



Italian Journal of Medicine

<https://www.italjmed.org/ijm>

eISSN 1877-9352

Publisher's Disclaimer. E-publishing ahead of print is increasingly important for the rapid dissemination of science. The Early Access service lets users access peer-reviewed articles well before print/regular issue publication, significantly reducing the time it takes for critical findings to reach the research community.

These articles are searchable and citable by their DOI (Digital Object Identifier).

The **Italian Journal of Medicine** is, therefore, E-publishing PDF files of an early version of manuscripts that have undergone a regular peer review and have been accepted for publication, but have not been through the copyediting, typesetting, pagination, and proofreading processes, which may lead to differences between this version and the final one.

The final version of the manuscript will then appear in a regular issue of the journal.

The E-publishing of this PDF file has been approved by the authors.

Ital J Med 2026 [Online ahead of print]

Please cite this article as:

Angioni D, Tirotta D, Di Bello F, Montagnani A. **Preserving clinical method in artificial intelligence-assisted diagnosis: a cognitive-safety framework against premature diagnostic closure.** *Ital J Med* doi: 10.4081/itjm.2026.2648

Submitted: 29-06-2026

Accepted: 02-10-2026

© the Author(s), 2026
Licensee [PAGEPress](#), Italy

Note: The publisher is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries should be directed to the corresponding author for the article.

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Preserving clinical method in artificial intelligence-assisted diagnosis: a cognitive-safety framework against premature diagnostic closure

Daniele Angioni,¹⁻⁴ Daniela Tirotta,⁵ Fabio Di Bello,⁶⁻⁸ Andrea Montagnani⁹⁻¹¹

¹ASL Novara, Italy; ²National Commission for Digital Innovation and Technology of AME, with delegation for artificial intelligence, Udine, Italy; ³Permanent Committee on Artificial Intelligence, Federation of the Orders of Physicians and Dentists of Piedmont and Aosta Valley, Novara, Italy; ⁴Italian Society of Ultrasound in General Practice and Clinical Ultrasound, Parma, Italy; ⁵Internal Medicine Unit, Morgagni-Pierantoni Hospital, AUSL Romagna, Forlì, Italy; ⁶Wiley, Hoboken, NJ, USA; ⁷Master in Humanization of Care, University of Udine, Italy; ⁸University Bicocca Academy of Milan, Italy; ⁹Internal Medicine Unit, Misericordia Hospital, Grosseto, Italy; ¹⁰Department of Internal Medicine and Medical Specialties, AUSL Toscana Sud Est, Italy; ¹¹Federation of Associations of Hospital Doctors on Internal Medicine (FADOI), Rome, Italy

Correspondence: Daniele Angioni, ASL Novara, Novara, Italy.
E-mail: studiomedicoangioni@gmail.com

Key words: generative artificial intelligence, large language models, differential diagnosis, diagnostic error, cognitive safety.

Contributions: Daniele Angioni: study concept and draft of the manuscript. Fabio Di Bello: conceptual development and critical revision. Daniela Tirotta: critical revision of the clinical content. Andrea Montagnani: supervision of the internal medicine perspective and critical revision of the manuscript. All authors read and approved the final manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare that they have no competing interests.

Ethics approval and consent to participate: not applicable.

Informed consent: not applicable. This narrative review and conceptual framework did not involve human participants or individual patient data; the clinical example is hypothetical.

Availability of data and materials: not applicable.

Declaration of generative artificial intelligence and artificial intelligence-assisted technologies in the writing process: artificial intelligence-assisted tools were used for language editing and support with manuscript shortening. All scientific content, source selection, interpretation, and final wording were reviewed and approved by the authors, who retain full responsibility for accuracy, integrity, and originality.

Abstract

Generative artificial intelligence (AI) can support differential diagnosis, but final-answer accuracy does not by itself establish diagnostic safety. This narrative review examines how AI exposure may alter hypothesis generation and proposes AI-associated differential diagnosis narrowing as a testable process-level construct of the clinician-AI system. The construct describes reduced differential breadth, flexibility, or counterfactual exploration after AI exposure; it is not a validated syndrome and does not imply causation. Anchoring, automation bias, fluent language, and interaction order are possible mechanisms, while premature closure and diagnostic error are potential downstream consequences. We organize candidate safeguards through the GENERATE-Dx mnemonic, which preserves independent hypothesis generation, separates likelihood from danger, makes missing and discordant data explicit, reopens alternatives, and defines verification and reassessment. Candidate endpoints include pre/post differential breadth, retention of expert-adjudicated rare-but-dangerous diagnoses, counterfactual exploration, confidence calibration, management decisions, and reassessment. The framework, metrics, and figures are heuristic proposals. Prospective studies should compare interaction order and interface design and evaluate whether AI supports accurate decisions without weakening the clinical method that makes diagnosis safe.

Introduction

Generative artificial intelligence (AI) is now used in clinical work not simply as a search tool, but as a system that writes diagnostic hypotheses, explanations, plans, and summaries. Traditional decision-support tools usually return probabilities, alerts, or classifications; large language models (LLMs) produce narratives. Those narratives may help organize complexity, but they may also give uncertainty a premature sense of resolution.

Differential diagnosis is the clinical structure that protects patients from closure error. Diagnosis requires progressive narrowing; the danger begins when narrowing takes place before serious alternatives have been sought, discordant data have been weighed, and residual uncertainty has been made explicit.

Differential diagnosis functions as a clinical safety tool. It keeps alternative hypotheses active, brings contradictory data into view, preserves attention to time-dependent disease, and protects the patient from premature diagnostic closure. Clinical method precedes AI and should remain the standard by which AI is judged. Diagnosis is not the act of naming a disease; it is the disciplined management of uncertainty.

Diagnostic-error literature has long emphasized faulty synthesis, anchoring, and premature closure.^{1,2} The National Academies described diagnosis as a collaborative and vulnerable process across care settings.³ AI adds automation bias, reduced vigilance, excessive trust, and verification difficulties.⁴⁻⁸ Recent LLM studies remain heterogeneous: GPT-4 access did not significantly improve physicians' diagnostic reasoning in one randomized trial,⁹ whereas the Articulate Medical Intelligence Explorer performed strongly on complex differential-diagnosis tasks in an experimental setting.¹⁰

These findings point to the question addressed in this review: not simply whether an AI system can produce correct diagnoses, but whether its use preserves the quality of the differential process. A model may improve the final list in one context and increase anchoring in another, depending on timing, interface, wording, and workflow. Safety is not a property of the model alone; it belongs to the clinician-AI system.

This issue is especially important in internal medicine and community care, where multimorbidity, frailty, evolving presentations, low prevalence, and longitudinal decisions make premature simplification hazardous.

This article is a targeted narrative review and conceptual framework. Its primary objective is to synthesize evidence from diagnostic cognition and clinician-AI interaction and define a testable process-level risk construct. Secondary objectives are to organize candidate safeguards, measurable endpoints, and an empirical research agenda. The examples illustrate, but do not validate, the framework; the article is neither a systematic review, regulatory standard, nor validated measurement instrument.

Key points

- The diagnostic safety of AI does not coincide with model accuracy: a system may suggest a plausible diagnosis while impoverishing the clinician's differential process.
- Differential narrowing is appropriate when it follows adequate verification; it becomes potentially unsafe when it is premature, insufficiently verified, lacks counterfactual testing, and is associated with AI exposure.
- The construct is proposed as a falsifiable cognitive-safety hypothesis, not as a validated nosographic entity.
- Protective generative techniques should be evaluated for differential breadth, retention of rare-but-dangerous diagnoses, contradiction checking, and confidence calibration.
- Protecting diagnostic method is crucial in internal medicine, primary care, and general practice, where nonspecific symptoms may represent the onset of serious time-dependent conditions.

Literature search strategy

A targeted, non-systematic narrative search was completed on 31 May 2026 using PubMed (including MEDLINE-indexed records) and Scopus as core databases. Because the review combines current clinical AI with older cognitive and human-factors research, two evidence streams were used. The contemporary clinical-AI stream covered 1 January 2015 through 31 May 2026. This pragmatic cutoff captures contemporary clinical AI, including work predating generative systems, and coincides with the 2015 National Academies report that framed diagnostic error as a patient-safety priority.³ It is a scope anchor, not a claim that AI or diagnostic-safety scholarship began in 2015. The diagnostic-cognition, human-factors, and automation stream had no lower date limit. Backward and forward citation searching identified further landmark sources. Thus, eligible pre-2015 work on anchoring, automation bias, premature closure, diagnostic momentum, trust in automation, and reflective reasoning entered through a prespecified unlimited stream, not as exceptions.

Database-specific search strings

PubMed contemporary-AI queries were run separately. Query A: (("artificial intelligence"[Title/Abstract] OR "generative artificial intelligence"[Title/Abstract] OR "large language model"[Title/Abstract] OR "large language models"[Title/Abstract]) AND ("diagnostic reasoning"[Title/Abstract] OR "differential diagnosis"[Title/Abstract] OR "diagnostic error"[Title/Abstract] OR "diagnostic errors"[Title/Abstract] OR "clinical decision support"[Title/Abstract])). Query B: (("artificial intelligence"[Title/Abstract] OR "machine learning"[Title/Abstract] OR automation[Title/Abstract]) AND ("automation bias"[Title/Abstract] OR anchoring[Title/Abstract] OR "premature closure"[Title/Abstract] OR "cognitive bias"[Title/Abstract] OR "human-AI collaboration"[Title/Abstract] OR "human machine collaboration"[Title/Abstract]) AND (diagnos*[Title/Abstract] OR clinical[Title/Abstract])). Query C: (("large language model"[Title/Abstract] OR "large language models"[Title/Abstract] OR "generative artificial intelligence"[Title/Abstract]) AND (verification[Title/Abstract] OR counterfactual*[Title/Abstract] OR "self-consistency"[Title/Abstract] OR prompt*[Title/Abstract] OR "multi-agent"[Title/Abstract]) AND (medical[Title/Abstract] OR clinical[Title/Abstract] OR diagnos*[Title/Abstract])). Each was combined with ("2015/01/01"[Date - Publication] : "2026/05/31"[Date - Publication]) AND English[Language]. The no-lower-date PubMed query was: (("diagnostic error"[Title/Abstract] OR "diagnostic errors"[Title/Abstract] OR "premature closure"[Title/Abstract] OR "premature diagnostic closure"[Title/Abstract] OR anchoring[Title/Abstract] OR "automation bias"[Title/Abstract] OR "diagnostic momentum"[Title/Abstract] OR "reflective reasoning"[Title/Abstract]) AND (diagnos*[Title/Abstract] OR clinical[Title/Abstract] OR clinician*[Title/Abstract] OR physician*[Title/Abstract])) AND English[Language].

Scopus contemporary-AI adaptations were: A, TITLE-ABS-KEY(("artificial intelligence" OR "generative artificial intelligence" OR "large language model*") AND ("diagnostic reasoning" OR "differential diagnosis" OR "diagnostic error*" OR "clinical decision support")); B, TITLE-ABS-KEY(("artificial intelligence" OR "machine learning" OR automation) AND ("automation bias" OR anchoring OR "premature closure" OR "cognitive bias" OR "human-AI collaboration" OR "human machine collaboration") AND (diagnos* OR clinical)); and C, TITLE-ABS-KEY(("large language model*" OR "generative artificial intelligence") AND (verification OR counterfactual* OR "self-consistency" OR prompt* OR "multi-agent") AND (medical OR clinical OR diagnos*)). Each was combined with PUBYEAR AFT 2014 AND PUBYEAR BEF 2027. The no-lower-date Scopus query was TITLE-ABS-KEY(("diagnostic error*" OR "premature closure" OR "premature diagnostic closure" OR anchoring OR "automation bias" OR "diagnostic momentum" OR "reflective reasoning") AND (diagnos* OR clinical OR clinician* OR physician*)). The Scopus interface filter Language = English was applied to all four searches. Because Scopus uses year-level

rather than day-level limits in this syntax, the upper boundary was enforced by closing the search on 31 May 2026; records identified after that date were not included.

Google Scholar was used only as a supplementary source and for forward citation searching, without field restrictions, using the exact combinations "large language model" AND "differential diagnosis" AND clinician; "artificial intelligence" AND "automation bias" AND diagnosis; "human-AI collaboration" AND diagnosis AND order; and "large language model" AND diagnostic reasoning AND (verification OR counterfactual). Eligible material comprised peer-reviewed clinical studies, reviews, major reports, guidelines, and landmark conceptual papers relevant to diagnostic reasoning or clinician-AI interaction. Duplicates, conference abstracts, non-clinical applications, and technical model-development papers without implications for clinical reasoning were excluded. Selection was conceptual rather than exhaustive; no quantitative synthesis or formal risk-of-bias assessment was performed. One author paper accepted on 21 May 2026 and published online on 23 June 2026 was added during revision as a targeted post-search conceptual citation; no other records published after the closing date were included.

From diagnostic error to cognitive safety

Diagnostic error is usually recognized downstream, as a missed, wrong, or delayed diagnosis. Cognitive safety looks earlier in the process: which hypotheses were generated, which alternatives disappeared, which discordant elements were ignored, which missing data were not requested, and which warning signs were normalized by a coherent but incomplete explanation.

In this review, cognitive safety denotes the capacity of the clinician-AI system to preserve sound reasoning conditions: adequate differential breadth, hypothesis verification, counterfactual exploration, awareness of missing information, confidence calibration, and attention to rare-but-serious diagnoses. A model may perform well on benchmarks and still be unsafe if it weakens this process in real workflows.

Proposed construct and conceptual boundaries

We newly propose AI-associated differential diagnosis narrowing as a process-level cognitive-safety phenomenon in which exposure to AI-generated diagnostic suggestions is followed by reduced breadth, flexibility, tolerance of residual uncertainty, or counterfactual exploration in the clinician's differential diagnosis. The construct concerns the clinician-AI interaction rather than an output error in isolation.

The term does not denote a disease, syndrome, or established cognitive bias. It is best treated as an intermediate cognitive-process configuration and candidate outcome of AI exposure. Anchoring, automation bias, linguistic authority, and interaction order are possible antecedents or mechanisms. Loss of alternatives may form an intermediate pathway toward premature closure; diagnostic momentum may propagate the narrowed frame across later encounters. Hallucination is different: it is an output property and can be absent even when narrowing occurs.

Causal language must match study design. AI-induced narrowing should be reserved for controlled designs that support causal attribution. In observational or descriptive work, AI-associated differential diagnosis narrowing is the preferred term. This terminology is used consistently throughout the manuscript and figures.

A clear distinction is necessary between appropriate and potentially unsafe narrowing. Competent clinicians necessarily narrow the differential; this is the purpose of diagnostic reasoning. Narrowing becomes potentially unsafe when it occurs early, is insufficiently verified, lacks counterfactual testing, or removes relevant alternatives before rare, serious, treatable, or time-dependent diagnoses have been considered.

The construct is therefore neither an antecedent bias nor the final clinical outcome. It describes an intermediate process configuration at the level of the clinician-AI system. It would gain support if controlled exposure to specified AI outputs produced reproducible pre/post reductions in clinically relevant hypotheses, contradiction checking, or counterfactual exploration compared with suitable

controls. It would be weakened if those changes were absent, explained by baseline case difficulty, or offset by improved retention of expert-adjudicated alternatives.

Potential mechanisms and interaction order

AI-associated diagnostic narrowing may follow established mechanisms such as anchoring and automation bias, amplified by fluent, immediate, and apparently reasoned output. The first AI response can frame subsequent thinking; coherent prose can be mistaken for evidential strength; and false reassurance can weaken searches for red flags or contradictory findings.

Other plausible processes include greater availability of diagnoses named by the model, reduced salience of rare alternatives, loss of temporal or atypical context, less counterfactual testing, and overconfidence. These are descriptive pathways, not additional proposed constructs. Clinically, probability ranking must remain distinct from priority: a low-probability diagnosis may still demand attention when dangerous, treatable, reversible, or time-dependent.

In community-care contexts, these mechanisms may be amplified by low prevalence, nonspecific initial symptoms, workload pressure, delayed access to investigations, and the need to decide whether to observe, treat, reassess, or refer. An apparently ordered AI output may therefore modify not only the final diagnosis, but also the physician's threshold for escalation.

Interaction order is a central experimental variable, but existing evidence does not support a universal sequence. Cabitza and colleagues compared human-first and AI-first collaboration protocols in knee magnetic resonance imaging and electrocardiogram interpretation. Order affected performance, and in those specific tasks, AI-first protocols were associated with higher accuracy.¹¹ The study shows that timing and information configuration matter, while cautioning against presenting independent human formulation before AI exposure as an evidence-established rule for every task.

For generative differential diagnosis, recording a provisional problem representation and differential before viewing the model output remains a plausible safeguard against anchoring and a useful way to measure pre/post change. It should be compared directly with AI-first and other collaboration protocols rather than prescribed as universally superior. Interface design, user expertise, case type, AI accuracy, and explanation format may all modify the effect.

The operational risk mechanisms considered in the framework are summarized in Table 1.

Generative artificial intelligence as risk and remedy

Several techniques can be reframed as cognitive safeguards. Chain-of-verification reduces hallucinations in non-clinical natural language processing tasks,¹² self-consistency improves general reasoning benchmarks,¹³ and self-refinement, tree-of-thoughts, and multi-agent debate have been proposed to improve reasoning or factuality.¹⁴⁻¹⁶ Clinical prompting studies are promising but insufficient to prove reduction of diagnostic error or premature closure in real-world care.¹⁷⁻¹⁹

These techniques should therefore not be presented as validated clinical interventions. Their current value is architectural and testable: they may help shift AI from a producer of answers to a source of useful cognitive friction. Safety does not require more persuasive AI. It requires systems that make premature closure harder.

The same concern applies to retrieval-augmented and clinically oriented AI. Curated or user-provided sources do not guarantee clinical truth. Errors may still arise from retrieval of irrelevant or incorrect material, incomplete chunking, faulty synthesis, over-inference, conflicts among sources, or override by parametric knowledge. Knowledge-base integration may reduce some risks, but it does not by itself prevent evidence-like error or premature diagnostic closure.²⁰

Candidate generative safeguards

One candidate safeguard is to request multiple independent and diagnostically diverse lists. The output should separate the most likely diagnoses, dangerous diagnoses not to miss, and diagnoses

that explain atypical elements. This structure prevents a single ranking from becoming the whole diagnostic space.

Chain-of-verification requires each proposed diagnosis to be checked against supportive data, contrary data, missing data, alternatives, and warning signs. In medicine, verification should be designed to falsify rather than merely confirm. The point is not to rationalize the first answer, but to test whether it survives contradiction. Studies in internal medicine residents suggest that reflective reasoning can improve diagnostic accuracy in complex cases and mitigate availability bias; its efficacy in AI-assisted diagnosis remains untested.^{21,22}

Recursive reasoning asks the model to produce an initial differential, critique it for anchoring and premature closure, reopen excluded hypotheses, simulate a dissenting reviewer, and integrate the revised analysis. Its value lies not in anthropomorphizing the model but in imposing a de-anchoring workflow.

Counterfactual prompting starts from the assumption that the leading diagnosis is false and asks which alternatives, underweighted findings, or additional data would overturn it. Adversarial prompting introduces a skeptical reviewer. Self-consistency and repeated independent generations compare multiple paths; minority outputs should not be dismissed when they contain rare-but-dangerous diagnoses.

These prompts are not autonomous clinical protocols. They are examples of how cognitive safeguards may be translated into structured instructions for a generative system, with the aim of preserving differential breadth, attention to contrary evidence, residual uncertainty, and retention of do-not-miss diagnoses.

GENERATE-Dx: a proposed mnemonic

We newly use GENERATE-Dx as a mnemonic for eight observable steps: Generate an independent provisional differential; Explicitly separate likelihood, danger, urgency, and actionability; Note missing, discordant, and time-dependent data; Examine supporting evidence and contradictions; Reopen alternatives; Audit anchoring, automation bias, and overconfidence; Test do-not-miss diagnoses and escalation thresholds; and Establish reassessment, safety-netting, and residual uncertainty. Dx denotes diagnostic application and is not a ninth component.

GENERATE-Dx is a heuristic design framework, not a validated score or clinical protocol. Its components reorganize established principles of diagnostic reasoning and candidate generative techniques for evaluation in AI-assisted workflows. Table 2 maps each step to its conceptual derivation and principal target. Any implementation should be tested for diagnostic benefit, cognitive load, time cost, and unintended over-investigation.

Heuristic representation and candidate measurement

Figure 1 uses author-generated schematic bars to illustrate possible changes in cognitive salience across four discrete, unordered hypothesis clusters before and after AI exposure. No patient or clinician data were used. Bar heights and differences are arbitrary and do not estimate probabilities, distributions, effect sizes, or distances between diagnostic categories.

We also newly propose a Tail-Preservation Index (TPI) for each clinician-case pair. Let T be the expert-adjudicated set of non-dominant, clinically high-relevance diagnoses present in the pre-AI differential. TPI is the weighted proportion of T retained after AI exposure (retained weight divided by total weight). It ranges from 0 to 1 and is undefined when T is empty. Here, tail denotes non-dominant alternatives, not an assumed distributional shape.

We propose equal weights for primary analyses. Prespecified sensitivity analyses may weight diagnoses by severity, time sensitivity, treatability, and case relevance. A blinded expert panel should define the case-specific high-relevance set, reconcile diagnostic synonyms, and adjudicate retention. TPI measures preservation of pre-existing non-dominant hypotheses; new do-not-miss diagnoses added after AI exposure should be reported separately.

Other candidate endpoints include differential breadth and semantic diversity, shift toward the first AI recommendation, counterfactual exploration, confidence calibration, verification completeness, premature closure, referral appropriateness, reassessment quality, and cognitive load. These remain proposals and require reliability testing, construct validation, and association with clinical outcomes. Candidate metrics are summarized in Table 3.

Empirical research agenda

Empirical verification is difficult because model version, parameters, interface, context, and prompt formulation change outputs.^{18,23,24} Early studies should use standardized vignettes and predefined or frozen AI outputs to isolate response-format effects, then compare standard answers with structured answers that separate probability, danger, missing data, contradictions, red flags, and verification requests.

The purpose should not be to prove that one model is consistently superior to another. The more useful question is whether specific cognitive design choices reduce anchoring, premature closure, and suppression of rare-but-dangerous diagnoses. This approach fits the dynamic nature of generative AI and the need to evaluate the clinician-AI system rather than the isolated model.

A feasible first study could randomize clinicians to clinician-alone, AI-first, human-first, and safeguarded-interface conditions using standardized cases and frozen outputs. Final between-arm endpoints would include expert-adjudicated breadth, do-not-miss coverage, verification, confidence, management, and reassessment. Pre/post change and TPI would be analyzed only in arms recording an independent pre-AI differential; clinician-alone participants would complete matched repeated assessments without AI. Figure 2 summarizes the proposed pathway and explicitly presents its arrows as hypotheses for testing rather than established causal relations.

Clinical implications across care settings

The framework concerns the whole diagnostic pathway, from first contact in the community to emergency assessment and internal medicine. General practice is especially relevant because low prevalence, nonspecific symptoms, and longitudinal decisions make risk thresholds central.

In primary care, the physician does not decide only which diagnosis is most likely, but also when to reassure, when to monitor, when to request investigations, when to reassess, and when to refer. A safe AI system should therefore support the longitudinal management of uncertainty, not replace it with an apparently definitive answer.

Internal medicine faces multimorbid complexity: overlapping signals, discordant tests, drug effects, frailty, chronic organ dysfunction, and coexisting diseases. AI is useful when it helps organize these data; it is risky when it turns a multi-causal picture into an explanation that is too clean.

Emergency medicine adds time pressure and incomplete information. A system that privileges the common pattern may delay time-dependent diagnoses; one that separates probability from danger may be protective. Across settings, retained hypotheses, missing data, referral reasons, safety-netting, and reassessment criteria should be documented.

Clinical example

Consider a hypothetical older multimorbid patient with persistent low-grade fever, fatigue, weight loss, normocytic anemia, elevated inflammatory markers, high ferritin, increased lactate dehydrogenase, and a focal liver abnormality detected by ultrasound. A convergent AI prompt may steer the work-up toward occult bacterial infection, liver abscess, or metastatic solid tumor. These hypotheses become unsafe if they dominate before discordant elements are tested.

A cognitive-safety response should test each dominant hypothesis against contradictory data: liver abscess against cultures, leukocytosis, imaging morphology, antibiotic response, and need for drainage; metastatic disease against imaging pattern, oncological work-up, and tissue diagnosis. At the same time, lymphoma, hemophagocytic lymphohistiocytosis/macrophage activation syndrome,

systemic autoimmune disease, infective endocarditis, tuberculosis, drug-related fever, and hematologic infiltrative disease should remain visible until excluded with proportionate evidence. In this scenario, generating multiple independent lists separates common, dangerous, and atypical explanations; chain-of-verification identifies supportive data, contrary data, missing tests, and red flags; recursive reasoning critiques the first explanation and reopens downgraded alternatives. The clinical target is not diagnostic proliferation, but safer narrowing: defining missing data, time-dependent diagnoses, referral thresholds, need for biopsy or specialist input, and reassessment criteria.

Training, governance and post-deployment monitoring

Introducing generative AI into clinical practice requires training beyond tool use. Clinicians need to recognize how AI can modify their own reasoning: anchoring, false reassurance, overconfidence, loss of alternatives, and implicit delegation of verification. Training should integrate AI literacy with diagnostic reasoning, uncertainty management, and cognitive safety.

This is consistent with a human-in-the-loop model understood as an operational standard, not as a slogan. The physician must remain responsible for clinical method, verification, final decision-making, and reassessment. Human supervision should be effective, documented, and capable of challenging or overriding algorithmic output when it is not coherent with the clinical picture.²⁵

Regulation should incorporate cognitive endpoints. Accuracy, sensitivity, and specificity are necessary but insufficient. Validation and surveillance should assess how AI changes clinician behavior: differential diagnosis, do-not-miss retention, contrary-data search, confidence calibration, referral appropriateness, and reassessment quality.

Post-deployment monitoring should examine not only incorrect outputs but also recurrent changes in reasoning after AI use. Local audits can compare pre-AI and post-AI differentials, retention of red flags, contradiction checking, confidence, referral decisions, and reassessment plans. These are governance proposals; thresholds for regulatory use require empirical validation and professional consensus.

Limitations

This is a targeted narrative review rather than a systematic review. The search is reported transparently but was not exhaustive, and no formal risk-of-bias assessment or quantitative synthesis was performed. The proposed construct, mnemonic, metric, and pathway may therefore reflect author selection and interpretation and require independent replication.

The main limitation is that the construct is proposed, not validated. Its frequency, determinants, specialty-specific vulnerability, and clinical consequences remain unknown. The protective value of generative safeguards is also unproven in real-world diagnostic pathways. Recent literature on clinician-LLM collaboration remains heterogeneous, with results depending heavily on task, setting, interface, and experimental design.^{9,23,26}

A second limitation concerns causality. Demonstrating that AI induces narrowing requires controlled pre/post measures, randomization, order control, and interface analysis. A third limitation is the opposite risk: maintaining an overly broad differential can increase cognitive load, costs, inappropriate investigations, and patient anxiety. Safety requires appropriate breadth, not maximal breadth. Metrics must distinguish useful diagnostic richness from noise.

A fourth limitation is the intrinsic variability of generative systems. Even for the same case, changes in model, version, temperature, prompt, memory, and interface may change both output and cognitive effect. Research should describe the full technical context of the interaction. Finally, neither prompt engineering nor knowledge-base integration should be treated as a sufficient safeguard. Safeguards should be embedded in interfaces, workflows, training, governance, and regulatory evaluation.²⁰

Figure 1 is a heuristic visualization rather than a measured distribution, and the TPI definition, weighting options, and expert-adjudication procedure require reliability and validity testing before

clinical interpretation. Future studies should evaluate the Tail-Preservation Index for inter-rater reliability, construct validity, responsiveness across different clinical scenarios, and its association with clinically meaningful outcomes.

What this manuscript adds

Prior work has examined diagnostic bias, human-AI collaboration, and generative verification largely separately. This manuscript integrates them through a named process-level construct, a literature-mapped mnemonic, candidate measures, and a governance focus on AI-related changes in clinician reasoning. These are testable proposals, not established terms or validated instruments.

- Introduces AI-associated differential diagnosis narrowing as a cognitive-safety construct in the clinician-AI system.
- Distinguishes model diagnostic accuracy from preservation of human differential reasoning.
- Defines potentially unsafe narrowing as early restriction with insufficient verification and counterfactual testing, distinguishing it from appropriate evidence-based narrowing.
- Reorganizes generative techniques already described in the literature into a heuristic framework for cognitive safeguarding, without presenting them as an already validated clinical protocol.
- Proposes candidate measures for differential breadth, retention of dangerous diagnoses, anchoring, verification, premature closure, and the Tail-Preservation Index.
- Shifts regulatory evaluation of clinical AI from output alone to modification of the clinician's cognitive process.
- Extends the concept of cognitive safety to internal medicine, community care, and general practice, including referral thresholds, reassessment, safety-netting, and longitudinal management of uncertainty.

Conclusions

Generative AI brings a diagnostic-safety question to the foreground: not only whether the model reaches the correct diagnosis, but whether it preserves the clinical reasoning that leads toward it. Medicine is not a competition between final outputs. It is disciplined management of uncertainty, hypothesis verification, and protection from premature closure.

AI-associated differential diagnosis narrowing describes a possible risk of the clinician-AI system: AI may appear useful while prematurely restricting the physician's diagnostic field. The construct should be tested, not assumed. Its value depends on whether it leads to better metrics, studies, interfaces, and workflows.

The point is not to encourage distrust of AI, but to design and use systems that respect clinical method: generate hypotheses, test them, search for contradictions, preserve dangerous alternatives, and narrow the differential only when narrowing is clinically justified. Cognitive safety should therefore be evaluated alongside model performance as an endpoint of AI-assisted medicine.

References

1. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med* 2005;165:1493-9.
2. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med* 2003;78:775-80.
3. Committee on Diagnostic Error in Health Care; Board on Health Care Services; Institute of Medicine, et al. Improving diagnosis in health care. Washington, DC: National Academies Press; 2015.
4. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012;19:121-7.
5. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017;24:423-31.

6. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. *JAMA* 2023;330:2255-7.
7. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. *Hum Factors* 1997;39:230-53.
8. Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors* 2004;46:50-80.
9. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open* 2024;7:e2440969.
10. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *Nature* 2025;642:451-7.
11. Cabitza F, Campagner A, Ronzio L, et al. Rams, hounds and white boxes: investigating human-AI collaboration protocols in medical diagnosis. *Artif Intell Med* 2023;138:102506.
12. Dhuliawala S, Komeili M, Xu J, et al. Chain-of-verification reduces hallucination in large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. Bangkok: Association for Computational Linguistics; 2024. pp. 3563-78.
13. Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations; 2023. Available from: <https://openreview.net/forum?id=1PL1NIMMrw>.
14. Madaan A, Tandon N, Gupta P, et al. Self-refine: iterative refinement with self-feedback. *Adv Neural Inf Process Syst* 2023;36:46534-94.
15. Yao S, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. *Adv Neural Inf Process Syst* 2023;36:11809-22.
16. Du Y, Li S, Torralba A, et al. Improving factuality and reasoning in language models through multiagent debate. *Proc Mach Learn Res* 2024;235:11733-63.
17. Savage T, Nayak A, Gallo R, et al. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit Med* 2024;7:20.
18. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024;30:2613-22.
19. Schmidgall S, Harris C, Essien I, et al. Evaluation and mitigation of cognitive biases in medical language models. *NPJ Digit Med* 2024;7:295.
20. Di Bello F, Angioni D. Evidence-like error and physician responsibility in retrieval-augmented clinical artificial intelligence. *AboutOpen* 2026;13:22-30.
21. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. *Med Educ* 2008;42:468-75.
22. Mamede S, van Gog T, van den Berge K, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA* 2010;304:1198-203.
23. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. *Nat Med* 2026;32:1152-9.
24. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023;29:1930-40.
25. Tirotta D, Di Bello F, Dentali F, et al. FADOI official position on artificial intelligence in internal medicine. *Ital J Med* 2026;20:2482.
26. Wang G, Zhang K, Jiang J, et al. Human-large language model collaboration in clinical medicine: a systematic review and meta-analysis. *NPJ Digit Med* 2026;9:195.

Table 1. Proposed mechanisms, cognitive pathways, clinical risks, and safeguards in artificial intelligence-associated differential diagnosis narrowing.

Possible mechanism	Cognitive pathway	Potential clinical risk	Candidate safeguard
Initial-output anchoring	First AI output frames subsequent hypotheses	Premature convergence	Record a provisional differential before exposure and compare alternative interaction orders in studies
Perceived authority of fluent output	Fluency is mistaken for reliability	Overestimation of output	Require explicit uncertainty and limitations
False reassurance	Plausible narrative reduces perceived risk	Delay in serious diagnoses	Search for red flags, contrary data, and missing information
Reduced salience of rare alternatives	Probability ranking hides severity and time-dependence	Omission of rare-but-dangerous diagnoses	Rank separately by probability, danger, treatability, and urgency
Loss of clinical context	Complex story becomes a few salient elements	Loss of temporal or atypical signals	Highlight trajectory, discordance, and missing data
Displaced verification	Clinician delegates verification to the system	Circular confirmation	Use chain-of-verification with supports, contradictions, missing data, and red flags

Table 2. GENERATE-Dx components mapped to their conceptual derivation.

Mnemonic step	Observable operation	Conceptual derivation	Principal target
G - Generate	Record an independent provisional differential	Interaction order and reflective reasoning ^{11,21,22}	Initial-output anchoring
E - Explicitly separate	Separate likelihood, danger, urgency, and actionability	Risk-based differential diagnosis ¹⁻³	Failure to prioritize danger
N - Note	Identify missing, discordant, and time-dependent data	Diagnostic error and longitudinal diagnosis ¹⁻³	Context loss
E - Examine	Check supporting evidence, contradictions, and alternatives	Verification and reflective reasoning ^{12,21,22}	Confirmation and false reassurance
R - Reopen	Use counterfactual and adversarial review	Self-refinement, tree search, and debate ^{14-16,19}	Premature closure
A - Audit	Audit anchoring, automation bias, and overconfidence	Automation bias and human factors ⁴⁻⁸	Inappropriate reliance
T - Test	Test do-not-miss diagnoses and escalation thresholds	Diagnostic safety and differential evaluation ^{3,10}	Dangerous omission
E - Establish	State reassessment, safety-netting, and residual uncertainty	Diagnosis as a process across time ³	Unmonitored uncertainty

Table 3. Candidate metrics for empirical evaluation.

Candidate metric	Operational definition	Possible measure
Differential breadth	Number and semantic diversity of clinically relevant diagnoses	Diagnosis count plus expert-reviewed semantic clustering
Rare-but-dangerous retention	Proportion of expert-defined do-not-miss diagnoses retained	Case-specific blinded expert reference set
Tail-Preservation Index	Weighted retention of pre-AI, non-dominant, high-relevance diagnoses	Clinician-case TPI; equal weights in the primary analysis and prespecified weighted sensitivity analysis
Anchoring shift score	Shift of clinician ranking toward first AI output	Pre/post ranking distance
Counterfactual exploration score	Quality of hypotheses against leading diagnosis	Expert rubric
Verification completeness	Coverage of support, contradiction, missing data, and red flags	Structured checklist scored by blinded reviewers
Premature closure rate	Closure before adequate exploration	Blinded assessment of reasoning trace
Confidence calibration error	Gap between confidence and correctness	Brier score or calibration curves
Management and reassessment	Appropriateness of testing, referral, safety-netting, and follow-up	Blinded expert rating and time-to-action endpoints
Cognitive load cost	Load added by safeguards	National Aeronautics and Space Administration Task Load Index (NASA-TLX) or equivalent scale

Heuristic reweighting of diagnostic hypotheses

HEURISTIC ILLUSTRATION — NOT EMPIRICAL DATA

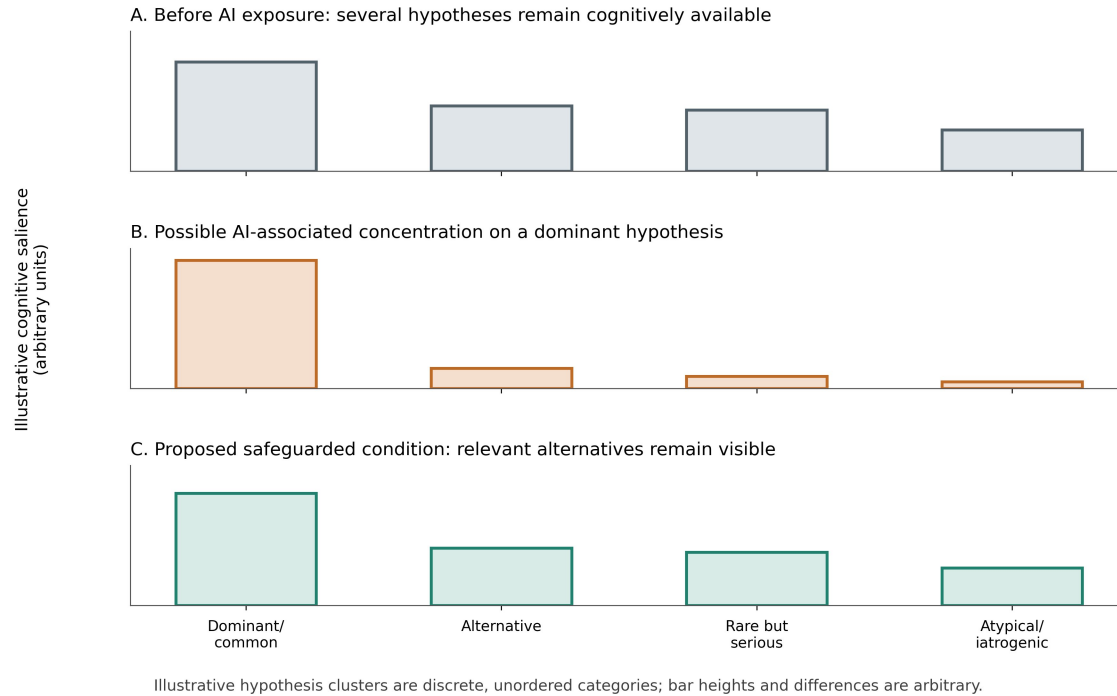


Figure 1. Heuristic reweighting of diagnostic hypotheses before and after artificial intelligence exposure. The author-generated bars are schematic and arbitrary; no patient or clinician data were used. The four hypothesis clusters are discrete, illustrative, and unordered, and their placement does not define a common clinical scale. Bar heights and differences represent possible changes in cognitive salience, not observed probabilities, distributions, effect sizes, rankings, or distances. Panel A shows several hypotheses before exposure; panel B shows a possible AI-associated concentration on a dominant hypothesis; panel C shows a proposed safeguarded condition in which clinically relevant alternatives remain visible. Bar heights and category ordering are not to scale.

Proposed pathway for AI-associated differential diagnosis narrowing

TESTABLE CONCEPTUAL MODEL - ARROWS DO NOT ESTABLISH CAUSATION

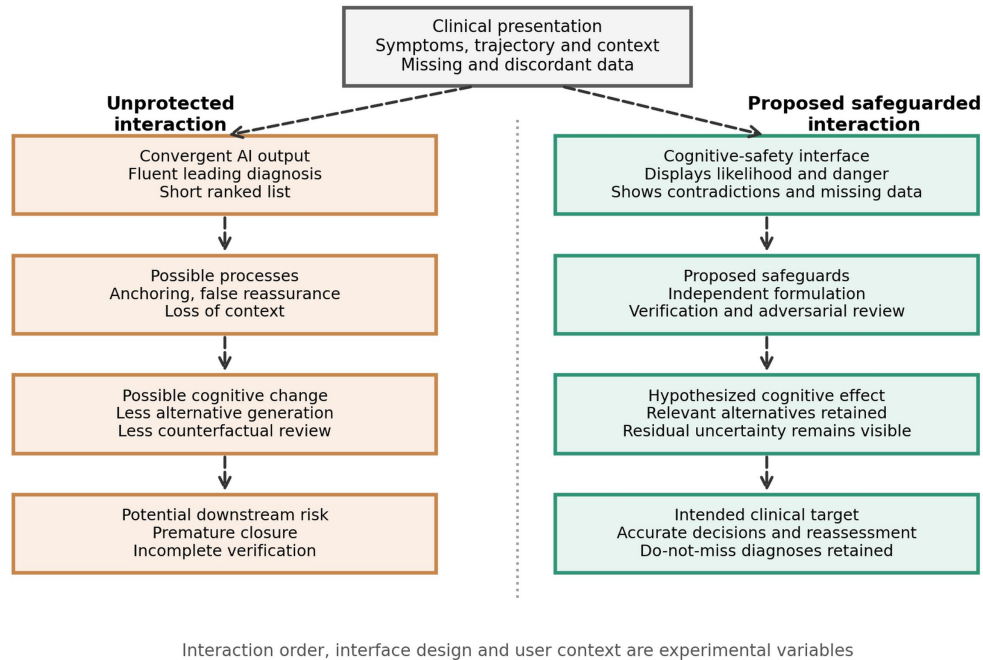


Figure 2. Proposed and testable pathway for AI-associated differential diagnosis narrowing. Arrows express hypotheses for empirical testing, not established causal relationships. The unprotected branch illustrates how a convergent AI output might be associated with anchoring, false reassurance, reduced alternative generation, and premature closure. The safeguarded branch illustrates interface features hypothesized to preserve relevant alternatives, contradictory evidence, and residual uncertainty. The intended target is accurate, appropriately narrowed reasoning with reassessment and retention of do-not-miss diagnoses.