International Journal of Computer (IJC)
https://ijcjournal.org
ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

Designing Responsible AI Agent Systems for Clinical Applications

Sachin Bajpai*

Managing Director, PwC, Irvine, California, USA

** Corresponding author: sachin.bajpai@gmail.com*

Accepted: 20 May 2026 **Published:** 25 August 2026

Abstract

Clinical AI agents combine prediction, retrieval, tool use, and workflow action, creating safety risks that model-level evaluation alone does not capture. This article derives a responsible architecture through a structured narrative evidence synthesis and examines its control logic in an end-to-end sepsis-deterioration scenario. The synthesis yielded three cross-cutting requirements: separation of orchestration from bounded execution, risk-defined human approval and override, and continuous audit and post-deployment monitoring. In the sepsis walkthrough, the architecture blocked unauthorized actions, suppressed recommendations when inputs or evidence were insufficient, escalated an unacknowledged high-risk alert, and preserved the event trail required to reconstruct the decision. The scenario verifies controllability and auditability at design level. Prospective silent and early-stage clinical evaluation remain necessary to measure missed deterioration, inappropriate escalation, unsupported recommendations, clinician override, and time to intervention.

Keywords: AI agents, clinical decision support, responsible AI, human-in-the-loop, agentic AI, patient safety, orchestration, clinical workflow, healthcare governance, medical AI

1. Introduction

Clinical artificial intelligence has shifted from isolated prediction tools toward systems that plan tasks, call external tools, retrieve data, and coordinate multi-step workflows. Healthcare changes the design problem. A model that classifies an image or predicts risk has a narrow operating boundary. A clinical agent that retrieves patient records, selects tools, drafts recommendations, and escalates urgency enters the workflow where physicians, nurses, pharmacists, and case reviewers carry legal and professional accountability.

The article aims to define a responsible architecture for clinical AI agent systems that support medical decision-making under explicit human control. Three research objectives guide the work: to define the architectural distinctions among orchestration, execution, and human oversight in clinical AI agents; to determine how evidence, evaluation, and governance requirements translate into agent design; and to propose implementation logic for vendor-agnostic and vendor-specific clinical deployments.

The novelty of the study lies in treating human-in-the-loop control as a structural component of the agent system. Oversight spans task routing, tool access, evidence presentation, approval rights, and post-

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

deployment monitoring. The working hypothesis states that clinical AI agents become more controllable when system designers separate orchestration, execution, and human approval, then connect these layers through auditable workflow rules.

2. Materials and Methods

2.1. Review design and search strategy

A structured narrative evidence synthesis was used because the research question concerned architectural and governance requirements distributed across empirical studies, scoping reviews, reporting standards, consensus guidelines, and regulatory documents rather than the pooled effect of a single intervention. Search reporting followed the PRISMA-S items applicable to evidence synthesis, and the completed manuscript was checked against the SANRA criteria for narrative reviews [1, 2]. Searches were run in PubMed/MEDLINE, Scopus, Web of Science Core Collection, and IEEE Xplore on 1 August 2026 for records published from 1 January 2019 through the search date. The principal search combined ("agentic AI" OR "AI agent" OR "autonomous AI agent" OR "large language model agent" OR "multi-agent system") AND (healthcare OR clinical OR medicine) AND (safety OR governance OR evaluation OR audit OR "human-in-the-loop" OR oversight OR regulation OR "clinical decision support"). Supplementary searches combined the agent terms with sepsis, oncology, dataset transparency, multimodal model, tool use, escalation, and post-deployment monitoring. Reference lists of included reviews and consensus papers were searched backward.

2.2. Eligibility and screening

Eligible records were English-language peer-reviewed studies, consensus or reporting standards, and documents issued by a national regulator or international health organization. A source had to address clinical AI agents or a directly transferable control problem and report at least one of the following: system architecture, tool permissions, human oversight, data provenance, workflow evaluation, auditability, regulation, escalation, or post-deployment monitoring. Records were excluded when they concerned non-health applications, reported model accuracy without a workflow or governance implication, duplicated a later peer-reviewed version, lacked accessible full text, or consisted only of an editorial opinion without an explicit analytic basis.

2.3. Source appraisal and data extraction

Because the corpus contained heterogeneous evidence types, appraisal was matched to source function. Reviews were examined for declared databases, eligibility criteria, screening, and synthesis methods; empirical studies for case or dataset definition, comparator, outcome definition, and reproducibility; consensus and reporting documents for stakeholder breadth, development method, and explicit recommendations; and regulatory or organizational guidance for issuing authority, scope, and currency.

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

Sources were classified as high-weight, moderate-weight, or contextual. Contextual sources could define terminology or illustrate a use case but could not independently establish an architectural requirement. The extraction matrix recorded the clinical problem, system function, reported failure or risk, proposed control, lifecycle stage, evidence type, appraisal category, and architectural requirement supported.

2.4. Derivation of architectural requirements

Directed coding began with six domains derived from the review question: data provenance, orchestration and tool access, human authority, evaluation, auditability, and post-deployment monitoring. Additional codes were added when a source described a control not represented in the initial scheme. Codes were consolidated into an architectural requirement only when four conditions were met: support from at least two independent sources; support from at least one authoritative regulatory, consensus, or reporting source and one empirical study or review; expression as an implementable system-level control; and relevance to at least two clinical use cases or lifecycle stages. This procedure produced three cross-cutting requirements: separation of orchestration from bounded execution; explicit human approval, amendment, rejection, and override at risk-defined transition points; and continuous audit and post-deployment monitoring. Data lineage, evidence exposure, uncertainty display, and vendor independence were retained as enabling controls attached to these requirements. Supplementary Table S1 presents the source-to-requirement evidence matrix, including the evidentiary weight assigned to each source, the architectural requirement supported, whether that support was direct or indirect, and the additional governance control derived from the source.

2.5. Scenario-based design verification

The architecture was then examined through a structured sepsis-deterioration walkthrough and compared with a conventional monolithic clinical decision-support design. The comparator generated a risk score and a single alert but did not separate task routing, tool execution, evidence packaging, permission checks, or escalation ownership. The proposed architecture was tested against a baseline deterioration sequence and six injected failures: a stale vital-sign feed, missing laboratory data, a unit mismatch, retrieval of an unsupported recommendation, an attempted order outside the executor permission set, and an unacknowledged high-risk alert. For this engineering scenario, high-risk escalation was triggered by a risk value of at least 0.80 or an increase of at least 0.20 within 30 minutes. Input staleness beyond 15 minutes or missingness above 20% of required variables forced an insufficient-data state. Order entry and patient messaging required clinician approval, and an unacknowledged high-risk alert was routed to the backup clinician after 5 minutes. Prespecified pass conditions were zero executed permission violations, complete capture of required audit events, no recommendation without a traceable evidence source and patient-data basis, and delivery of a high-risk

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

alert within 60 seconds. Clinical evaluation outcomes were defined prospectively as inappropriate escalation, missed deterioration, unsupported recommendation rate, clinician override rate, and time from qualifying signal to clinical intervention.

3. Results

Recent healthcare AI literature treats agentic systems as a distinct design class because they pursue goals through tool access, planning, interaction, and task coordination. A scoping review of agentic AI in healthcare identifies autonomy, advanced reasoning, dynamic interaction, and multi-agent collaboration as features that separate agentic systems from narrow single-task tools [3]. Clinical settings make this distinction concrete. A clinical agent does more than output a score. It selects data, calls tools, drafts recommendations, escalates cases, and presents action options to a clinician.

The first result concerns architecture. A responsible clinical agent system needs three separated components: an orchestrator, executor agents, and a human-in-the-loop control layer. The orchestrator manages task routing, data access rules, tool permissions, and escalation logic. Executor agents perform bounded tasks such as EHR retrieval, guideline lookup, risk stratification, alert prioritization, documentation drafting, and order-set preparation. The human-in-the-loop layer gives clinicians a defined point for review, amendment, approval, rejection, or interruption. Recent agentic AI literature supports this separation because clinical usefulness arises from coordinated, specialized functions [3, 4].

Figure 1 adapts the agentic workflow logic from the healthcare scoping review and places human control inside the operating chain [3].

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

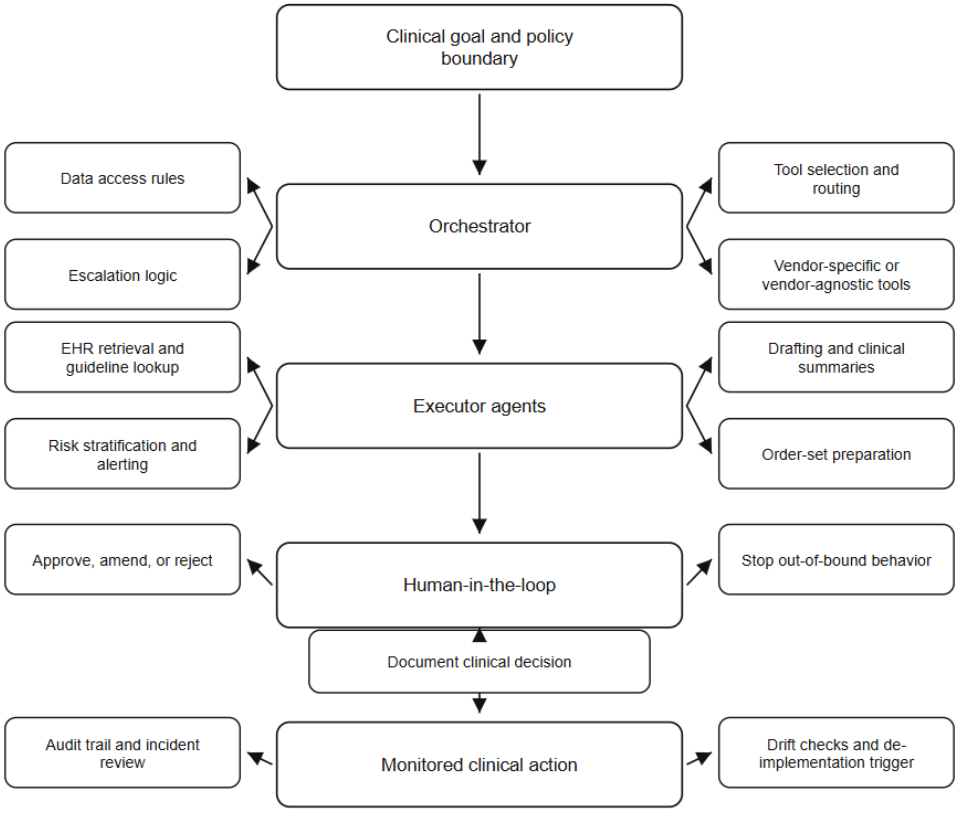

Figure 1: Responsible clinical AI agent architecture with controlled action paths [3]

Figure 1 separates the agent's technical work from clinical authority. A clinician cannot meaningfully supervise a final recommendation without seeing the task goal, retrieved evidence, selected tool, uncertainty boundary, and proposed action. A review button has limited value if the physician cannot inspect why the agent escalated a patient, ignored a datum, or preferred one guideline source over another.

The second result concerns data provenance. The STANDING Together recommendations address transparency, representativeness, and documentation of health datasets, as biased or poorly described data can perpetuate inequities in AI systems [5]. Clinical agents intensify this problem. They may combine electronic records, laboratory results, monitor streams, patient-reported symptoms, wearable data, imaging reports, and guideline repositories. One weak data channel can distort downstream reasoning. Agent design needs data lineage at the workflow level. Model documentation alone does not cover the agent's full reasoning chain.

Reporting standards lead to a similar conclusion. TRIPOD+AI guides reporting clinical prediction models that use regression or machine learning methods [6]. Clinical agents often embed such models inside broader task sequences. A sepsis agent, for example, may combine deterioration prediction, antibiotic timing prompts, fluid assessment cues, and escalation pathways. Each embedded component

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

needs documentation. The orchestrator needs a separate account for coordination rules. Without both layers, users see fragmented explanations rather than a traceable clinical pathway.

Early-stage evaluation guidance adds the human factors dimension. DECIDE-AI focuses on small-scale clinical evaluation, safety, workflow interaction, and reporting of how clinicians use AI-driven decision support [7]. Agent performance depends on more than benchmark accuracy. Clinicians must interpret alerts during busy shifts, assess whether the evidence is sufficient, and know when the agent has reached its permitted boundary. A technically accurate oncology recommendation still carries risk if the agent hides missing biomarker data or fails to identify the guideline source. Evaluation has to inspect the agent inside clinical work, with attention to user response, escalation timing, and override patterns.

The oncology literature gives a concrete example of tool-based agent reasoning. A 2025 study developed and validated an autonomous clinical AI agent for oncology that used GPT-4 together with multimodal precision oncology tools, document retrieval, and external knowledge resources [8]. The architectural lesson lies in the use of specialized tools under a coordinated workflow. High-risk clinical domains favor bounded tool orchestration, traceable retrieval, and component-level validation. The agent has to record which tool it used, what evidence it retrieved, and where the clinician entered the decision path.

Sepsis care shows a different pressure point. Reviews of AI in sepsis management report that machine learning systems support earlier detection and treatment planning by leveraging complex clinical data from electronic records. At the same time, the adoption faces barriers related to data diversity, clinician trust, interpretability, and regulatory readiness [9]. Sepsis suits agentic monitoring because vital signs, laboratory values, medication history, prior admissions, triage notes, and bedside data change across short time intervals. The same urgency raises the risk of automation errors. A responsible sepsis agent should escalate suspected deterioration, prepare evidence-based recommendations, and require clinician confirmation before action.

Trustworthy AI guidance translates ethical principles into design controls. FUTURE-AI frames trustworthy healthcare AI through fairness, universality, traceability, usability, robustness, and explainability [10]. Clinical agent design maps these principles onto practical controls. Fairness requires subgroup monitoring across data channels. Universality requires testing across local workflows, devices, and patient groups. Traceability requires event logs for tool calls, source retrieval, recommendation changes, and human decisions. Usability requires interfaces that reduce cognitive burden. Robustness requires safe degradation when data feeds fail. Explainability requires evidence-grounded reasoning that clinicians can inspect.

Regulatory guidance narrows the boundary between assistive and regulated clinical decision support. FDA guidance clarifies the scope of oversight for clinical decision support software. It explains that some functions fall outside device regulation when healthcare professionals can independently review

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

the basis for recommendations [11]. Agentic systems challenge that condition when they rely on hidden reasoning, inaccessible evidence, or automated action. Responsible architecture requires exposing the basis for recommendations and limiting automation based on clinical risk.

The WHO guidance on large, multimodal models in health addresses systems that accept multiple data types and generate outputs beyond the format of the original inputs [12]. Clinical agents increasingly fit this description because they combine text, structured clinical fields, sensor streams, imaging outputs, and guideline documents. Multimodal input expands clinical usefulness, but it spreads error sources across channels. A failure may come from the input, the fusion step, the retrieval source, the model's reasoning path, or the interface. Layered control provides each modality with a quality check, grants the orchestrator a permission rule, assigns each executor a bounded task, and provides clinicians with an inspection point.

The evidence matrix did not treat every source as an equivalent vote. Regulatory, consensus, and reporting documents defined the controls expected at the clinical and organizational boundary, whereas empirical and review evidence showed where those controls become operationally relevant. Tool specialization, permission boundaries, and traceable task routing recurred across the agentic and oncology evidence and formed the first requirement: orchestration must be separated from bounded execution. Independent review of the recommendation basis, observed clinician interaction, and risk-dependent intervention points supported the second requirement: clinical action must remain behind an explicit human approval and override boundary. Dataset lineage, event reconstruction, drift, missing-data events, and workflow response supported the third requirement: monitoring and audit must continue after deployment. A candidate requirement was excluded when it appeared only in a conceptual source or only in one use case. The resulting architecture reflects cross-source convergence under predefined retention rules.

4. Discussion

Clinical organizations need to treat the AI agent as a supervised participant in the workflow. This framing provides the agent with enough structure to support surveillance, retrieval, triage, and drafting, while clinicians retain authority at the point where a recommendation becomes clinical action. The design task starts with the pathway, since prior authorization, sepsis escalation, oncology toxicity monitoring, discharge follow-up, and documentation support differ in urgency, data reliability, and harm profile.

The same architecture serves several clinical settings only when the orchestrator adjusts permissions based on risk. A documentation agent may draft notes for clinician review. A sepsis agent may escalate alerts and prepare recommended actions. A chemotherapy-monitoring agent may contact the oncology

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

team and mark a patient for urgent evaluation. These examples require different approval points because denial of coverage, delayed antibiotics, and missed neutropenic fever result in distinct patient harms.

Table 1 compares clinical decision logic across selected use cases. It separates the agent's contribution from the human authority point and from the condition that interrupts the workflow.

Table 1: Decision logic for responsible clinical AI agent deployment

Clinical use case	Agent contribution	Human authority point	Stop condition	Audit object
Prior authorization	Gathers clinical criteria, payer rules, and patient record evidence	Qualified reviewer approves or denies the request	Missing evidence, conflict between the rule and the clinical record, and high-risk therapy	Evidence packet, rule version, reviewer decision
Sepsis monitoring	Tracks vital signs, laboratory trends, medication history, admissions, and bedside data	Physician confirms escalation and treatment plan	Data feed failure, unstable recommendation, unexplained risk jump	Alert chain, data inputs, action timing
Oncology care monitoring	Detects toxicity signals from laboratory values, symptoms, wearable data, and adherence patterns	Oncology team confirms urgency and next step.	Missing laboratory trend, conflicting symptom report, unverified fever risk	Patient signal history, escalation note, clinician response
Clinical documentation	Drafts structured notes from encounters	Clinician edits and signs the note	Uncertain attribution, missing speaker identity, unsupported diagnosis	Draft version, edit history, signed note
Guideline retrieval	Retrieves protocol fragments and evidence summaries	The clinician decides whether the protocol fits the patient	Outdated guideline, missing patient-specific contraindication	Source version, retrieval path, final citation

The table supports risk-based autonomy. Administrative workflows tolerate broader drafting and retrieval functions, although denial, delay, or restriction of care still requires review. High-risk clinical

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

pathways require earlier interruption points and stronger exposure to evidence. Technical confidence should not set autonomy by itself. Clinical consequences should set the boundary.

The end-to-end walkthrough used a simulated adult medical-ward encounter. At 09:00, the patient had a temperature of 38.1 °C, heart rate of 104 beats per minute, blood pressure of 112/70 mmHg, respiratory rate of 20 breaths per minute, and no organ-support requirement. At 09:24, the verified monitor feed showed a heart rate of 122, blood pressure of 94/58, and respiratory rate of 26; a new laboratory result increased the configured deterioration score from 0.61 to 0.84.

The orchestrator first verified patient identity, timestamps, unit consistency, source availability, and executor permissions. It then called a risk-stratification executor and a local-pathway retrieval executor. Neither executor could submit an order, message the patient, or modify the record. The orchestrator assembled an evidence packet containing the input trend, missing-data status, risk value and change, tool outputs, pathway version, retrieved passages, uncertainty field, and proposed escalation. The clinician interface allowed acceptance, amendment, rejection, or deferral. Acceptance released a prepared order set for a second confirmation; amendment preserved both the agent draft and clinician revision; rejection required a reason code but did not stop monitoring. If the alert remained unacknowledged for 5 minutes, the orchestrator routed it to the backup clinician and recorded both routing events.

Six failures were injected. A vital-sign timestamp older than 15 minutes forced an insufficient-data state and suppressed the recommendation. Missingness above 20% triggered a data-quality notice. A unit mismatch prevented the affected laboratory value from entering the score. Retrieval from an unapproved source produced no clinical recommendation. An executor request to place an order was denied because the task card permitted drafting only. When the primary clinician did not acknowledge a high-risk alert, the backup route activated at 5 minutes. Under the conventional comparator, the same cases produced a score and alert, but the design did not expose whether the alert depended on stale data, which retrieval source had been used, whether an unauthorized action had been attempted, or why escalation had stopped.

The walkthrough verified four design properties. First, no tested pathway connected model output directly to clinical action. Second, data-quality and permission failures changed system state before the recommendation reached the clinician. Third, each human decision was recorded as a separate event. Fourth, the escalation chain remained reconstructable after non-acknowledgement. The exercise does not estimate sensitivity, specificity, mortality reduction, workload, or alert fatigue. It shows that safety controls can be tested with explicit pass conditions before prospective silent evaluation.

Table 2: End-to-end scenario comparison

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

Workflow stage	Conventional monolithic CDS	Proposed architecture	Observable output
Input validation	Score may be calculated from available fields without a distinct data-quality state	Orchestrator checks identity, timestamp, units, required-variable missingness, and source status before executor calls	Validated-input event or insufficient-data event
Risk calculation	Single model returns a score	Bounded executor returns score, version, input list, and uncertainty to the orchestrator	Model-call event with version and inputs
Evidence retrieval	Optional text explanation may be embedded in the alert	Approved retrieval executor returns local pathway version and cited passages	Retrieval source, version, passage identifiers
Recommendation assembly	Score and recommendation are produced by one component	Orchestrator combines validated inputs, tool outputs, evidence, and permission state	Evidence packet and proposed action
Tool permissions	Action capability may be coupled to the alerting component	Executors may retrieve, calculate, or draft only; order entry and messaging remain blocked before approval	Permission grant or denial event
Clinician response	Alert is acknowledged or dismissed	Clinician may accept, amend, reject, defer, or stop; original output and revision are preserved	Actor, timestamp, action, reason code, edited fields
Non-acknowledgement	May remain as an unopened alert	Backup clinician receives the high-risk alert after 5 minutes	Primary and backup routing events
Failure handling	Final alert may not reveal the failed component	Stale data, missingness, unit mismatch, unsupported source, and unauthorized calls produce distinct safe states	Failure code, suppressed action, recovery route

Table 3: Validation outcomes and operational definitions

Outcome	Operational definition	Scenario / pilot data required	Use
Inappropriate escalation	Escalations that do not satisfy the prespecified case definition divided by all escalations	Adjudicated scenario labels or clinician review	Safety and alert burden
Missed deterioration	Qualifying deterioration episodes not escalated within the allowed interval divided by all qualifying episodes	Time-stamped ground truth and alert log	Sensitivity of the workflow, not only the model
Unsupported recommendation rate	Recommendations lacking a traceable patient-data basis, approved source, or tool output divided by all recommendations	Evidence packets and audit log	Grounding and auditability
Clinician override rate	Rejected or substantively amended recommendations divided by all reviewed recommendations	Clinician action and edit history	Fit with workflow; not an isolated accuracy measure
Time to clinical intervention	Time from qualifying signal to the first documented clinician action	Signal timestamp, alert delivery, acknowledgement, action timestamp	Latency and workflow impact
Executed permission violations	Unauthorized tool actions completed despite policy	Permission-denial and action logs	Must equal zero

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

Outcome	Operational definition	Scenario / pilot data required	Use
Audit completeness	Required event fields present divided by all required event fields	Audit schema validation	Target 100% for scenario verification

Table 4 gives an implementation sequence for clinical organizations. It compares design choices across deployment stages and identifies the artifact that governs each stage.

Table 4: Implementation sequence for controlled clinical AI agents

Stage	Design choice	Governance question	Required artifact
1. Pathway selection	Select a bounded workflow with clear clinical ownership	Who owns escalation rules and pathway changes	Use-case charter
2. Risk classification	Grade harm potential and urgency	Which actions require approval before execution	Risk-control matrix
3. Data mapping	Define permitted sources and quality checks	Which data streams support agent reasoning	Data lineage map
4. Orchestrator design	Set tool permissions and routing rules	Which tools the agent may call and under which conditions	Orchestration specification
5. Executor configuration	Limit each executor to one bounded task	Which output formmay each executory produce	Executor task card
6. Human-in-the-loop design	Define review, amendment, and stop points	Where clinician judgment enters the workflow	Human authority map
7. Pilot deployment	Run the agent in a limited clinical setting	How the agent behaves during real clinical work	Pilot protocol
8. Production monitoring	Track overrides, incidents, missing data events, and drift	When the organization adjusts or withdraws the agent	Monitoring dashboard and incident log

The minimum audit schema contains encounter and patient identifiers, event timestamp, actor, workflow state, input-source identifiers and freshness, model and tool versions, prompt or task-card version where

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

applicable, permission decision, retrieved source and version, evidence passages, output and uncertainty, proposed action, clinician disposition, edited fields, reason code, escalation recipient, acknowledgement time, action time, and incident flag. Protected clinical content may be stored by reference when direct duplication would exceed the organization's minimum-necessary policy, but the log must still permit reconstruction of which information was available at each decision point.

The sequence places governance before scaling. Agentic systems expand quickly once teams connect them to records, messaging systems, and clinical tools. Control becomes harder after broad deployment. A vendor-specific build may integrate tightly with one EHR, cloud environment, and model provider. A vendor-agnostic build may route tasks across several tools. Both options support responsible use when the organization fixes the governance layer: data lineage, tool permissions, human authority, audit logs, and monitoring thresholds.

Vendor-agnostic architecture offers practical advantages in clinical settings. It prevents the organization from tying safety logic to one model vendor. The orchestrator can send retrieval, drafting, and classification tasks to different tools while preserving the same review rule. If the team changes a language model, the agent still has to expose evidence, identify uncertainty, and route high-risk outputs to human review. Vendor flexibility becomes a control mechanism when interchangeable execution tools remain separate from clinical governance.

Post-deployment monitoring deserves the same design attention as initial validation. Clinical workflows shift, guidelines change, data feeds fail, and clinicians develop override habits. A responsible agent system monitors override rates, unsupported recommendations, silent failures, missing data events, user edits, alert fatigue, and subgroup drift. These signals show whether the agent still fits its clinical environment. Repeated unsafe behavior should trigger restriction, redesign, or deactivation.

Human-in-the-loop control has practical value only when teams can observe and enforce it. Clinicians need access to the evidence behind recommendations, a clear intervention point before clinical action, a record of acceptance or rejection, and a stop mechanism for unsafe behavior. Without those operating features, human oversight becomes a policy phrase rather than a clinical safeguard.

The validation scenario-based demonstrates that the proposed controls create observable and enforceable state transitions under prespecified failures, but it does not show that the architecture improves diagnosis, treatment selection, mortality, clinician workload, or alert fatigue. The next evidentiary stage is a prospective silent deployment in the intended clinical environment, followed by early live evaluation under DECIDE-AI. Silent-trial literature shows that technical performance is frequently reported more completely than in situ ground truth, human-computer interaction, and sociotechnical outcomes [13]. The pilot should therefore retain the process outcomes defined in Table

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

4 and add subgroup analysis, usability assessment, and incident review before any expansion of agent permissions.

5. Conclusion

The structured synthesis produced three requirements that remained stable across regulatory, consensus, empirical, and review evidence: orchestration must be separated from bounded execution; clinical action must remain behind risk-defined human approval and override; and audit and monitoring must continue after deployment. Data lineage, evidence exposure, uncertainty display, and stable governance rules connect these requirements across different vendors and clinical pathways.

The sepsis walkthrough converted the architecture from a diagram into a testable workflow. Stale or incomplete inputs produced an insufficient-data state, unsupported retrieval did not reach the clinician as a recommendation, an unauthorized order attempt was blocked, non-acknowledgement activated a backup route, and the log preserved the evidence and human decisions needed to reconstruct the event. These results support design-level claims about controllability and auditability only. They do not demonstrate superior clinical performance relative to conventional decision support.

Clinical adoption requires a staged evaluation sequence: scenario and fault-injection testing, prospective silent deployment, early live clinical evaluation, and post-deployment monitoring with predefined restriction or withdrawal rules. Comparative studies should measure missed deterioration, inappropriate escalation, unsupported recommendations, override behavior, intervention latency, workload, subgroup performance, and adverse incidents before broader autonomy is considered.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Funding: The authors received no external funding for this work.

Competing interests: The authors declare that they have no competing interests.

Author contributions: All authors contributed to this research and approved the final manuscript.

References

[1] M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, et al., “PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews,” Systematic Reviews, vol. 10, Art. no. 39, 2021, doi: 10.1186/s13643-020-01542-z.

[2] C. Baethge, S. Goldbeck-Wood, and S. Mertens, “SANRA—A scale for the quality assessment of narrative review articles,” Research Integrity and Peer Review, vol. 4, Art. no. 5, 2019, doi: 10.1186/s41073-019-0064-8.

[3] B. G. Collaco, S. A. Haider, S. Prabha, et al., “The role of agentic artificial intelligence in healthcare: A scoping review,” npj Digital Medicine, vol. 9, Art. no. 345, 2026, doi: 10.1038/s41746-026-02517-5.

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

[4] N. Karunanayake, "Next-generation agentic AI for transforming healthcare," Informatics and Health, vol. 2, no. 2, pp. 73-83, 2025, doi: 10.1016/j.infoh.2025.03.001.

[5] J. E. Alderman, J. Palmer, E. Laws, et al., "Tackling algorithmic bias and promoting transparency in health datasets: The STANDING Together consensus recommendations," The Lancet Digital Health, vol. 7, no. 1, pp. e64-e88, 2025, doi: 10.1016/S2589-7500(24)00224-3.

[6] G. S. Collins, P. Dhiman, C. L. A. Navarro, et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.

[7] B. Vasey, M. Nagendran, B. Campbell, et al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," Nature Medicine, vol. 28, pp. 924-933, 2022, doi: 10.1038/s41591-022-01772-9.

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

[9] D. O'Reilly, J. McGrath, and I. Martin-Loeches, "Optimizing artificial intelligence in sepsis management: Opportunities in the present and looking closely to the future," Journal of Intensive Medicine, vol. 4, no. 1, pp. 34-45, 2024, doi: 10.1016/j.jointm.2023.10.001.

[10] K. Lekadir, A. Feragen, A. J. Fofanah, et al., "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," BMJ, vol. 388, Art. no. e081554, 2025, doi: 10.1136/bmj-2024-081554.

[11] U.S. Food and Drug Administration, Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. Silver Spring, MD, USA: FDA, 2026.

[12] World Health Organization, Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. Geneva, Switzerland: WHO, 2025.

[13] L. Tikhomirov, C. Semmler, N. Prizant, et al., "A scoping review of silent trials for medical artificial intelligence," Nature Health, vol. 1, pp. 532-554, 2026, doi: 10.1038/s44360-025-00048-z.

[14] R. Adams, K. E. Henry, A. Sridharan, et al., "Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis," Nature Medicine, vol. 28, pp. 1455-1460, 2022, doi: 10.1038/s41591-022-01894-0.

[15] D. Ferber, L. Hilgers, C. Höper, et al., "Towards autonomous medical artificial intelligence agents," Nature, vol. 655, pp. 1282-1291, 2026, doi: 10.1038/s41586-026-10675-5.

Supplementary Table S1

Source	Type / evidentiary weight	R1: orchestration vs bounded execution	R2: human authority / override	R3: audit and monitoring	Additional control supported
[5] Alderman et al., 2025	Consensus recommendations / high	Indirect	—	Direct	Dataset transparency, representativeness, lineage
[3] Collaco et al., 2026	Scoping review / moderate	Direct	Indirect	Indirect	Agent autonomy, tool use, multi-agent coordination
[6] Collins et al., 2024	Reporting guideline / high	Indirect	Indirect	Direct	Component-level reporting for embedded prediction models
[8] Ferber et al., 2025	Case-based empirical evaluation / moderate	Direct	Indirect	Direct	Tool traceability and multimodal evidence retrieval

ISSN: 2307-4523
Vol. 57, No. 1, 2026 | pp. 585-599

Source	Type / evidentiary weight	R1: orchestration vs bounded execution	R2: human authority / override	R3: audit and monitoring	Additional control supported
[4] Karunanayake, 2025	Conceptual review / contextual	Illustrative only	Illustrative only	Illustrative only	Terminology and deployment examples
[10] Lekadir et al., 2025	International consensus guideline / high	Indirect	Direct	Direct	Fairness, robustness, traceability, usability, explainability
[9] O'Reilly et al., 2024	Clinical review / moderate	Indirect	Direct	Direct	Sepsis data diversity, trust, interpretability, readiness
[11] FDA, 2026	Regulatory guidance / high	Indirect	Direct	Direct	Independent review of recommendation basis and risk-based regulation
[7] Vasey et al., 2022	Reporting guideline / high	Indirect	Direct	Direct	Early live evaluation, human factors, workflow interaction
[12] WHO, 2025	International governance guidance / high	Indirect	Direct	Direct	Multimodal risk, accountability, data and model governance
[1] Rethlefsen et al., 2021	Search-reporting guideline / methodological	—	—	—	Reproducible database and search reporting
[2] Baethge et al., 2019	Narrative-review appraisal / methodological	—	—	—	Review quality and scientific reasoning
[14] Adams et al., 2022	Prospective multi-site cohort / high	Indirect	Direct	Direct	Provider acknowledgement, time to action, clinical workflow outcome
[13] Tikhomirov et al., 2026	Scoping review of silent trials / moderate-high	Indirect	Indirect	Direct	Prospective noninterventional validation and sociotechnical outcomes
[15] Ferber et al., 2026	Large sandboxed EHR evaluation / high for simulation	Direct	Indirect	Direct	End-to-end tool use, full-workflow evaluation, structured auditability