots and closed-loop physiological controllers. We exclude purely passive diagnostic classifiers that operate on a single fixed input without iterative reasoning or action selection.

## 1.5 Paper organisation

Table 2 maps each research question to the sections that address it and to the principal finding reported there.

**Table 2** Mapping of research questions to sections and principal findings

| RQ | Section | Principal finding |
|---|---|---|
| RQ1 | §3 | Agentic systems rest on four reusable components—planning, memory, tool use and coordination—formalised within (PO)MDP decision theory. |
| RQ2 | §4 | Five paradigms recur and nearly all deployed systems share a four-layer stack; we propose a six-level autonomy scale that partitions the space by where human judgement enters. |
| RQ3 | §5 | Coverage is broad but uneven; demonstrated benefit concentrates at high autonomy in narrow domains, research attention at intermediate autonomy in broad ones. |
| RQ4 | §6 | Benchmarks measure knowledge, not agency; evaluation requirements should be indexed to autonomy level, against which most published systems are under-evaluated. |
| RQ5 | §7, §8, §9 | Ten recurring failure modes, of which the tractable ones are engineering problems and the neglected ones are socio-technical; regulation is not yet indexed to autonomy. |
| RQ6 | §10 | Eight research priorities, of which prospective multi-site validation is the most urgent. |

# 2 Review methodology

This work is a *scoping review*, conducted according to the framework of Arksey and O'Malley (2005) as refined by Levac et al. (2010), and reported in accordance with the PRISMA extension for Scoping Reviews (PRISMA-ScR) (Tricco et al., 2018). A scoping design was chosen deliberately: the literature is heterogeneous in study type, outcome definition and reporting

quality, and the objective is to map the extent, range and nature of the evidence rather than to estimate a pooled effect. Consistent with PRISMA-ScR guidance, no formal risk-of-bias assessment of individual sources was undertaken; the implications of this choice are discussed in Section 11.2. The completed PRISMA-ScR checklist is provided as Online Resource 1.

The review protocol was registered with the Open Science Framework on 9 August 2026 (registration DOI: https://doi.org/10.17605/OSF.IO/WY8BN). Registration was *retrospective*: the review was substantially complete before the protocol was deposited, and the registered protocol therefore documents the methods as executed rather than as planned in advance of data collection. Two analytical instruments reported here—the autonomy scale of Section 4.2 and the evaluation dimensions of Section 6—were developed inductively during synthesis and are declared as such in the registration; neither was used to determine eligibility. Beyond these, the eligibility criteria, information sources and charting form were not altered after the search was designed and the charting form piloted. Readers should note that retrospective registration provides no independent assurance that the eligibility criteria were fixed before the results were known.

## 2.1 Information sources and search strategy

We searched PubMed/MEDLINE, IEEE Xplore, arXiv (cs.AI, cs.CL, cs.LG, cs.MA, eess.IV, q-bio), the ACL Anthology, Semantic Scholar, Scopus and Google Scholar (first 200 results). The initial search was executed between January and March 2025 and updated in June 2026 to capture material published during preparation of the manuscript; the updated search used an identical strategy, and records retrieved by it are included in the counts reported in Fig. 1.

The search combined agent-specific terms (*AI agent*, *LLM agent*, *autonomous agent*, *agentic AI*, *multi-agent system*, *reinforcement learning agent*, *tool-augmented LLM*) with medical and biomedical terms (*medicine*, *clinical*, *healthcare*, *diagnosis*, *treatment*, *drug discovery*, *radiology*, *surgery*, *electronic health record*). A second, separately reported strand retrieved foundational computer science literature on agent theory, large language models, planning frameworks, multi-agent systems and reinforcement learning, on the grounds that these underpin the clinical systems reviewed. Complete, database-specific search strings, given in the native syntax of each platform so that they can be executed verbatim, are provided as Online Resource 2. Reference lists of included reviews were hand-searched, and forward citation tracking was performed on the twenty most-cited included records.

## 2.2 Eligibility criteria

**Included:** peer-reviewed articles; preprints from recognised servers where no peer-reviewed version existed; technical reports from research organisations describing systems with sufficient methodological detail to be characterised; clinical trial registrations and results; regulatory guidance documents; and reference textbooks on agent theory. English language only. Published 1 January 2017 to 30 June 2026.

**Excluded:** opinion pieces without empirical or formal content; duplicate reports of the same system (the most complete version was retained); purely passive ML classifiers without iterative reasoning, tool use or action selection; works restricted to animal models without translational discussion; and records for which no full text could be obtained.

Eligibility was defined solely by these criteria. No source was included or excluded on the basis of authorship.

## 2.3 Selection of sources of evidence

Records were deduplicated in a reference manager and screened in two stages. Title and abstract screening, and subsequent full-text assessment, were each performed independently by two reviewers, with disagreements resolved by discussion and, where necessary, by adjudication from a third reviewer. Disagreement was infrequent and was resolved in all cases without recourse to a formal adjudication procedure beyond discussion. The selection process and the number of sources at each stage are summarised in Fig. 1.

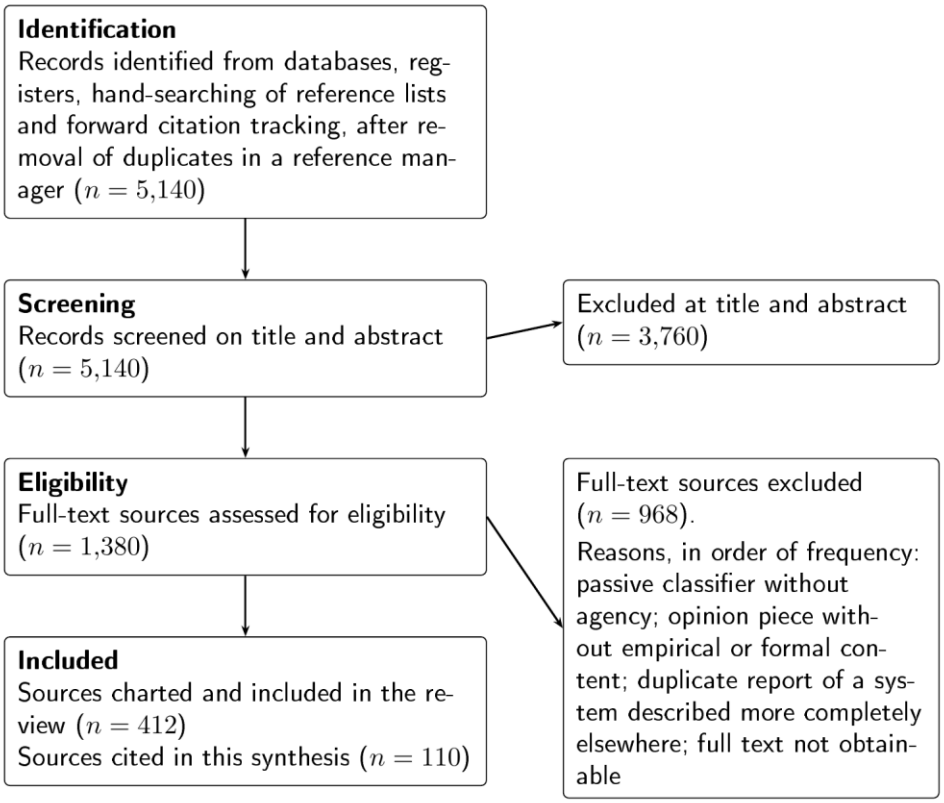



**Fig. 1** Flow of sources through the review, reported according to PRISMA-ScR. Boxes on the left give the number of records at each stage; boxes on the right give the number excluded and, at full-text stage, the reasons for exclusion. Records identified through hand-searching and forward citation tracking are included in the identification count

## 2.4 Data charting and synthesis

A charting form was piloted on twenty records and refined before full application. For each charted source we recorded: agent type; technical modality; clinical domain; data source and sample size; evaluation design; reported performance; whether any safety analysis was presented; and regulatory status. Foundational computer science sources were charted for their contribution to planning, memory, tool use or multi-agent coordination. Autonomy level was not a charting variable: the scale of Section 4.2 was developed from the charted corpus rather than applied during charting, and was then assigned at the level of the application domain (Table 4) using the coding rules of Section 4.2. Claims below about where the field's attention lies therefore rest on the distribution of domains and evaluation designs in Table 4, not on a count of publications per level.

Synthesis is narrative and organised around the research questions of Section 2. Of the 412 charted sources, this article cites a purposively selected subset of 110, chosen to illustrate each paradigm, domain and evaluation design; selection favoured sources reporting prospective or external validation and, where several described equivalent systems, the most completely reported. The criteria governing that selection are stated here rather than left implicit, so that the basis for inclusion in the narrative can be judged; the charted corpus itself is not deposited, and is available from the corresponding author on reasonable request.

## 2.5 Reconciliation with concurrent reviews

Published reviews of this literature report widely divergent counts of eligible sources—seven (Collaco et al., 2026a), forty-three (Njei et al., 2026), more than two hundred (Xu et al., 2026)—and readers are entitled to know why the present count differs from each. The divergence is

driven almost entirely by how *agent* is operationalised, and we state our position explicitly so that our corpus can be compared rather than merely contrasted.

Collaco et al. (2026a) require demonstrated autonomous operation in a described healthcare application, which excludes benchmark studies, simulated environments and systems not yet applied clinically; this yields the smallest count and the most clinically interpretable one. Njei et al. (2026) admit applied studies of agentic systems in healthcare across a longer window without requiring clinical deployment. Xu et al. (2026) include technical contributions describing agent architectures irrespective of evaluation.

Our eligibility criteria are broader than all three along three dimensions and narrower along none. First, we admit foundational computer science literature on agent theory, planning, memory, tool use and reinforcement learning, on the grounds that RQ1 concerns exactly this material; the reviews above are confined to healthcare applications. Second, we admit regulatory guidance, reporting standards and methodological literature on the evaluation of clinical AI, because RQ4 and RQ5 cannot be addressed without them. Third, our definition of agentic AI is not restricted to LLM-based systems: it encompasses offline reinforcement learning agents, embodied robotic agents and closed-loop physiological controllers, which the LLM-centred reviews exclude by construction. A corpus assembled on these criteria is necessarily larger, and is not comparable to those above without adjustment.

The consequence for interpretation is that our corpus should not be read as a larger inventory of the same object. Restricted to the intersection with Collaco et al. (2026a)—clinically deployed agentic systems with reported outcomes—our findings agree with theirs: such systems are few, most evaluation is exploratory, and patient-level evidence is confined to a small number of

narrowly scoped applications. That agreement, reached from a wider corpus and a broader definition, is a form of triangulation rather than a redundancy.

### 2.6 Use of generative AI in manuscript preparation

Generative AI tools were used for language editing and for checking bibliographic consistency. They were not used to select sources, to chart data, to generate results, or to draft scientific content. All text was written and verified by the authors, who take full responsibility for its accuracy and integrity.

## 3 Theoretical and algorithmic foundations

### 3.1 The agent paradigm

The intelligent agent has been a foundational construct in AI since the field's inception. Russell and Norvig (2020) define an agent as anything that perceives its environment through sensors and acts upon it through effectors; a *rational* agent selects actions expected to maximise its performance measure. This definition encompasses a rich taxonomy—simple reflex agents, model-based reflex agents, goal-based agents, utility-based agents and learning agents—that remains useful for classifying contemporary systems.

**Definition 1 (AI agent).** An AI agent is an entity that (i) perceives its environment through sensors or data feeds; (ii) maintains an internal state representation; (iii) reasons and plans over a finite or indefinite horizon; (iv) executes actions that may alter the environment or trigger external processes; and (v) updates its behaviour in response to feedback. Agents may be reactive, deliberative or hybrid.

**Definition 2 (agentic AI system).** An agentic AI system is a computational architecture in which one or more AI agents operate with substantial autonomy over multi-step tasks, orchestrating external tools, specialised sub-models or physical actuators, with limited per-step human intervention.

Formally, a single-agent problem is modelled as a Markov Decision Process (MDP) $\langle S, A, T, R, \gamma \rangle$, where $S$ is the state space, $A$ the action space, $T: S \times A \to \Delta(S)$ the transition function, $R: S \times A \to \mathbb{R}$ the reward function and $\gamma \in [0,1)$ the discount factor (Sutton and Barto, 2018). Under partial observability this generalises to a POMDP $\langle S, A, T, R, \Omega, O \rangle$ with observation space $\Omega$ and observation function $O$ (Kaelbling et al., 1998). The POMDP is the natural formulation for clinical agents, which never observe a patient's complete physiological state and must act under structured uncertainty. It also makes explicit why clinical agency is hard in a way that game-playing agency is not: the celebrated successes of deep reinforcement learning (RL) (Mnih et al., 2015; Silver et al., 2016) operated in environments with complete or cheaply sampled state and unlimited simulated experience, neither of which is available at the bedside.

## 3.2 Large language models as agent backbones

The emergence of LLMs as general-purpose reasoning engines has reshaped the agent landscape. Models such as GPT-4 (OpenAI, 2023) and Gemini (Gemini Team, 2023), together with medical adaptations, encode extensive world and clinical knowledge from pretraining and support zero-shot and few-shot generalisation across specialties. Singhal et al. (2023) showed that instruction-tuned LLMs encode substantial clinical knowledge, and clinical-domain pretraining on de-

identified electronic health record (EHR) text yields further gains on clinical natural language tasks (Yang et al., 2022).

Three capabilities underpin the use of LLMs as agent backbones. *In-context learning* (Brown et al., 2020) allows the model to condition its behaviour on task instructions and retrieved context without parameter updates. *Chain-of-thought reasoning* (Wei et al., 2022), strengthened by sampling several reasoning paths and taking the majority answer (Wang et al., 2023), improves performance on multi-step problems analogous to diagnostic workup. *Instruction following* aligned to human preferences (Ouyang et al., 2022) makes the model controllable enough to be placed in a loop at all. All three rest on the transformer architecture (Vaswani et al., 2017) and none is specific to medicine, which is precisely why medical agents inherit both the strengths and the failure modes of the general-purpose systems beneath them (Thirunavukarasu et al., 2023).

## 3.3 Planning and reasoning frameworks

### 3.3.1 ReAct: reasoning and acting

The ReAct framework (Yao et al., 2022) interleaves reasoning traces (*Thought*) with executable actions (*Act*) and environmental observations (*Observe*). This loop maps onto clinical reasoning: an agent can reason about a differential diagnosis, act by querying a drug interaction database, observe the result, and iterate towards a revised recommendation. ReAct and its variants are the dominant control loop in deployed medical LLM agents.

### 3.3.2 Tree-of-thought and Monte Carlo tree search

Yao et al. (2023) extend chain-of-thought to a tree-structured search over reasoning paths, allowing an agent to explore several diagnostic hypotheses in parallel and to backtrack from

unproductive branches. Monte Carlo tree search supplies a principled exploration algorithm combining rollout-based value estimation with upper-confidence-bound selection. The clinical appeal is evident; the computational cost, and the difficulty of specifying a value function over diagnostic states, are the practical obstacles.

### 3.3.3 Reflection and self-critique

Reflexion (Shinn et al., 2023) equips agents with the ability to reflect verbally on past failures and to encode the resulting lessons in episodic memory. In a clinical setting, an agent that proposes an unsafe treatment plan can incorporate feedback from a drug interaction checker and revise its recommendation on a subsequent attempt. Self-critique improves performance on tasks with verifiable feedback signals; where no such signal exists—as in much of diagnostic reasoning—its benefit is less well established.

## 3.4 Memory systems

Effective agents require structured memory beyond the context window of the underlying model. Four memory types are conventionally distinguished: *working memory*, the active context window holding the current encounter; *episodic memory*, logs of past encounters retrievable by similarity search; *semantic memory*, factual knowledge held in vector databases populated with literature and guidelines; and *procedural memory*, learned skills or tool-use policies encoded in model weights or as retrievable programs.

Retrieval-augmented generation (RAG) (Lewis et al., 2020) operationalises semantic memory by grounding model outputs in retrieved documents, reducing unsupported generation and enabling attribution to specific sources. Medical RAG systems query vector stores populated with

abstracts, clinical guidelines and drug compendia. Grounding mitigates but does not eliminate fabrication: an agent can retrieve a correct document and still misread it.

## 3.5 Tool use and function calling

Modern LLM interfaces expose structured function-calling capabilities that let agents invoke external tools against a declared schema, making actions verifiable and auditable. Schick et al. (2023) showed that models can learn when and how to call APIs in a self-supervised manner, and subsequent frameworks orchestrate heterogeneous tool collections (Shen et al., 2023) or multiple conversing agents each holding distinct tools (Wu et al., 2024). Medical tool libraries typically include FHIR resource retrieval; literature and trial registry search; ICD-10 and SNOMED-CT code lookup; drug interaction checking; imaging analysis sub-models; genomic annotation services; and laboratory reference range databases. Retrieval over curated clinical corpora with explicit source attribution has been shown to improve both accuracy and clinician-rated trust relative to unconstrained generation (Zakka et al., 2024). Schema validation at the tool boundary is one of the few points in an agentic pipeline where correctness can be enforced mechanically rather than statistically, which is why it recurs below as a mitigation for several failure modes.

## 3.6 Multi-agent systems

Multi-agent systems (MAS) comprise several interacting agents, each with its own perceptions, goals and actions (Wooldridge, 2009; Shoham and Leyton-Brown, 2009). In medical settings, MAS architectures allow complex clinical tasks to be decomposed across specialised agents—for example a differential diagnosis agent, a pharmacology agent, an imaging agent and a care-plan agent—on the hypothesis that the ensemble outperforms any single generalist model. The

theoretical foundations draw on game theory, distributed AI and mechanism design, and the design questions they raise—how to allocate roles, how to aggregate conflicting recommendations, how to prevent one agent's error from propagating—are directly analogous to those addressed in multidisciplinary team deliberation.

Two cautions apply when transferring MAS results to medicine. First, gains reported for agent ensembles are frequently confounded with additional inference-time computation: an ensemble that queries the same model five times must be compared against a single-agent baseline granted the same budget, and this control is often absent. Second, simulation studies of populations of interacting language agents (Park et al., 2023) demonstrate emergent coordination but also error propagation, and the architectural conditions under which one agent's mistake is caught rather than amplified remain poorly characterised—a gap that Section 7 treats as the cascading error failure mode.

## 3.7 Reinforcement learning for sequential decision-making

Reinforcement learning provides a principled framework for learning sequential decision policies through interaction with an environment. The agent observes state $s_t$, takes action $a_t$, receives reward $r_t$ and transitions to $s_{t+1}$, iterating towards a policy $\pi^*$ maximising expected cumulative discounted reward $\mathbb{E}[\sum_{t=0}^{T} \gamma^t r_t]$.

In clinical medicine online RL is rarely admissible, because exploration means experimenting on patients. *Offline RL* addresses this by learning from historical trajectories without live interaction (Levine et al., 2020). Conservative Q-Learning (Kumar et al., 2020) and Implicit Q-Learning (Kostrikov et al., 2022) are the algorithms most often applied to medical data; both penalise or avoid out-of-distribution actions, guarding against the over-optimistic value estimates that arise

when a learned policy recommends actions never observed in the data. Gottesman et al. (2019) set out the methodological conditions under which offline RL results in healthcare can be considered credible—adequate state representation, sufficient action overlap, and honest off-policy evaluation—conditions that a substantial part of the published literature does not meet.

# 4 Architectural paradigms in medical agentic AI

## 4.1 The agentic AI stack

Across the charted corpus, medical AI agents converge on four layers (Fig. 2). A *foundation layer* provides the pretrained model. A *reasoning layer* directs multi-step inference through chain-of-thought, ReAct, tree-of-thought or tree search. A *tool and memory layer* accesses external interfaces, knowledge bases, EHR systems and structured memory stores. An *orchestration layer* coordinates sub-agents, manages memory, enforces safety guard-rails and maintains the human oversight interface. The consistency of this decomposition is itself a finding: systems that differ markedly in clinical purpose are architecturally similar, which suggests that evaluation and governance instruments can be defined at the level of the stack rather than per application.

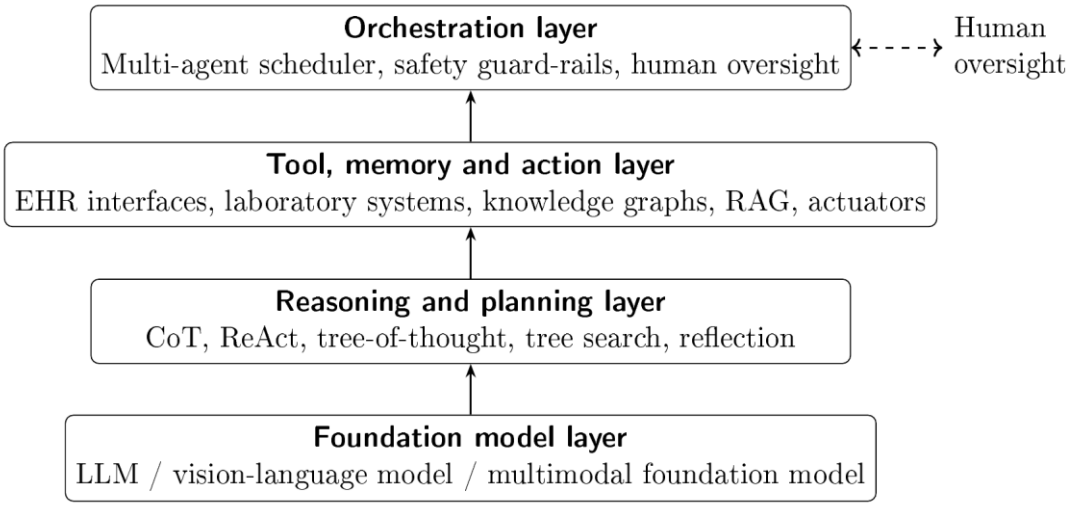



**Fig. 2** The four-layer agentic AI stack recurring across medical applications. The dashed bidirectional arrow denotes the human-in-the-loop oversight channel

## 4.2 Levels of clinical agent autonomy

The literature describes systems as "autonomous" with little consistency, and the resulting ambiguity has practical consequences: evaluation requirements, oversight obligations and regulatory classification all depend on how much judgement the system exercises without a human, yet no shared scale exists to express this. The five paradigms below differ in algorithm; they do not partition the design space in the way that matters for governance. A system built on offline RL and one built on an LLM with tool use may pose identical oversight problems if they act at the same point in the clinical pathway.

Precedent exists for this move. Yang et al. (2017) proposed a six-level scale for medical robotics—no autonomy, robot assistance, task autonomy, conditional autonomy, high autonomy, full autonomy—explicitly to support regulatory, ethical and legal reasoning, and it has been influential in surgical robotics (Attanasio et al., 2021). Both that scale and its analogue in driving automation derive from the older human-factors literature on levels of automation (Parasuraman

et al., 2000), which distinguishes automation of information acquisition, analysis, decision selection and action implementation.

We adopt the same six-level structure deliberately, so that the two scales can be read together, but the dimension being graded is different, and the difference is what makes a separate scale necessary. Yang et al. (2017) grade the *physical execution* of a task: the question is how much of the motion the robot performs without the surgeon's hand. For reasoning agents the motion is not the difficulty. A system that merely displays text can carry more unreviewed consequence than one that moves an instrument, if the text is the conclusion of a chain of inference no clinician will reconstruct. The variable that matters is therefore *where human judgement enters relative to the reasoning*, and—because oversight is only meaningful while an action can still be withdrawn—whether the action is reversible at the point the human sees it. Our scale, summarised in Table 3, grades on these two properties rather than on the division of physical labour. The two scales are orthogonal in principle: an L1 surgical robot in the sense of Yang et al. (2017) may sit at L2 on ours if its trajectory is planned by an unreviewable model, and this decoupling is precisely what the robotics scale was not designed to express.

**Table 3** Levels of clinical agent autonomy (LCAA). The reversibility column records whether an action at that level can be withdrawn before it affects the patient, which is the property that determines whether human oversight is meaningful

| Level | Label | Human judgement enters | Reversible |
|---|---|---|---|
| L0 | Passive prediction | Clinician interprets the raw output; the system does not recommend | Yes |
| L1 | Assisted interpretation | Clinician evaluates a recommendation and its stated rationale before acting | Yes |
| L2 | Supervised multi-step | System plans and executes retrieval and tool calls; clinician approves the final recommendation only | Yes |
| L3 | Conditional autonomy | System acts within a pre-approved envelope; clinician reviews exceptions and post hoc | Partly |
| L4 | Bounded full autonomy | System acts without per-case review within a narrow, well-characterised domain; humans set the bounds | No |
| L5 | Open-ended autonomy | System sets its own sub-goals across an unbounded action space | No |

Three observations follow from applying the scale to the systems reviewed here, and they recur throughout the remainder of this article.

First, *demonstrated patient benefit is concentrated at L3–L4 in narrow domains*. Closed-loop insulin delivery (L4) and autonomous retinopathy screening (L4) have randomised or pivotal prospective evidence; both operate on a single well-defined target variable with an unambiguous ground truth and a bounded action space. Adaptive radiotherapy replanning within pre-approved parameter bounds (L3) is the same pattern. Narrowness is not a limitation of these systems but the precondition of their approval.

Second, *research attention is concentrated at L2, where the evidence is weakest*. Seven of the thirteen domains in Table 4 code to L2, and two more span L2: multi-agent clinical reasoning, conversational diagnosis, tool-augmented differential diagnosis, clinical documentation, imaging reporting and offline-RL treatment optimisation all plan and execute autonomously but submit a recommendation for human approval. This is the level at which automation bias becomes the dominant risk (Section 7), because the human retains nominal authority while the reasoning that produced the recommendation is too long to audit case by case. It is also the level at which no prospective patient-outcome evidence exists: the strongest evidence at L2 is a randomised comparison against simulated patients and a non-randomised pre-post cohort.

Third, *L5 is at present a research aspiration rather than a described system*. No reviewed system sets its own clinical sub-goals across an unbounded action space, and the EU AI Act's requirement of effective human oversight (Section 9) makes it doubtful that L5 systems could be lawfully deployed in clinical care in their current conception. The autonomous chemistry agents of Boiko et al. (2023) approach L5 in a domain where errors are recoverable and outcomes objectively verifiable—conditions that do not transfer.

### 4.2.1 Coding rules

A scale that cannot be applied consistently is decorative. We therefore state the assignment procedure explicitly. A system is coded by answering four questions in order, each about the system as deployed rather than as described in its abstract.

**Q1.** *Does the system select among actions, or only produce an estimate?* If it emits a score, class or segmentation that a human must interpret into a decision, it is L0. Producing a ranked differential without recommending a next step is still L0.

**Q2.** *Does a human evaluate the reasoning, or only the conclusion?* If the rationale presented is short enough to be checked against the case in the time available, and the clinician is expected to do so, the system is L1. If the system executes multiple retrieval or tool-invocation steps and surfaces only the resulting recommendation for approval, it is L2. The distinguishing test is whether rejecting the recommendation requires redoing the system's work.

**Q3.** *Can the system's action reach the patient before a human sees it?* If not, coding stops at L1 or L2. If it can, but only within parameters a human approved in advance and with exceptions routed back for review, the system is L3. Escalation pathways, alert thresholds and pre-approved replanning bounds are the characteristic mechanisms.

**Q4.** *Who sets the goal, and is the action space enumerable?* A system acting without per-case review, pursuing a goal set by humans, over an action space that can be enumerated in advance is L4. A system that generates its own sub-goals over an action space that cannot be enumerated is L5.

Four conventions resolve the cases we found ambiguous while charting. *Code the deployed configuration, not the maximal capability*: a system technically capable of autonomous action but deployed with mandatory sign-off is coded at the level at which it runs. *Code the highest level reached by any routine action*, not the modal one, since safety is governed by the worst case. *A human in the loop who is not given the means to intervene does not lower the level*: nominal oversight without an interface for exercising it is coded as though absent, which is the practical content of the human-oversight requirement discussed in Section 9. And *reversibility is judged at the point the human sees the action*, not at the point it becomes clinically irreversible: an insulin dose already delivered is unreversed even if its effect can later be counteracted.

Two worked cases illustrate the procedure. Consider a sepsis early-warning system that fires an alert a clinician may act on or dismiss. It passes Q1, since it selects an action: to alert. At Q2 the clinician evaluates the alert rather than the model's inference. At Q3 the action—interrupting the clinician—has already reached them before review, within a threshold set in advance. It is therefore L3.

Now consider a multi-agent system that assembles a differential by querying the record, calling a drug-interaction checker and reconciling disagreement between sub-agents, then presents a recommendation for physician approval. It passes Q1. It is L2 at Q2, because rejecting the recommendation would require reproducing the retrieval chain. It stops there at Q3, because nothing reaches the patient unapproved.

We have not established inter-rater reliability for this procedure, and say so again in Section 11.2; doing so requires independent coders applying it to a common sample, which we regard as a worthwhile short study in its own right.

The scale is used in Table 4 to characterise each application domain, returns in Section 6, where we argue that evaluation requirements should be indexed to autonomy level rather than to algorithm class, and again in Section 9.

## 4.3 Single-agent LLM systems with tool use

The simplest agentic paradigm augments a single LLM with a curated toolkit and a ReAct-style planning loop. The foundation is typically either a general-purpose model or one pretrained on clinical text: GatorTron, trained on more than 90 billion words including over 82 billion words of de-identified clinical narrative and scaled to 8.9 billion parameters, improved performance across five clinical natural language processing tasks relative to earlier biomedical transformers (Yang et al., 2022). Adding a FHIR retrieval layer and retrieval-augmented access to guidelines converts such a model from a text encoder into an agent that can assemble the evidence for a recommendation rather than merely score text. Van Veen et al. (2024) provide one of the clearer demonstrations that adapted LLMs can match or exceed clinician performance on a bounded clinical text task, in their case summarisation. Wornow et al. (2023) caution, however, that foundation models for electronic health records are frequently evaluated on tasks and data of limited clinical relevance, so that reported capability overstates readiness—a caveat that applies with greater force once the model is granted the ability to act.

The precision pharmacokinetic workflow of Frasca et al. (2026) illustrates the clinical pipeline that this paradigm is positioned to automate: patient-level covariates feed gradient-boosted and neural dosing models, whose outputs are rendered interpretable through local surrogate explanation to support paediatric antibiotic dose selection. The agentic extension would have such a system retrieve the patient data, query the dosing model, validate the result against

contraindication rules and deliver an explained recommendation without manual coordination between steps. That extension has not, to our knowledge, been prospectively evaluated, and the distinction between a demonstrated pipeline and its projected agentic form is one we maintain throughout this review.

## 4.4 Multi-agent systems

### 4.4.1 Debate and verification architectures

Architectures in which several model instances argue towards a conclusion, with a judge selecting the best-supported answer, improve factual accuracy and reasoning on general benchmarks relative to single-model baselines (Du et al., 2023). The mechanism—forcing an explicit defence of intermediate claims—is attractive for clinical reasoning, but we found no prospective clinical evaluation of debate architectures, and their benefit in medical settings should presently be regarded as extrapolated from general-domain results rather than demonstrated.

### 4.4.2 Role-specialised agent teams

A second family assigns clinical roles to separate model instances that communicate through structured messages. Tang et al. (2024) instantiate a multi-disciplinary team of domain-expert agents for zero-shot medical reasoning, and Kim et al. (2024) show that adapting the collaboration structure to task complexity—solo for simple queries, group deliberation for complex ones—outperforms fixed structures across medical benchmarks. Both are retrospective benchmark studies; neither establishes clinical utility.

### 4.4.3 Hierarchical orchestration and conversational diagnostic AI

AMIE employs a hierarchical structure in which the system decomposes a consultation, delegates components, and synthesises the output (Tu et al., 2024). In a randomised, blinded, text-based study using simulated patients across 149 case scenarios, AMIE outperformed primary care physicians on diagnostic accuracy and on most measures of consultation quality as rated by specialist physicians and by the simulated patients. The result is notable, and its boundaries are equally so: text-only interaction through a chat interface is not the medium in which primary care is practised, a point the authors themselves emphasise.

### 4.4.4 Biomedical discovery agents

Gao et al. (2024) set out a vision of multi-agent discovery systems in which LLM agents, domain-specific ML tools and human experts collaborate to plan experiments, execute them through tool calls to biological databases or laboratory platforms, assess the results and iterate. Working implementations exist outside medicine—M. Bran et al. (2024) augment an LLM with chemistry tools to plan and execute syntheses, and Boiko et al. (2023) demonstrate autonomous design and execution of chemical experiments—and establish that the paradigm is technically feasible where actions are machine-executable and outcomes are objectively verifiable. Neither condition holds straightforwardly in clinical medicine.

## 4.5 Reinforcement learning agents

### 4.5.1 Offline RL for treatment optimisation

Komorowski et al. (2018) pioneered offline RL for sepsis management, learning a policy over fluid and vasopressor dosing from intensive care trajectories and reporting that patient mortality

was lowest when clinician actions matched the learned policy. The work established the template for the field and, equally, the template for its critique: the estimate is derived from off-policy evaluation on retrospective data, not from prospective comparison, and subsequent methodological work has examined how sensitive such estimates are to state representation and evaluation choices (Gottesman et al., 2019). No offline RL treatment policy has to date been prospectively validated in a randomised trial.

### 4.5.2 Closed-loop physiological control

Closed-loop insulin delivery is the clearest case of an autonomous agent delivering demonstrated patient benefit. In a 13-week multicentre randomised trial of 326 participants aged six years and over, the bionic pancreas reduced glycated haemoglobin from 7.9% to 7.3% while the standard-care group remained at 7.7%, an adjusted between-group difference of $-0.5$ percentage points (Russell et al., 2022). Comparable benefit had earlier been shown for hybrid closed-loop control in a six-month randomised trial (Brown et al., 2019). Both systems combine model-predictive control with adaptive parameter estimation rather than deep RL, a distinction often blurred in secondary literature.

### 4.5.3 Radiation therapy planning

RL agents for automated treatment planning learn to adjust beam configurations and dose-volume objectives against clinical constraints, with the aim of compressing a planning process that is otherwise iterative and labour-intensive. Across the sources charted in this domain, reported reductions in planning time are substantial, but they derive from retrospective replanning of previously treated cases, and plan acceptability is generally assessed by the

developing centre rather than independently. We are not aware of a prospective evaluation of an autonomous planning agent against clinical outcomes.

## 4.6 Embodied robotic agents

The Smart Tissue Autonomous Robot achieved autonomous laparoscopic intestinal anastomosis in a porcine model with greater consistency than expert surgeons, using structured-light three-dimensional sensing, force feedback and a learned suturing policy (Saeidi et al., 2022). This remains preclinical work: no autonomous soft-tissue surgical agent has been evaluated in humans. Rehabilitation exoskeletons increasingly incorporate adaptive controllers that modulate assistance to patient effort in real time, and represent a lower-autonomy, higher-maturity counterpart.

## 4.7 Autonomous monitoring and alert agents

Deployed early-warning agents constitute the largest body of real-world evidence in this review. The TREWS sepsis system was evaluated in a prospective, multi-site study covering 590,736 patients monitored across five hospitals, in which physician confirmation of the alert within three hours was associated with reduced in-hospital mortality (Adams et al., 2022); a companion analysis examined the determinants of clinician adoption and the association with treatment timing (Henry et al., 2022). These are observational deployment studies, not randomised trials, and the association between alert interaction and outcome is subject to confounding by indication—an ambiguity that recurs throughout Section 6 and that the field has not yet resolved. Wearable-based arrhythmia detection is the most widely deployed consumer-facing instance of continuous autonomous monitoring, with regulatory clearance obtained on the basis of diagnostic accuracy against ambulatory electrocardiography rather than on clinical outcomes.

# 5 Clinical application domains

Table 4 summarises the domains reviewed in this section against the dominant architectural paradigm, the typical autonomy level of Section 4.2, and the strongest form of evidence identified. The pattern noted above is visible in a single view: prospective randomised evidence appears only where autonomy is high but the domain is narrow, and is absent precisely where reasoning is most ambitious.

**Table 4** Clinical application domains by dominant paradigm, typical autonomy level (Table *3*) and strongest identified evidence

| Domain | Dominant paradigm | LCAA | Strongest evidence identified |
|---|---|---|---|
| Autonomous retinopathy screening | Perception + decision rule | L4 | Prospective pivotal, multi-site |
| Closed-loop glucose control | Control + adaptive estimation | L4 | Randomised controlled trial |
| Adaptive radiotherapy replanning | RL / optimisation | L3 | Retrospective, bounded deployment |
| Sepsis early warning | Predictive model + alerting | L3 | Prospective multi-site cohort |
| Clinical documentation | LLM, single agent | L2 | Pre-post and non-randomised cohort |
| Diagnostic imaging reporting | Multimodal LLM + tools | L2 | Retrospective reader studies |
| Conversational diagnosis | Hierarchical multi-agent | L2 | Randomised, simulated patients |
| Differential diagnosis support | LLM + tool use | L2 | Benchmark only |
| Multi-agent clinical reasoning | Role-specialised MAS | L2 | Benchmark only |
| Treatment optimisation | Offline RL | L2 | Retrospective off-policy evaluation |
| Conversational psychotherapy | LLM, single agent | L3 | Randomised vs waitlist/information control |
| Drug discovery and biomarkers | Tool-using discovery agent | L2–L4 | Demonstrated in chemistry; not clinical |
| Digital twins and federated care | Infrastructure + modelling | L1–L2 | Architectural demonstration |

## 5.1 Diagnostic imaging

Diagnostic imaging is the most mature domain for agentic deployment. Imaging agents span the reporting pipeline: retrieval, prior comparison, structured report generation, critical finding flagging and communication. Multimodal foundation models capable of joint image-text reasoning (Moor et al., 2023) supply the technical basis, though evaluation to date is predominantly retrospective and reader-study based. Autonomous diabetic retinopathy screening established the regulatory precedent for the field: the pivotal study of an autonomous diagnostic

system enrolled 900 participants at ten primary care sites and reported 87.2% sensitivity and 90.7% specificity against a reading-centre reference standard, supporting the first authorisation of a system that renders a diagnosis without physician interpretation (Abramoff et al., 2018).

## 5.2 Clinical decision support

### 5.2.1 Differential diagnosis generation

Instruction-tuned LLMs exceed the passing threshold on medical licensing examination material (Nori et al., 2023; Singhal et al., 2023), a result that established the plausibility of the field but that measures recall of examinable knowledge rather than the capacity to manage an unfolding clinical problem. Agentic extensions that iteratively gather patient-specific data from the record are the natural next step, and are the subject of the agent benchmarks discussed in Section 6.

### 5.2.2 Precision pharmacology

The precision dosing framework of Frasca et al. (2026) demonstrates an interpretable ML pipeline for paediatric antibiotic dosing amenable to agentification, as discussed in Section 4. Analogous pipelines are emerging in oncology pharmacology, particularly for drug repurposing and individualised chemotherapy dosing.

### 5.2.3 Clinical documentation

Ambient documentation agents transcribe and structure the clinical encounter. They are among the most widely adopted applications, on the rationale that documentation burden is a principal driver of clinician burnout. Published evaluations are largely pre-post or non-randomised cohort studies at single health systems. They report reductions in documentation time and in clinician-reported burden; randomised evidence, and evidence at the level of patient outcomes, remains

limited. This is the clearest instance in the review of a technology whose adoption has outpaced its evaluation, and it is instructive that the barrier is not technical difficulty but the absence of an agreed outcome against which to measure benefit.

## 5.3 Neurology and psychiatry

Ricetti et al. (2025) review ML applications in neurological and psychiatric disease diagnosis, documenting strong reported performance in Alzheimer's disease, Parkinson's disease and mood disorders alongside persisting challenges in data quality, privacy and the ethics of algorithmic decision-making. These challenges intensify in agentic settings, where the system acts on model output rather than presenting it for clinician appraisal.

Autonomous psychiatric monitoring agents analyse continuous speech and text for markers of psychosis, suicidality and manic relapse. Conversational cognitive-behavioural agents have shown reductions in depressive symptoms against waitlist control in randomised trials (Fitzpatrick et al., 2017), operating as adjuncts to licensed clinicians with escalation pathways. Waitlist control establishes efficacy against no treatment, not against existing care, and the trials are short; both limitations are material to deployment decisions.

## 5.4 Oncology and precision medicine

Hybrid quantum-classical architectures have been applied to dermatological lesion classification, combining segmentation with a hybrid convolutional/quantum classifier (Frasca et al., 2025). Reported accuracies on this benchmark should be interpreted with care: dermoscopic datasets of this kind contain multiple images of the same lesion, and random train-test partitioning inflates

apparent performance unless splitting is performed at lesion level. The methodological interest of the approach lies in the architecture rather than in the headline accuracy.

AI in proton therapy—spanning planning, delivery optimisation and adaptive quality assurance (Isaksson et al., 2024)—exemplifies the multi-step workflows emerging in radiation oncology, where commercial adaptive systems now replan fractions within pre-approved parameter bounds. This is bounded autonomy of exactly the kind regulators can accommodate, and is worth contrasting with the open-ended autonomy of conversational agents.

## 5.5 Drug discovery and biomarker identification

Structure prediction extended to protein-ligand and protein-nucleic acid complexes (Abramson et al., 2024) supplies a computational substrate for agentic docking and lead optimisation. Working autonomous discovery agents have been demonstrated in chemistry (M. Bran et al., 2024; Boiko et al., 2023), where hypotheses can be tested by machine-executable experiments. Multi-omic integration agents that reason over biomedical knowledge graphs are an active research direction; published biomarker findings from such systems have not to our knowledge been independently validated in prospective cohorts.

## 5.6 Predictive digital twins and federated healthcare

Predictive digital twins—virtual patient representations that forecast outcomes and support resource allocation—are an emerging agentic paradigm whose principal obstacle is not modelling but data access. Blockchain-integrated digital twin frameworks address secure multi-institutional exchange through decentralised ledgers providing verifiable provenance (Repetto et al., 2025), and federated learning (Rieke et al., 2020) trains models across institutions without

centralising raw patient data. Together these supply the infrastructure on which multi-site agentic deployment depends.

### 5.7 Public health and epidemiological modelling

Multi-agent simulation has a long history in epidemic policy analysis, where agents represent individuals or institutions and the object of study is the emergent population dynamic. This use of the term *agent* is distinct from the LLM-based agents that are the principal subject of this review, and the two literatures have to date developed largely independently. The connection is nonetheless substantive: policy optimisation under uncertainty is a sequential decision problem of the type formalised in Section 3, and agentic systems that reason over surveillance data and simulation output are a plausible near-term development.

## 6 Evaluation frameworks

### 6.1 Benchmark landscape

Table 5 summarises benchmarks used to evaluate medical AI agents. The distribution is informative in itself: most measure knowledge recall or single-turn task performance, and only a minority evaluate multi-step interaction with a clinical environment—which is the defining property of an agent. A system can therefore score highly on the majority of this landscape without demonstrating any agentic capability at all.

**Table 5** Benchmarks used to evaluate medical AI systems, by what they measure

| Benchmark | Description | Measures | Metric |
|---|---|---|---|
| MedQA | USMLE-style multiple-choice questions | Knowledge | Accuracy |
| MedMCQA | Indian medical entrance examination questions | Knowledge | Accuracy |
| PubMedQA | Biomedical question answering over abstracts | Knowledge | Accuracy/F1 |

| Benchmark | Description | Measures | Metric |
|---|---|---|---|
| MIMIC-CXR | Chest radiographs with paired free-text reports | Perception | BLEU, F1 |
| PathVQA | Pathology visual question answering | Perception | Accuracy |
| AgentBench | General multi-turn agentic task completion | Agency | Success rate |
| MedAgentBench | EHR-based clinical agent tasks in a FHIR environment | Agency | Task success |
| AgentClinic | Simulated clinical encounters with patient, examiner and tool agents | Agency | Diagnostic accuracy |
| EHRAgent | Multi-step reasoning over structured EHR tables | Agency | Execution accuracy |

Agent-specific benchmarks are recent and remain few. MedAgentBench situates tasks in a FHIR-compliant environment requiring genuine multi-step retrieval and action (Jiang et al., 2025); AgentClinic constructs simulated clinical encounters in which a doctor agent interacts with patient, measurement and moderator agents across multiple specialties and languages (Schmidgall et al., 2024); EHRAgent evaluates code-mediated reasoning over structured records (Shi et al., 2024). These are the appropriate instruments for the systems this review describes, and their novelty is the reason so few published systems have been assessed with them.

## 6.2 Prospective clinical evaluation

Benchmark performance is necessary but not sufficient, and the gap between the two is now quantified. Bedi et al. (2025), reviewing health care applications of large language models, found that evaluation is dominated by accuracy on question-answering tasks, that few studies assess systems on real patient care data, and that dimensions such as fairness, bias and toxicity are assessed in a small minority of studies. The pattern is long-standing rather than specific to language models: Nagendran et al. (2020) found that studies comparing AI systems with clinicians were at high risk of bias and rarely reported to existing standards, and Wu et al. (2021) showed that most authorised AI devices had been evaluated retrospectively, with a minority reporting multi-site evaluation. Kelly et al. (2019) set out why this matters for clinical impact rather than merely for publication quality.

Three findings sharpen the problem specifically for agents. First, performance on examination-style benchmarks does not transfer to interactive clinical reasoning: Johri et al. (2025) showed, using simulated patient dialogue, that models scoring highly on structured vignettes degrade substantially when required to elicit the information themselves through conversation—which is what an agent must do. Second, deployment studies repeatedly find that performance measured in the laboratory does not survive contact with the clinical workflow (Beede et al., 2020).

Third, and most directly, the prospective evidence base for agentic pipelines specifically is effectively empty. Karunanayake (2026) observes that prospective trials represent a very small fraction of the clinical literature on large language models, and that agentic pipelines specifically are not represented among them. Independently, Collaco et al. (2026a) applied strict eligibility criteria to clinically applied agentic systems and identified seven, of which one involved patients. Two reviews using different methods thus converge on the same conclusion. There is no indication that agentic systems are evaluated more rigorously than the non-agentic devices examined by Wu et al. (2021); the added autonomy makes the deficit more consequential.

Appropriate prospective designs include randomised trials measuring patient outcomes; non-inferiority studies against specialist consensus with pre-specified margins; human-agent teaming studies measuring clinician performance with and without the system; and process metrics such as time-to-decision. Reporting guidance now exists for each stage—CONSORT-AI for trial reports (Liu et al., 2020), DECIDE-AI for early-stage live clinical evaluation (Vasey et al., 2022), and TRIPOD+AI for prediction model studies (Collins et al., 2024)—but none of these addresses the properties specific to agents, a gap taken up in Section 10.

Table 6 summarises the principal prospective evaluations identified in this review, reported with their actual study designs. The first four were conducted in patients; the last two in volunteers

recruited outside clinical care and in simulated patients respectively, and that distinction is material to what they establish.

**Table 6** Principal prospective evaluations of the medical AI agent and closed-loop systems reviewed here. Study design is stated explicitly because it varies substantially across these evaluations and is frequently misreported in secondary literature. *N* denotes enrolled participants unless the row states otherwise; the first four rows report evaluations in patients, the last two in volunteers recruited outside clinical care and in simulated patients respectively

| System and intervention | Study design | N | Principal result |
|---|---|---|---|
| TREWS: sepsis early-warning agent | Prospective multi-site cohort | 590,736† | Alert confirmation within 3 h associated with reduced in-hospital mortality (Adams et al., 2022) |
| Autonomous retinopathy diagnosis | Prospective single-arm pivotal vs reference standard | 900 | 87.2% sensitivity; 90.7% specificity (Abramoff et al., 2018) |
| Bionic pancreas: autonomous insulin delivery vs standard care | Randomised controlled trial, 13 weeks | 326 | HbA1c −0.5 percentage points vs standard care (Russell et al., 2022) |
| Hybrid closed-loop control vs sensor-augmented pump | Randomised controlled trial, 6 months | 168 | Increased time in target glucose range (Brown et al., 2019) |
| Conversational CBT agent vs information-only control | Randomised controlled trial, 2 weeks | 70 | Significant reduction in depressive symptoms (Fitzpatrick et al., 2017) |
| AMIE: conversational diagnostic agent vs primary care physicians | Randomised, blinded study with simulated patients | 149‡ | Superior diagnostic accuracy and consultation quality ratings (Tu et al., 2024) |

†patient encounters, not distinct patients. ‡simulated case scenarios presented to patient actors, not patients.

Two observations follow. First, only three of these evaluations are randomised trials with patient-level outcomes, and two of the three concern closed-loop glucose control—a domain with an unambiguous, continuously measurable target variable and decades of device regulation behind it. Second, the systems with the greatest reasoning autonomy have the weakest outcome evidence: AMIE was evaluated against simulated rather than real patients, and no multi-agent clinical reasoning system has been tested prospectively at all.

## 6.3 Agentic-specific evaluation dimensions

Standard ML metrics do not capture agency. We propose seven complementary dimensions, each of which is measurable and none of which is addressed by the knowledge benchmarks of Table 5:

1. *Task completion rate*: the fraction of multi-step tasks completed without human intervention.
2. *Tool use fidelity*: the correctness and appropriateness of tool invocations, including refusals to invoke.
3. *Factual grounding*: the proportion of clinical claims traceable to a retrievable and correct source.
4. *Reasoning coherence*: internal consistency of the reasoning chain across steps, including consistency between stated rationale and executed action.
5. *Safety action rate*: the frequency of appropriate safety behaviours—escalation, refusal, requests for clarification—measured against cases where such behaviour is warranted.
6. *Longitudinal consistency*: stability of recommendations for identical cases presented at different times.
7. *Human calibration*: agreement between the agent's expressed confidence and its empirical accuracy.

Dimensions (4) and (5) are the least tractable and the most important: an agent whose stated reasoning diverges from its executed action is not auditable, and an agent that never refuses is unsafe irrespective of its accuracy.

## 6.4 Indexing evaluation requirements to autonomy

We argue that evaluation requirements should be indexed to autonomy level (Table 3) rather than to algorithm class. The rationale is that the quantity determining what evidence is needed is not how the output was computed but how much unreviewed consequence it carries. Table 7 sets out the correspondence we propose.

**Table 7** Proposed minimum evaluation requirements by autonomy level. Each level inherits the requirements of the levels below it

| Level | Minimum evaluation | Agentic dimensions (Section 6) required |
|---|---|---|
| L0–L1 | External validation on independent data; subgroup-stratified reporting | (3) factual grounding; (7) calibration |
| L2 | Multi-step task evaluation in a simulated clinical environment; human-agent teaming study | adds (1) task completion; (2) tool fidelity; (4) reasoning coherence |
| L3 | Prospective deployment study with pre-specified exception handling and drift monitoring | adds (5) safety action rate; (6) longitudinal consistency |
| L4–L5 | Randomised trial with patient-level outcomes; independent external validation; post-market surveillance | all seven, reported continuously rather than at a single time point |

Applied to the current literature, this scheme exposes a systematic mismatch. The L4 systems in Table 4 broadly satisfy their row, because device regulation compelled it. The L2 systems—which constitute the majority of published research—are almost universally evaluated to the L0–L1 standard, reporting benchmark accuracy where the scheme requires simulated-environment task evaluation and a teaming study. The gap is not that the field evaluates poorly by its own lights, but that it evaluates systems as if they were predictive models when they are decision-making agents.

# 7 Safety and reliability

## 7.1 Fabrication and factual error

The generation of plausible but incorrect statements poses acute risks in medicine, and the risk compounds in agentic settings where an unverified statement becomes the input to a subsequent action. Med-HALT provides a purpose-built test of this failure in medical LLMs and demonstrates that it occurs at rates incompatible with unsupervised clinical use (Umapathi et al., 2023). Bibliometric analysis identifies interpretability and explainability as dominant frontiers in the medical AI literature (De Nicolò and La Torre, 2024a), and both bear directly on the detection of fabricated content. Mitigations include retrieval grounding with source attribution, auditing of reasoning traces, confidence-gated generation and ensemble verification; none is sufficient alone.

## 7.2 Distributional shift

Models trained at one health system frequently degrade at another, owing to differences in case mix, coding practice and instrumentation (Finlayson et al., 2021). Federated learning (Rieke et al., 2020) and verifiable multi-institutional data frameworks (Repetto et al., 2025) address the development side of the problem. The deployment side requires prospective performance monitoring and automated drift detection, which remain uncommon: most reported systems are validated once and thereafter assumed stable.

## 7.3 Reward hacking in RL agents

Agents optimised against proxy rewards can learn to satisfy the metric without achieving the intended goal—a failure mode formalised in the AI safety literature as specification gaming

(Amodei et al., 2016). In a clinical context, a policy rewarded on survival within an observation window can appear to improve outcomes by favouring actions correlated with, rather than causal of, survival. The methodological safeguards are constraint-based formulations, careful reward design and adversarial testing, together with the offline evaluation conditions set out by Gottesman et al. (2019).

## 7.4 Automation bias

Clinicians are susceptible to automation bias: uncritical acceptance of system recommendations even where these conflict with clinical judgement. This is a long-established finding in clinical decision support (Goddard et al., 2012) and rests on a broader literature documenting both over-reliance and disuse of automation as a function of trust calibration (Parasuraman and Riley, 1997). It is aggravated by alert fatigue from low-specificity alerts. Agent design must therefore incorporate calibrated uncertainty communication, deliberate friction at high-stakes decisions, structured model documentation at the point of use (Sendak et al., 2020), and mechanisms that sustain active clinician engagement.

Automation bias is a property of the socio-technical system rather than of the model, and cannot be measured by any benchmark in Table 5. Its significance grows with autonomy level in a non-obvious way: it is not maximal at L4, where no human is asked to approve anything, but at L2, where a human is nominally accountable for a recommendation whose derivation is too long to audit. Rendering that derivation reviewable is therefore an interface and workload problem as much as an interpretability one, and Beede et al. (2020) illustrate how readily such problems are missed when evaluation stops at model performance.

## 7.5 Security and adversarial attack

Medical agents face prompt injection through malicious instructions embedded in free-text fields (Greshake et al., 2023), data poisoning, evasion attacks using imperceptible perturbations to medical images (Finlayson et al., 2019), and model extraction. Tool-using agents widen the attack surface considerably, since a successful injection can cause not only a wrong answer but a wrong action. Certification standards analogous to those used in other safety-critical domains have been proposed for trustworthy AI (Wing, 2021).

## 7.6 Failure mode taxonomy

Table 8 classifies the failure modes recurring across the charted corpus, with indicative severity and the mitigations most commonly proposed. Severity reflects the potential consequence for an individual patient if the failure is undetected.

**Table 8** Taxonomy of medical AI agent failure modes with indicative severity, the layer of the agentic stack (Fig. *2*) at which each is best addressed, and the corresponding mitigation. F = foundation, R = reasoning, T = tool and memory, O = orchestration, S = socio-technical (outside the stack)

| Failure mode | Description | Sev. | Layer | Mitigation |
|---|---|---|---|---|
| Fabrication | Unsupported clinical assertion | Critical | F, T | Retrieval grounding, source attribution |
| Distributional shift | Degradation outside training distribution | High | F, O | Prospective monitoring, drift detection |
| Reward hacking | Optimisation of proxy rather than goal | Critical | F, R | Constrained RL, adversarial testing |
| Automation bias | Uncritical clinician over-reliance | High | S | Friction design, calibrated uncertainty |
| Tool misuse | Incorrect or inappropriate invocation | High | T | Schema validation, invocation audit |
| Cascading error | Error propagation across agents | Very high | O | Circuit breakers, independent verification |
| Adversarial attack | Deliberate input manipulation | High | T, O | Input sanitisation, adversarial training |
| Latency failure | Missed real-time clinical window | High | O | Service-level monitoring |
| Consent violation | Use of data outside authorised purpose | Critical | T, O | Purpose binding, de-identification |
| Equity failure | Disparate performance across groups | High | F, S | Stratified evaluation, disparity monitoring |

Two features of this mapping deserve emphasis. The failure modes addressable within the tool and orchestration layers—tool misuse, cascading error, latency, consent violation—are those for which mechanical rather than statistical guarantees are available, and are therefore the most tractable targets for engineering effort and for certification. Conversely, automation bias and, in part, equity failure are socio-technical: they arise from the interaction between the system and its users and organisation, cannot be measured by any benchmark in Table 5, and are systematically under-addressed in the literature precisely because they fall outside the artefact the developers control. This asymmetry between what is easy to fix and what is easy to overlook is, in our reading, the central safety problem of the field.

# 8 Ethical and societal dimensions

## 8.1 Algorithmic fairness and health equity

Systems trained predominantly on data from well-resourced health systems exhibit performance disparities across racial, ethnic, socioeconomic and gender groups. Obermeyer et al. (2019) demonstrated how an algorithm using healthcare cost as a proxy for health need systematically underestimated the illness burden of Black patients. The problem persists in current models: Omiye et al. (2023) found that commercially available LLMs reproduced debunked race-based clinical claims, and Zack et al. (2024) found race- and gender-based stereotyping in clinical vignettes. Agentic systems inherit these biases and add a further layer: an agent that acts on a biased assessment removes the point at which a clinician might have caught it, and an agent that chains several biased judgements compounds them without any single step appearing anomalous.

Equity also has a human side. A multicentre cross-sectional survey of Italian physicians' attitudes and trust towards AI in healthcare (De Nicolò and La Torre, 2024b) found that trust and willingness to adopt vary across professional demographics. Uneven adoption produces uneven benefit even where the underlying model is fair, which is why stratified evaluation should extend to use as well as to performance. Fairness audits across demographic strata should be a prerequisite for deployment (Chen et al., 2021).

## 8.2 Informed consent and transparency

Patients have a legitimate interest in knowing when AI agents participate in their care. Transparency encompasses disclosure that such tools are in use, intelligible explanation of their function, and access to the reasoning behind specific recommendations. Local surrogate explanation of the kind applied to dose prediction by Frasca et al. (2026) illustrates one workable form of clinician-facing explanation, though post-hoc explanation of a single prediction is a weaker guarantee than a traceable reasoning chain, which is what agentic architectures could in principle provide.

There is, however, a difficulty specific to agency that the explainability literature has not addressed. Post-hoc explanation attaches to a prediction; an agent produces a sequence of decisions, each conditioned on the outputs of the previous ones, and explaining any single step does not account for the trajectory. An agent may take five individually defensible steps to an indefensible conclusion, and a per-step explanation will show nothing wrong. The unit requiring explanation is therefore the trajectory, and the property required of it is not intelligibility alone but *faithfulness*: that the trace presented is in fact the computation performed rather than a plausible reconstruction generated afterwards. Whether current models' stated reasoning is

faithful in this sense is an open problem in interpretability research, and its resolution bears directly on whether L2 autonomy can be supervised at all.

Consent raises a parallel structural problem. Consent is ordinarily given to a defined intervention, but an agent that decides at run time which tools to invoke and which data to retrieve generates an interaction pattern that cannot be fully enumerated in advance. The workable response is probably to shift the object of consent from the specific data flow to the bounded envelope within which the agent operates—which is another reason for the action space to be declared and bounded, as Section 10.2 requires of any study reporting such a system.

## 8.3 Professional accountability and liability

Three liability models are canvassed in the literature: *product liability*, placing primary responsibility on the developer; *professional liability*, placing it on the deploying clinician; and *shared* frameworks distributing accountability according to the degree of human oversight (Gerke et al., 2020). Under current tort doctrine, a clinician who follows an AI recommendation that turns out to be wrong is more exposed than one who disregards it, which creates an incentive structure at odds with the clinical value of the technology (Price et al., 2019). Autonomy complicates all three models, since the clinician's capacity to supervise diminishes precisely as the system's value proposition increases (Char et al., 2018). The regulatory status of systems that give clinical advice without device authorisation is a related and unsettled question (Gilbert et al., 2023).

## 8.4 Data privacy and sovereignty

Privacy-preserving techniques—federated learning, differential privacy and secure multi-party computation—permit model development without centralising raw patient data (Rieke et al., 2020; Kairouz et al., 2021). Multi-site feasibility is established: federated training has matched centralised performance for tumour segmentation across dozens of institutions (Sheller et al., 2020) and supported outcome prediction across twenty hospitals during the COVID-19 pandemic (Dayan et al., 2021). Blockchain-integrated digital twin architectures add cryptographically verifiable provenance for multi-institutional systems (Repetto et al., 2025), addressing privacy and auditability together. The residual legal and normative questions concerning secondary use of health data are not resolved by any of these techniques (Price and Cohen, 2019). Agentic systems raise a further question that these techniques do not answer: an agent that retrieves, combines and transmits patient data across tools creates data flows that are difficult to enumerate in advance and therefore difficult to consent to specifically.

# 9 Regulatory frameworks

## 9.1 United States: Software as a Medical Device

The FDA's Software as a Medical Device framework stratifies risk by intended use. The Predetermined Change Control Plan pathway, set out in final guidance issued in December 2024 (FDA, 2024), allows developers to pre-specify anticipated model modifications together with the methodology for developing, validating and implementing them, reducing the burden of post-market review for changes within the declared envelope; it is the principal existing mechanism for accommodating models that change after authorisation. The authorised population of AI-

enabled devices now exceeds one thousand and continues to grow (FDA, 2025), though it remains dominated by systems that produce estimates rather than take actions. Agentic systems that act autonomously face more demanding pathways, with cybersecurity assessment and post-market surveillance obligations. Wu et al. (2021) documented substantial limitations in how this device population has been evaluated, and those limitations bear directly on agentic successors.

## 9.2 European Union: the AI Act

The EU AI Act (European Parliament and Council, 2024) classifies AI systems intended for use in medical contexts influencing clinical decisions as high-risk, requiring conformity assessment, data governance, human oversight, accuracy and robustness standards, and post-market monitoring. The requirement of Article 14, that high-risk systems be designed to permit effective human oversight including the ability to interrupt or override, is a structural constraint on agentic design rather than a procedural one: it excludes architectures in which an action becomes irreversible before a human can intervene.

## 9.3 International harmonisation

IMDRF guidance on machine learning-enabled medical devices supplies a common regulatory vocabulary across major jurisdictions (IMDRF, 2022). WHO guidance on the ethics and governance of AI for health (WHO, 2021) articulates complementary normative principles: protecting autonomy, promoting well-being, ensuring transparency, fostering responsibility, and guaranteeing inclusiveness, equity and sustainability.

Neither instrument yet addresses the compositional character of agentic systems, and this is the respect in which they are hardest to regulate. A medical device is authorised as an artefact with a

defined intended use. An agentic system is an assembly: a foundation model trained by one party, a retrieval corpus curated by a second, tools maintained by a third, and orchestration logic written by a fourth, any of which may change without the others being informed. Three consequences follow, none of which existing frameworks resolve.

First, the object of authorisation is unclear. If the underlying model is updated by its provider, the authorised system has changed in a way its deployer may not observe, let alone control—a situation the Predetermined Change Control Plan pathway anticipates for the developer's own updates but not for those of an upstream supplier. Second, responsibility for an output is distributed across components that may be separately authorised, or not authorised at all: general-purpose models used within clinical workflows frequently fall outside device regulation entirely (Gilbert et al., 2023). Third, the properties regulators most need to assess—the action space, the oversight mechanism, the error-handling behaviour—are properties of the assembly rather than of any component, and are therefore invisible to component-level review.

We suggest that autonomy level, being a property of the assembly as deployed and declarable by the deployer, is a more tractable regulatory anchor than either the algorithm class or the component inventory, for reasons developed below.

## 9.4 Autonomy level as a regulatory variable

The autonomy scale of Section 4.2 maps onto these frameworks more directly than the algorithmic categories currently used. Existing risk stratification turns on intended use and on the seriousness of the clinical situation; it does not distinguish a system whose recommendation a clinician evaluates in full from one whose recommendation arrives at the end of a twenty-step

reasoning chain that no clinician will reconstruct. Both are nominally under human oversight, and only the first actually is.

The distinction the EU AI Act draws—that high-risk systems must permit effective human oversight and the ability to interrupt—is best read as a constraint at the L3/L4 boundary, where actions cease to be reversible before reaching the patient. Below that boundary, oversight requirements are about whether the human can meaningfully evaluate what is presented, which is an interface and workload question rather than a control question. Framing regulatory obligations in these terms would have two practical consequences: it would make explicit that L2 systems carry an oversight burden that current approval routes do not test for, and it would give developers a stable target, since a system's autonomy level is a design decision that can be declared and audited, whereas its algorithm class changes with every model update.

# 10 Future research agenda

Drawing on the gaps identified above, we propose eight research priorities, ordered by our assessment of urgency.

## 10.1 Prospective multi-site validation

The predominance of retrospective, single-site, developer-conducted evaluation is the most consequential gap in the field, and Table 6 shows how thin the prospective evidence base remains. Priorities are: multi-site prospective trials with agent-specific endpoints; independent external validation consortia not funded by developers; and open benchmarks for multi-step clinical reasoning that are held out from training corpora.

## 10.2 Reporting standards for agentic systems

CONSORT-AI (Liu et al., 2020), DECIDE-AI (Vasey et al., 2022) and TRIPOD+AI (Collins et al., 2024) collectively cover trials, early live evaluation and prediction models, but none accommodates systems that select actions, invoke tools and operate over multiple turns. The consequence is concrete: from a typical published report of an agentic system it is at present impossible to determine what the system was permitted to do, which tool versions produced its results, or what happened when it erred—so the study cannot be replicated, and its safety profile cannot be assessed.

We regard closing this gap as the single most tractable contribution the methodological community could make in the near term, and offer Table 9 as a starting specification rather than a finished instrument. It is organised as an extension applicable alongside whichever existing guideline fits the study design, and every item is one whose absence we encountered while charting.

**Table 9** Proposed reporting items for studies of agentic AI systems, for use as an extension to CONSORT-AI, DECIDE-AI or TRIPOD+AI according to study design. Items marked [a] have no counterpart in the existing guidelines

| No. | Domain | Item |
|---|---|---|
| A1[a] | Autonomy | State the autonomy level as deployed, with the coding justified against an explicit scale, and report separately any higher level the system is technically capable of |
| A2[a] | Autonomy | Specify whether actions can reach the patient before human review, and if so which |
| A3 | Action space | Enumerate the actions available to the agent, or characterise the space and state how it is bounded |
| A4[a] | Tools | List every tool, database and sub-model the agent may invoke, with version and access date; state whether tool outputs were validated |
| A5 | Model | Report the underlying model, version, decoding parameters and any fine-tuning, together with the date of access |
| A6[a] | Orchestration | Describe the control loop, the termination criteria, and for multi-agent systems the communication topology and conflict-resolution rule |
| A7 | Oversight | State who supervises, what they are shown, how long they have, and by what mechanism they may override or interrupt |
| A8[a] | Errors | Report the number and nature of agent errors during the study, how each was detected, and by whom |

| No. | Domain | Item |
|---|---|---|
| A9[a] | Trajectory | Report task completion and tool-use fidelity, not only terminal accuracy; state whether reasoning traces were retained and whether they are available |
| A10 | Safety behaviour | Report the rate of escalation, refusal and clarification requests, against the cases in which each was warranted |
| A11 | Consistency | Report stability of output for repeated identical inputs, and across the study period |
| A12 | Calibration | Report agreement between expressed confidence and observed accuracy |
| A13 | Equity | Report performance stratified by the demographic groups relevant to the setting, with the strata pre-specified |
| A14 | Data flows | State what patient data the agent transmitted to which components, and under what authorisation |
| A15[a] | Provenance | State whether evaluation data could have appeared in the training corpus of the underlying model, and how this was assessed |

Items A1, A2, A4, A6, A8, A9 and A15 have no counterpart in existing guidance, and each corresponds to a failure mode in Table 8 that cannot otherwise be detected by a reader. A15 deserves particular emphasis: when the evaluation corpus may lie inside the training corpus, reported performance carries no information about generalisation, and for systems built on commercially trained foundation models this cannot be excluded by the investigators without a positive statement from the model provider.

Developing this into a usable instrument requires the process such guidelines normally follow—a Delphi exercise with methodologists, developers, clinicians and regulators, followed by piloting—rather than a proposal from two authors. We offer it in that spirit.

## 10.3 Equity-centred agent design

Stratified reporting of performance across demographic subgroups remains the exception rather than the rule. Equity analysis should be embedded across the development lifecycle: equity-aware data curation, fairness constraints in reward design, subgroup-stratified benchmarks and post-deployment disparity monitoring. The finding that physician attitudes towards AI vary systematically across professional groups (De Nicolò and La Torre, 2024b) motivates

complementary user-centred research, since a tool used unevenly produces uneven benefit however fair its outputs.

## 10.4 Human-agent teaming

Effective deployment requires deliberate design of the human-agent interface rather than treating it as a presentation layer. This is the priority we regard as most underserved relative to its importance, because it is the one on which the viability of L2 autonomy depends: if a supervisor cannot evaluate an agent's trajectory within the time a clinical decision allows, then intermediate autonomy offers the accountability of oversight without its substance.

The research questions here belong to cognitive science as much as to AI. How much of a reasoning trace can a practitioner evaluate under time pressure, and what representations make trajectories assessable at a glance—summary, counterfactual, uncertainty over steps, or flagged divergence from expectation? How should an agent's confidence be expressed so that reliance tracks accuracy, given that the human factors literature documents both over-reliance and disuse as failure modes of miscalibrated trust (Parasuraman and Riley, 1997; Goddard et al., 2012)? How does trust calibrate over months of use, and does it recover after an observed error? What shared mental model does a clinician form of an agent's competence boundary, and by what cues?

Priorities therefore include: handoff protocols for high-stakes decisions; trajectory representations empirically shown to support rejection of incorrect recommendations, which is the operative test rather than subjective satisfaction; validated instruments for shared mental model assessment; and longitudinal study of trust calibration under realistic workload. Frameworks for the non-adoption and abandonment of health technologies (Greenhalgh et al.,

2017) supply an analytic scaffold that the agentic literature has largely ignored, and the deployment study of Beede et al. (2020) remains among the few demonstrations of how much is learned by observing use rather than measuring output.

## 10.5 Mechanistic interpretability and formal safety verification

Interpretability for agentic systems must move beyond post-hoc attribution to mechanistic accounts of multi-step reasoning. Formal methods offer a complementary route: conformal prediction supplies distribution-free coverage guarantees under exchangeability (Angelopoulos and Bates, 2023), and model checking can enforce behavioural constraints, together enabling agents that are structurally incapable of executing certain unsafe action sequences. Verified constraints on the action space are more tractable, and arguably more valuable, than verified explanations of the reasoning.

## 10.6 Continual learning under regulatory constraint

Static models degrade as practice and populations evolve, yet unconstrained updating is incompatible with device authorisation. The research problem is therefore joint: continual learning methods that resist catastrophic forgetting, paired with change-control mechanisms that a regulator can evaluate in advance. The Predetermined Change Control Plan pathway indicates the direction; the technical methods to exploit it fully do not yet exist.

## 10.7 Multimodal foundation models for medicine

The next generation of medical agents requires foundation models that natively process text, images, audio, genomic sequences, physiological time series and structured records within a unified representation (Moor et al., 2023). Current siloed handling introduces information loss at

cross-modal interfaces. Priorities include unified biomedical tokenisation, cross-modal alignment pretraining, and evaluation on tasks that genuinely require multiple modalities rather than permitting a single dominant one.

### 10.8 Clinical world models and emerging computational substrates

Learned simulators of patient physiology would let agents evaluate the consequences of decisions before executing them—the counterfactual capability that separates planning from pattern matching. The obstacles are the physical plausibility of simulated trajectories, generalisation across subpopulations, and integration with planning algorithms.

Further ahead, hybrid quantum-classical architectures have shown feasibility on clinical classification tasks (Frasca et al., 2025), and quantum approaches to search and optimisation may in principle allow agents to explore larger hypothesis spaces in drug discovery and treatment optimisation. We include this direction with an explicit caveat: current demonstrations are small-scale, the benchmark results are sensitive to evaluation protocol, and no quantum advantage has been established for any clinical task. It is a research frontier rather than a near-term priority, and we rank it last accordingly.

## 11 Discussion

This review mapped the convergence of agent theory with medical application and found a field in transition from proof-of-concept to early deployment, but unevenly so. Returning to the research questions: the theoretical foundations (RQ1) are mature and shared with general-purpose AI, which means medical agents inherit both the capabilities and the failure modes of the systems beneath them. The architectures (RQ2) converge on a common four-layer stack,

which supports defining evaluation and governance instruments at the level of the stack rather than per application.

## 11.1 The autonomy inversion

Applying the autonomy scale of Section 4.2 to the evidence base yields the organising finding of this review, which we term the *autonomy inversion*. The systems with the greatest demonstrated patient benefit are those with the highest autonomy and the narrowest scope. The domains attracting the greatest research activity sit at intermediate autonomy, with the broadest scope and the weakest evidence (RQ3, RQ4). Closed-loop glucose control and autonomous retinopathy screening act without per-case review, and both have prospective randomised or pivotal evidence. Multi-agent clinical reasoning and conversational diagnosis retain a human approver, and have none.

Figure 3 plots each application domain of Table 4 against its autonomy level and the strongest form of evidence identified for it. The two clusters are the finding. Systems at L3 and L4, all operating in bounded domains, occupy the upper right: prospective and in three cases randomised evidence. Systems at L2, all addressing open-ended reasoning, occupy the lower left: benchmark and retrospective evidence, with a single exception. If capability and evidence advanced together, the points would lie along a rising diagonal from the lower left; instead the relationship runs through scope, not through time.

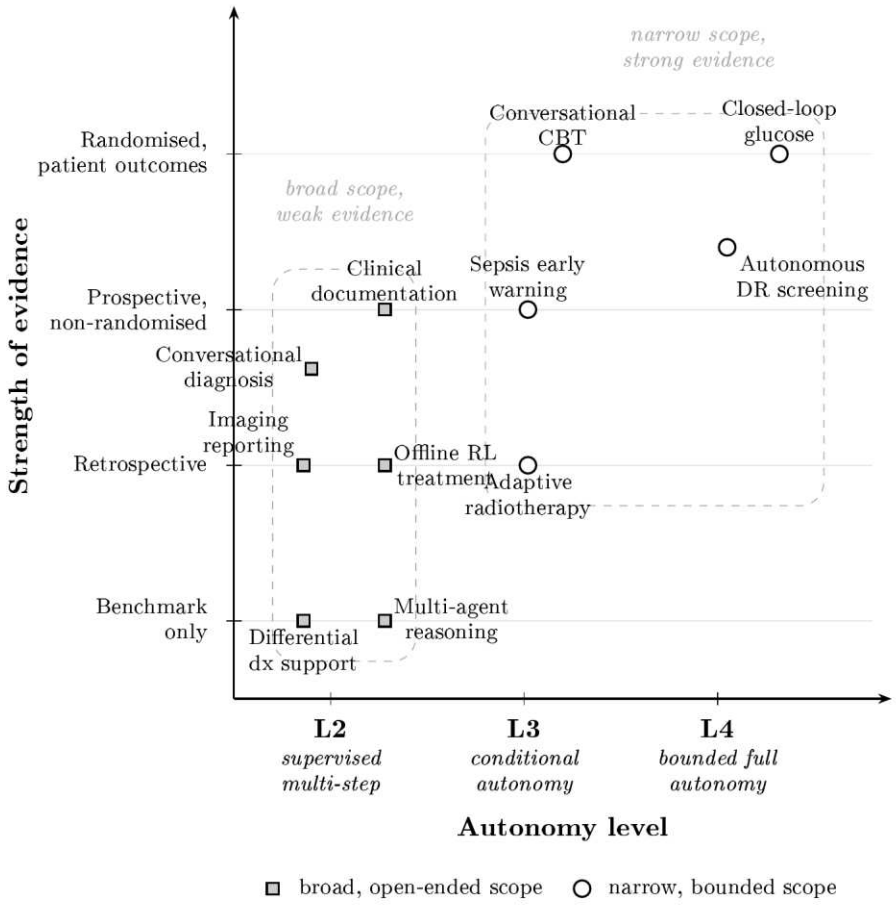



**Fig. 3** The autonomy inversion. Each application domain of Table *4* is plotted against its typical autonomy level (Table *3*) and the strongest form of evidence identified in this review. Square markers denote domains addressing broad, open-ended clinical reasoning; circular markers denote narrow, bounded tasks. Marker shape rather than colour carries the distinction, and dashed envelopes indicate the two clusters. Systems granted the most autonomy hold the strongest evidence and the narrowest scope; systems retaining a human approver hold the weakest evidence and the broadest scope. Conversational CBT is the one narrow-scope domain at intermediate-to-high autonomy with randomised evidence, and is discussed in the text

The constituent observations are not new. Collaco et al. (2026a) report that clinically applied agentic systems are few and mostly exploratory; Karunanayake (2026) reports that prospective evaluation has not kept pace with deployment; the maturity of closed-loop control and autonomous screening is uncontested. What has not previously been stated is that these facts stand in a systematic relation to one another, and that the relation is explained by autonomy level rather than by chronology.

The inversion is not accidental, and reading it as a mere maturity gap would be a mistake. Autonomy was granted to the L4 systems precisely because their domains are bounded enough for the evidentiary burden to be dischargeable: a single continuously measurable target variable, an unambiguous ground truth, a small action space. The L2 systems are attractive to researchers for the opposite reason—they address open-ended clinical reasoning—and that openness is what makes their evaluation intractable under current instruments. The nominal human approver, meanwhile, does not resolve the problem so much as relocate it: as Section 7 argues, an approver who cannot reconstruct a twenty-step reasoning chain provides the appearance of oversight rather than its substance, which is why automation bias is the dominant risk at exactly the level where the field is concentrating its effort.

This suggests a reframing of what progress in medical agentic AI would look like. The field's implicit trajectory is towards greater autonomy in broader domains. The evidence assembled here suggests instead that the near-term returns lie in two other directions: extending the L3–L4 pattern to further narrow, bounded tasks where the evidentiary burden is dischargeable, and developing the evaluation instruments (Table 7) and interface designs that would make L2 autonomy genuinely supervisable rather than nominally so.

On safety and governance (RQ5), ten failure modes recur, and their distribution across the architectural stack is itself informative: those addressable at the tool and orchestration layers admit mechanical guarantees and are therefore tractable, while automation bias and equity failure are socio-technical, invisible to every benchmark in current use, and correspondingly neglected. Regulation has begun to accommodate model updating but not distributed multi-agent responsibility, and is indexed to intended use rather than to autonomy.

The agenda (RQ6) follows from these gaps, and we would emphasise its second item over its first in terms of tractability: prospective multi-site validation is the most needed development, but agreed reporting standards for agentic systems are the most achievable, and would improve the interpretability of the evidence that does accumulate.

## 11.2 Limitations

Several limitations apply. First, this is a scoping review: consistent with PRISMA-ScR guidance we did not perform formal risk-of-bias assessment of individual sources, so the reported performance figures reproduced here should be read as claims made by primary sources rather than as validated estimates. Second, although screening was performed independently in duplicate, no quantitative measure of inter-rater agreement was computed, so the consistency of selection cannot be independently verified; relatedly, funding of the individual included sources was not extracted during charting, and PRISMA-ScR item 22 is therefore only partially satisfied. Third, the pace of publication in this area means that relevant work will have appeared after our search closed; we mitigated this with an updated search but cannot eliminate it. Fourth, restriction to English-language sources may under-represent work from non-anglophone health systems, which is a particular concern for a review that argues for equity-aware evaluation. Fifth, narrative synthesis of evidence ranging from formal theory to randomised trials necessarily involves judgement, and the article cites a purposively selected subset of the charted corpus. The selection criteria are stated in Section 2, but the charted corpus is not deposited, so a reader cannot independently verify that the cited subset is representative of it. Sixth, this review was conducted concurrently with several others on overlapping material (Collaco et al., 2026a; Njei et al., 2026; Xu et al., 2026; Collaco et al., 2026b); we have set out in Section 2.5 how our eligibility criteria differ and where our findings converge with theirs, but no coordination

between these efforts was possible, and readers seeking an inventory of clinically applied systems specifically should consult Collaco et al. (2026a) alongside this work. Finally, the proposed agenda reflects the authors' assessment of priority and should be complemented by community-wide priority-setting exercises such as consensus studies with clinicians, patients and regulators. The autonomy scale we propose is likewise a conceptual instrument rather than a validated one: its inter-rater reliability when applied to systems by independent assessors has not been established, and doing so is itself a research task. Relatedly, because the scale was derived from the corpus rather than applied during charting (Section 2), autonomy level was assigned to application domains rather than to individual sources; the autonomy inversion is accordingly a statement about the distribution of domains and their evidence, and a per-source coding of the charted corpus would put it on a firmer quantitative footing.

# 12 Conclusion

Agentic AI—systems that perceive, reason, plan and act over multi-step clinical tasks—is developing rapidly, and this review has catalogued five architectural paradigms and their applications across diagnostic imaging, clinical decision support, neurology and psychiatry, oncology, drug discovery, digital twins and public health. We grounded that landscape in agent theory, planning frameworks, multi-agent coordination and offline reinforcement learning, and contributed a six-level autonomy scale, seven evaluation dimensions specific to agency, a mapping of evaluation requirements to autonomy level, and a ten-item failure mode taxonomy assigned to the architectural layer at which each is addressable.

The dominant finding is the autonomy inversion: the systems with demonstrated patient benefit act with high autonomy in narrow, bounded domains, while those attracting the greatest research

investment retain a nominal human approver, address open-ended reasoning, and have no prospective outcome evidence at all. Closing that gap is the field's principal task, and it is a methodological task before it is a technical one. It calls for prospective multi-site validation, reporting standards adapted to agency, evaluation requirements indexed to autonomy rather than to algorithm, stratified reporting across subgroups, and regulatory frameworks that can accommodate systems which act rather than merely predict. It also calls for a degree of restraint: the most defensible near-term progress may lie in extending bounded autonomy to further well-characterised tasks, rather than in broadening the scope of systems whose supervision is already nominal.

Realising the potential of agentic AI in medicine safely requires that enthusiasm for capability be matched by rigour in evaluation, care in deployment, equity in design, and a sustained commitment to patient safety.

# Statements and Declarations

## Funding

The authors did not receive support from any organization for the submitted work.

## Competing interests

All authors have no relevant financial or non-financial interests to disclose.

## Ethics approval

Not applicable. This review synthesises published literature and did not involve human participants or animals.

## Consent to participate / Consent for publication

Not applicable.

## Data availability

No new data were generated by this study. All sources analysed are published and are identified in the reference list. The completed PRISMA-ScR checklist and the complete database-specific search strings are provided as Online Resources 1 and 2 and are available without restriction. The charted dataset compiled during the review is available from the corresponding author on reasonable request. The registered review protocol, the completed PRISMA-ScR checklist and the full database-specific search strategies are additionally deposited with the Open Science Framework and are openly accessible at https://doi.org/10.17605/OSF.IO/WY8BN (De Nicolò, 2026).

## Materials and code availability

Not applicable. This review did not generate materials or code. The variables recorded on the charting form are listed in Section 2.

## Author contributions

All authors contributed to the conception and design of the review. V.D.N., D.L.T. and I.I. performed the literature search, screening and data charting. V.D.N. drafted the manuscript.

All authors revised the manuscript critically for important intellectual content, approved the version to be published, and agree to be accountable for all aspects of the work.

## Use of generative AI

As stated in Section 2, generative AI tools were used for language editing and for checking bibliographic consistency. They were not used to select sources, to chart data, to generate results or to draft scientific content. The authors take full responsibility for the accuracy and integrity of the text, and confirm that the edits reflect their original work.

## Ethical responsibilities of authors

The work described has not been published before and is not under consideration for publication elsewhere. Its submission, the author group and the order of authors are approved by all authors. The manuscript reports the review honestly and without fabrication or selective presentation, and cites the concurrent reviews of this literature explicitly (Section 1.1). The authors' own prior publications are cited where substantively relevant and constitute a small minority of the reference list.

# References


Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC (2018) Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digit Med 1:39. https://doi.org/10.1038/s41746-018-0040-6

Abramson J, Adler J, Dunger J, et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630:493–500. https://doi.org/10.1038/s41586-024-07487-w

Adams R, Henry KE, Sridharan A, et al (2022) Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 28:1455–1460. https://doi.org/10.1038/s41591-022-01894-0

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv:1606.06565

Angelopoulos AN, Bates S (2023) Conformal prediction: a gentle introduction. Found Trends Mach Learn 16:494–591

Arksey H, O'Malley L (2005) Scoping studies: towards a methodological framework. Int J Soc Res Methodol 8:19–32

Attanasio A, Scaglioni B, De Momi E, Fiorini P, Valdastri P (2021) Autonomy in surgical robotics. Annu Rev Control Robot Auton Syst 4:651–679. https://doi.org/10.1146/annurev-control-062420-090543

Bedi S, Liu Y, Orr-Ewing L, et al (2025) Testing and evaluation of health care applications of large language models: a systematic review. JAMA. https://doi.org/10.1001/jama.2024.21700

Beede E, Baylor E, Hersch F, et al (2020) A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp 1–12

Boiko DA, MacKnight R, Kline B, Gomes G (2023) Autonomous chemical research with large language models. Nature 624:570–578. https://doi.org/10.1038/s41586-023-06792-0

Brown SA, Kovatchev BP, Raghinaru D, et al (2019) Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes. N Engl J Med 381:1707–1717. https://doi.org/10.1056/NEJMoa1907863

Brown TB, Mann B, Ryder N, et al (2020) Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33, pp 1877–1901

Char DS, Shah NH, Magnus D (2018) Implementing machine learning in health care — addressing ethical challenges. N Engl J Med 378:981–983. https://doi.org/10.1056/NEJMp1714229

Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M (2021) Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 4:123–144

Collaco BG, Haider SA, Prabha S, Gomez-Cabello CA, Genovese A, Wood NG, et al (2026a) The role of agentic artificial intelligence in healthcare: a scoping review. npj Digit Med 9:345. https://doi.org/10.1038/s41746-026-02517-5

Collaco BG, Gomez-Cabello CA, Wood NG, Hester EO, Forte AJ (2026b) Explainable agentic artificial intelligence in healthcare: a scoping review. Bioengineering 13:513. https://doi.org/10.3390/bioengineering13050513

Collins GS, Moons KGM, Dhiman P, et al (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378

Dayan I, Roth HR, Zhong A, et al (2021) Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 27:1735–1743. https://doi.org/10.1038/s41591-021-01506-3

De Nicolò V (2026) AI agents and agentic AI in medicine: an autonomy-centred scoping review and research agenda. Open Science Framework [registration]. https://doi.org/10.17605/OSF.IO/WY8BN

De Nicolò V, La Torre D (2024a) Global research trends of artificial intelligence and machine learning applied in medicine: a bibliometric analysis (2012–2022). In: Artificial Intelligence for Medicine: An Applied Reference for Methods and Applications. Elsevier, pp 151–177

De Nicolò V, La Torre D (2024b) Italian physicians' trust towards artificial intelligence: a multicentric cross-sectional study. In: 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS). IEEE, pp 1621–1625

Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I (2023) Improving factuality and reasoning in language models through multiagent debate. arXiv:2305.14325

European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union L, 2024/1689, 12 July 2024. http://data.europa.eu/eli/reg/2024/1689/oj

US Food and Drug Administration (2024) Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and Food and Drug Administration staff. Silver Spring, MD, issued 4 December 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence

US Food and Drug Administration (2025) Artificial intelligence-enabled medical devices [device list]. Silver Spring, MD. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices

Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS (2019) Adversarial attacks on medical machine learning. Science 363:1287–1289. https://doi.org/10.1126/science.aaw4399

Finlayson SG, Subbaswamy A, Singh K, et al (2021) The clinician and dataset shift in artificial intelligence. N Engl J Med 385:283–286. https://doi.org/10.1056/NEJMc2104626

Fitzpatrick KK, Darcy A, Vierhile M (2017) Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 4:e19. https://doi.org/10.2196/mental.7785

Frasca M, Cutica I, Pravettoni G, La Torre D (2025) Optimizing melanoma diagnosis: a hybrid deep learning and quantum computing approach for enhanced lesion classification. Intell Based Med 12:100264

Frasca M, Gazzaniga G, Graziosi A, et al (2026) Artificial intelligence and precision medicine: a pilot study predicting optimal ceftaroline dosage for pediatric patients. Front Artif Intell 8:1702087

Gao S, Fang A, Huang Y, et al (2024) Empowering biomedical discovery with AI agents. Cell 187:6990–7009. https://doi.org/10.1016/j.cell.2024.09.022

Gemini Team (2023) Gemini: a family of highly capable multimodal models. arXiv:2312.11805

Gerke S, Minssen T, Cohen G (2020) Ethical and legal challenges of artificial intelligence-driven healthcare. In: Artificial Intelligence in Healthcare. Elsevier, pp 295–336

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

Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19:121–127. https://doi.org/10.1136/amiajnl-2011-000089

Gottesman O, Johansson F, Komorowski M, et al (2019) Guidelines for reinforcement learning in healthcare. Nat Med 25:16–18. https://doi.org/10.1038/s41591-018-0310-5

Greenhalgh T, Wherton J, Papoutsi C, et al (2017) Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 19:e367. https://doi.org/10.2196/jmir.8775

Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M (2023) Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, pp 79–90

Henry KE, Adams R, Parent C, et al (2022) Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 28:1447–1454. https://doi.org/10.1038/s41591-022-01895-z

International Medical Device Regulators Forum (2022) Machine learning-enabled medical devices: key terms and definitions. IMDRF/AIMD WG/N67 Final:2022. https://www.imdrf.org/documents/machine-learning-enabled-medical-devices-key-terms-and-definitions

Isaksson LJ, Mastroleo F, Vincini MG, et al (2024) The emerging role of artificial intelligence in proton therapy: a review. Crit Rev Oncol Hematol 204. PMID 39233128

Jiang Y, Black KC, Geng G, et al (2025) MedAgentBench: a realistic virtual EHR environment to benchmark medical LLM agents. arXiv:2501.14654

Johri S, Jeong J, Tran BA, et al (2025) An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. https://doi.org/10.1038/s41591-024-03328-5

Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochastic domains. Artif Intell 101:99–134

Kairouz P, McMahan HB, Avent B, et al (2021) Advances and open problems in federated learning. Found Trends Mach Learn 14:1–210

Karunanayake N (2026) When chatbots become agents: the next phase of healthcare AI. J Med Syst 50:80. https://doi.org/10.1007/s10916-026-02402-4

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17:195. https://doi.org/10.1186/s12916-019-1426-2

Kim Y, Park C, Jeong H, et al (2024) MDAgents: an adaptive collaboration of LLMs for medical decision-making. In: Advances in Neural Information Processing Systems 37

Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA (2018) The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 24:1716–1720. https://doi.org/10.1038/s41591-018-0213-5

Kostrikov I, Nair A, Levine S (2022) Offline reinforcement learning with implicit Q-learning. In: International Conference on Learning Representations

Kumar A, Zhou A, Tucker G, Levine S (2020) Conservative Q-learning for offline reinforcement learning. In: Advances in Neural Information Processing Systems 33, pp 1179–1191

Levac D, Colquhoun H, O'Brien KK (2010) Scoping studies: advancing the methodology. Implement Sci 5:69. https://doi.org/10.1186/1748-5908-5-69

Levine S, Kumar A, Tucker G, Fu J (2020) Offline reinforcement learning: tutorial, review, and perspectives on open problems. arXiv:2005.01643

Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems 33, pp 9459–9474

Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26:1364–1374. https://doi.org/10.1038/s41591-020-1034-x

M. Bran A, Cox S, Schilter O, Baldassari C, White AD, Schwaller P (2024) Augmenting large language models with chemistry tools. Nat Mach Intell 6:525–535. https://doi.org/10.1038/s42256-024-00832-8

Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A (2024) Evaluating large language models as agents in the clinic. npj Digit Med 7:84. https://doi.org/10.1038/s41746-024-01083-y

Mnih V, Kavukcuoglu K, Silver D, et al (2015) Human-level control through deep reinforcement learning. Nature 518:529–533. https://doi.org/10.1038/nature14236

Moor M, Banerjee O, Abad ZSH, et al (2023) Foundation models for generalist medical artificial intelligence. Nature 616:259–265. https://doi.org/10.1038/s41586-023-05881-4

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

Njei B, Al-Ajlouni YA, Kanmounye US, et al (2026) Artificial intelligence agents in healthcare research: a scoping review. PLOS ONE 21:e0342182. https://doi.org/10.1371/journal.pone.0342182

Nori H, King N, McKinney SM, Carignan D, Horvitz E (2023) Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375

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

Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R (2023) Large language models propagate race-based medicine. npj Digit Med 6:195. https://doi.org/10.1038/s41746-023-00939-z

OpenAI (2023) GPT-4 technical report. arXiv:2303.08774

Ouyang L, Wu J, Jiang X, et al (2022) Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems 35, pp 27730–27744

Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. Hum Factors 39:230–253

Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A Syst Hum 30:286–297

Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp 1–22

Price WN, Cohen IG (2019) Privacy in the age of medical big data. Nat Med 25:37–43. https://doi.org/10.1038/s41591-018-0272-7

Price WN, Gerke S, Cohen IG (2019) Potential liability for physicians using artificial intelligence. JAMA 322:1765–1766. https://doi.org/10.1001/jama.2019.15064

Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. Nat Med 28:31–38. https://doi.org/10.1038/s41591-021-01614-0

Repetto M, Colapinto C, Jayaraman R, Appio FP, La Torre D (2025) Blockchain-enabled predictive digital twin approach for healthcare: enhancing accuracy and performance with federated learning. Int J Prod Econ 294:109768. https://doi.org/10.1016/j.ijpe.2025.109768

Ricetti C, Carrara L, La Torre D (2025) The potential of machine learning in diagnosing neurological and psychiatric diseases: a review. Discov Artif Intell 5:105

Rieke N, Hancox J, Li W, et al (2020) The future of digital health with federated learning. npj Digit Med 3:119. https://doi.org/10.1038/s41746-020-00323-1

Russell S, Norvig P (2020) Artificial Intelligence: A Modern Approach, 4th edn. Pearson Education

Russell SJ, Beck RW, Damiano ER, et al (2022) Multicenter, randomized trial of a bionic pancreas in type 1 diabetes. N Engl J Med 387:1161–1172. https://doi.org/10.1056/NEJMoa2205225

Saeidi H, Opfermann JD, Kam M, et al (2022) Autonomous robotic laparoscopic surgery for intestinal anastomosis. Sci Robot 7:eabj2908. https://doi.org/10.1126/scirobotics.abj2908

Schick T, Dwivedi-Yu J, Dessì R, et al (2023) Toolformer: language models can teach themselves to use tools. In: Advances in Neural Information Processing Systems 36

Schmidgall S, Ziaei R, Harris C, Reis E, Jopling J, Moor M (2024) AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv:2405.07960

Sendak MP, Gao M, Brajer N, Balu S (2020) Presenting machine learning model information to clinical end users with model facts labels. npj Digit Med 3:41. https://doi.org/10.1038/s41746-020-0253-3

Sheller MJ, Edwards B, Reina GA, et al (2020) Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 10:12598. https://doi.org/10.1038/s41598-020-69250-1

Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y (2023) HuggingGPT: solving AI tasks with ChatGPT and its friends in Hugging Face. In: Advances in Neural Information Processing Systems 36

Shi W, Xu R, Zhuang Y, et al (2024) EHRAgent: code empowers large language models for few-shot complex tabular reasoning on electronic health records. In: Proceedings of EMNLP 2024

Shinn N, Cassano F, Berman E, Gopinath A, Narasimhan K, Yao S (2023) Reflexion: language agents with verbal reinforcement learning. In: Advances in Neural Information Processing Systems 36

Shoham Y, Leyton-Brown K (2009) Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press, Cambridge

Silver D, Huang A, Maddison CJ, et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489. https://doi.org/10.1038/nature16961

Singhal K, Azizi S, Tu T, et al (2023) Large language models encode clinical knowledge. Nature 620:172–180. https://doi.org/10.1038/s41586-023-06291-9

Sutton RS, Barto AG (2018) Reinforcement Learning: An Introduction, 2nd edn. MIT Press, Cambridge

Tang X, Zou A, Zhang Z, et al (2024) MedAgents: large language models as collaborators for zero-shot medical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2024

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29:1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56. https://doi.org/10.1038/s41591-018-0300-7

Tricco AC, Lillie E, Zarin W, et al (2018) PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 169:467–473. https://doi.org/10.7326/M18-0850

Tu T, Palepu A, Schaekermann M, et al (2024) Towards conversational diagnostic AI. arXiv:2401.05654

Umapathi LK, Pal A, Sankarasubbu M (2023) Med-HALT: medical domain hallucination test for large language models. In: Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pp 314–334

Van Veen D, Van Uden C, Blankemeier L, et al (2024) Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 30:1134–1142. https://doi.org/10.1038/s41591-024-02855-5

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. Nat Med 28:924–933. https://doi.org/10.1038/s41591-022-01772-9

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in Neural Information Processing Systems 30, pp 5998–6008

Wang X, Wei J, Schuurmans D, et al (2023) Self-consistency improves chain of thought reasoning in language models. In: International Conference on Learning Representations

Wang L, Ma C, Feng X, et al (2024) A survey on large language model based autonomous agents. Front Comput Sci 18:186345. https://doi.org/10.1007/s11704-024-40231-1

Wei J, Wang X, Schuurmans D, et al (2022) Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems 35, pp 24824–24837

World Health Organization (2021) Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. WHO, Geneva

Wing JM (2021) Trustworthy AI. Commun ACM 64:64–71. https://doi.org/10.1145/3448248

Wooldridge M (2009) An Introduction to MultiAgent Systems, 2nd edn. Wiley, Chichester

Wornow M, Xu Y, Thapa R, et al (2023) The shaky foundations of large language models and foundation models for electronic health records. npj Digit Med 6:135. https://doi.org/10.1038/s41746-023-00879-8

Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J (2021) How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 27:582–584. https://doi.org/10.1038/s41591-021-01312-x

Wu Q, Bansal G, Zhang J, et al (2024) AutoGen: enabling next-gen LLM applications via multi-agent conversation. In: Conference on Language Modeling (COLM)

Xi Z, Chen W, Guo X, et al (2023) The rise and potential of large language model based agents: a survey. arXiv:2309.07864

Xu G, Li X, Chen Y, Duan Y, Wu S, Yu H, et al (2026) A comprehensive survey of AI agents in healthcare. J Biomed Inform 179:105045. https://doi.org/10.1016/j.jbi.2026.105045

Yang GZ, Cambias J, Cleary K, et al (2017) Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy. Sci Robot 2:eaam8638. https://doi.org/10.1126/scirobotics.aam8638

Yang X, Chen A, PourNejatian N, et al (2022) A large language model for electronic health records. npj Digit Med 5:194. https://doi.org/10.1038/s41746-022-00742-2

Yao S, Zhao J, Yu D, et al (2022) ReAct: synergizing reasoning and acting in language models. arXiv:2210.03629

Yao S, Yu D, Zhao J, et al (2023) Tree of thoughts: deliberate problem solving with large language models. In: Advances in Neural Information Processing Systems 36

Zack T, Lehman E, Suzgun M, et al (2024) Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 6:e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X

Zakka C, Shad R, Chaurasia A, et al (2024) Almanac — retrieval-augmented language models for clinical medicine. NEJM AI. https://doi.org/10.1056/AIoa2300068