EXPLAINABLE AI FOR THERAPEUTIC DECISION-MAKING AND PRESCRIPTION SAFETY: A LONGITUDINAL FRAMEWORK FOR CLINICAL DECISION SUPPORT

Abstract

Background: Clinical artificial intelligence has been evaluated mainly through diagnosis, triage, image interpretation, and isolated question answering. Therapeutic decision-making has a different structure: it converts clinical reasoning into action through drug choice, dose, route, timing, contraindication screening, monitoring, reassessment, escalation, de-escalation, and discontinuation. An explainable system that names a diagnosis but does not account for this action chain remains incomplete as clinical decision support.

Materials and Methods: A DOI-indexed narrative review was conducted across PubMed, Nature Portfolio, ACM Digital Library, ScienceDirect, SAGE, JMIR, Radiology, BMJ, JAMA, and Intelligent Medicine sources. The search focused on AI-assisted prescription, treatment recommendations, clinical decision support, explainable artificial intelligence, clinical LLM agents, clinician acceptance, reporting standards, human-AI collaboration, and workflow-level evaluation. Sources without DOI metadata were used only for orientation and were not cited as evidence.

Results: The literature indicates that AI recommendations influence prescribing behavior, that clinicians negotiate treatment recommendations rather than simply accept or reject them, and that large language model agents require workflow-level evaluation rather than static medical-question benchmarks. Recent reporting guidance for prediction models and LLM studies reinforces the need for transparent, reproducible, clinically applicable evaluation. Therapeutic XAI therefore requires a longitudinal structure linking indication, safety screening, dosing, timing, monitoring, and revision.

Conclusion: Prescription safety should become a primary endpoint for clinical XAI. A useful therapeutic explanation should state why an intervention is indicated, when it should be withheld, how it should be monitored, and which clinical changes require revision. The proposed Longitudinal Therapeutic XAI Framework translates explanation from a justification of advice into a safety layer for action.

Keywords: Explainable Artificial Intelligence; Therapeutic Decision Support; Prescription Safety; Clinical AI Agents; Treatment Recommendations; Large Language Models; Human-AI Collaboration; Clinical Workflow.

Introduction:

The most visible promise of clinical artificial intelligence is diagnostic acceleration. Models read images, rank differential diagnoses, classify risk, retrieve guidelines, summarize patient information, and generate case explanations. These functions matter, yet they stop before the act that exposes the patient to benefit or harm. In clinical practice, the physician does not treat a probability. The physician prescribes, withholds, monitors, escalates, de-escalates, and revises. Therapeutic decision-making is not a single output field placed after diagnosis. It is a longitudinal process in which the initial intervention remains conditional on physiology, comorbidity, laboratory evolution, drug interaction, organ function, patient preference, institutional protocol, and treatment response.

This distinction changes how explainable artificial intelligence should be designed and evaluated. Diagnostic XAI explains why a system suggests that a patient has sepsis, pulmonary embolism, pneumonia, melanoma, or myocardial infarction. Therapeutic XAI must explain why a specific action follows from that conclusion, why the same action is unsafe in a different patient, what parameter should be checked next, and at what point the plan should change. A system that correctly identifies sepsis but fails to specify antimicrobial timing, hemodynamic monitoring, source-control logic, lactate reassessment, renal dosing, allergy risk, or escalation criteria contributes only partially to care.

The literature on AI-assisted decision-making already demonstrates that AI advice changes clinician behavior. In intensive care prescribing tasks, AI suggestions shifted physician prescriptions, while simple feature-importance explanations did not exert a stronger influence than AI advice alone [1]. In treatment-focused sepsis decision support, clinicians did not behave as binary acceptors or rejectors of algorithmic advice; they negotiated recommendations, accepting, delaying, prioritizing, or discarding components of the suggested plan [2]. These findings are important because therapeutic reasoning naturally decomposes into parts. A physician may accept the indication but reject the dose, accept the antimicrobial class but delay de-escalation, accept the need for vasopressor support but demand further volume assessment, or accept anticoagulation but modify the agent because of renal function.

Large language models add a second layer to this problem. They are no longer evaluated only as static systems that answer medical questions. Recent work frames them as agents capable of interacting with clinical environments, users, tools, and sequential tasks [3,6]. This framing is closer to real care, but it also raises the standard for safety. If an agent influences a treatment path, it must be judged by the quality of that path, not by the fluency of isolated answers. Evidence also shows that current large language models remain limited for autonomous clinical decision-making, especially when instruction-following, order sensitivity, laboratory interpretation, guideline adherence, and clinical context are tested under realistic conditions [4].

The core gap is therefore not the lack of another diagnostic benchmark. The gap is the absence of a practical evaluation structure for therapeutic XAI: a structure that measures whether AI explanations improve the safety,

completeness, and revisability of clinical action. This review proposes such a structure. It synthesizes DOI-indexed literature on AI prescription behavior, treatment recommendation acceptance, clinical agents, human-AI collaboration, reporting standards, and XAI design. It then introduces a Longitudinal Therapeutic XAI Framework that treats explanation as a safety layer across the entire treatment cycle.

Materials and Methods:

This manuscript is a narrative review with a conceptual framework. The aim was not to pool effect sizes or perform a meta-analysis. The available literature combines experimental prescription simulations, qualitative studies, stakeholder-driven reviews, clinical decision support studies, large language model evaluations, reporting guidelines, and human-AI collaboration research. These designs are heterogeneous in task, participant profile, clinical domain, model type, explanation format, and outcome measurement. A narrative synthesis was therefore selected to identify mechanisms, gaps, and framework components relevant to therapeutic decision-making.

The search covered PubMed, Nature Portfolio, ACM Digital Library, ScienceDirect, SAGE Journals, JMIR, Radiology, BMJ, JAMA, Intelligent Medicine, and publisher records indexed through DOI metadata. Searches were conducted using combinations of the following terms: artificial intelligence, explainable artificial intelligence, clinical decision support, prescription decision-making, treatment recommendations, sepsis treatment, clinician acceptance, human-AI collaboration, large language model agents, medical decision-making, prescription safety, trust calibration, workflow evaluation, TRIPOD+AI, and TRIPOD-LLM.

The inclusion criteria were: (i) DOI-indexed article; (ii) relevance to clinical AI, XAI, therapeutic or diagnostic decision support, prescription behavior, clinician acceptance, human-AI collaboration, medical LLM agents, or reporting standards; (iii) publication from 2023 to 2026 whenever recent literature was available; and (iv) older DOI-indexed work when it provided a foundational concept for trust, reliance, clinical decision support, or collaboration protocols. The exclusion criteria were: no DOI, purely technical model development without clinician interaction or clinical decision relevance, non-medical AI papers without transferable decision-support mechanisms, and editorials or web documents used only for orientation.

The selected literature was mapped into five analytic domains: influence of AI advice on treatment decisions; clinician acceptance and negotiation of treatment recommendations; agentic and workflow-level evaluation of LLMs; reporting standards for clinical AI and LLM studies; and translation of XAI design into prescription safety. Each study was read for its task, clinical setting, participant type, intervention, explanation format, outcome, and contribution to the proposed framework.

Table 1: Search strategy and eligibility logic for the review.

Component	Operational definition in this review
Databases and sources	PubMed, Nature Portfolio, ACM Digital Library, ScienceDirect, SAGE Journals, JMIR, Radiology, BMJ, JAMA, Intelligent Medicine, and DOI-indexed publisher records.
Core concepts	AI-assisted prescription, treatment recommendations, clinical decision support, explainable AI, clinician acceptance, LLM agents, workflow evaluation, human-AI collaboration, prescription safety.
Inclusion criteria	DOI-indexed source; direct relevance to clinical AI, XAI, clinician behavior, therapeutic decision-making, prescription safety, reporting standards, or agentic workflows.
Exclusion criteria	No DOI; technical model-only article without clinician-facing decision relevance; non-clinical AI paper without transferable decision-support mechanism.
Synthesis method	Narrative thematic synthesis organized around therapeutic action, clinician negotiation, workflow-level agent evaluation, reporting quality, and safety-oriented explanation design.

Results:

AI recommendations modify therapeutic behavior, but explanation alone is not enough

The strongest entry point for therapeutic XAI is the observation that AI advice changes prescribing. Nagendran and colleagues examined prescription decision-making under four conditions: baseline, peer clinician information, AI recommendation, and AI recommendation with explanation [1]. The study showed that additional information influenced prescriptions, with AI producing a significant shift. Yet the simple explanation did not outperform AI advice alone. This finding is not a minor technical detail. It indicates that the mere addition of an explanation does not automatically create safer or more clinically reflective therapeutic behavior.

For prescription safety, this result has two implications. First, advice source matters. Clinicians respond differently when the recommendation is framed as AI advice rather than peer information. Second, explanation format matters. A simple feature-importance explanation attached to a prescription recommendation is not equivalent to a therapeutic rationale. It may tell the physician which variables influenced the model, but it does not necessarily answer

the clinical questions that determine safe action: why this drug, why this dose, why now, why not another route, what to monitor, and when to revise.

This distinction defines the boundary between explanatory transparency and therapeutic utility. A model explanation that highlights lactate, blood pressure, creatinine, or oxygen requirement may be informative. A therapeutic explanation must go further. It must convert these features into a safe plan: antimicrobial initiation, fluid or vasopressor logic, renal dose adjustment, anticoagulation threshold, monitoring interval, and stopping criteria. Prescription is not an interpretability problem alone. It is an action-governance problem.

Clinicians negotiate treatment advice rather than accept or reject it

Treatment recommendations enter a clinical mind already populated by competing priorities. A physician may agree with the direction of a recommendation while disagreeing with its intensity, timing, route, or sequence. The study by Sivaraman and colleagues is therefore central to the present review [2]. In AI-based treatment recommendations for sepsis, clinicians did not merely ignore or trust the system. They negotiated the recommendation. They selected components to follow, components to delay, and components to reject. This behavior is closer to clinical reality than the conventional measurement of "acceptance" as a binary endpoint.

Negotiation is not failure of adoption. It is the normal structure of therapeutic reasoning. A sepsis recommendation may include broad-spectrum antimicrobials, source control, fluid resuscitation, vasopressors, blood cultures, lactate reassessment, and hemodynamic monitoring. Each component has a different evidentiary basis, urgency, and safety profile. The physician may accept antimicrobial timing while changing the agent because of allergy. The physician may accept lactate monitoring while rejecting additional fluids because of heart failure. A therapeutic XAI system that cannot represent this partial agreement will misclassify clinical behavior.

This explains why therapeutic decision support needs a different interface logic from diagnostic support. The system should not ask only whether the physician accepts the recommendation. It should permit structured modification: accept indication, modify dose, delay execution, request more data, substitute agent, add monitoring, set reassessment time, and record clinical reason. The explanation should support this negotiation instead of treating deviation as non-compliance.

LLM agents require workflow-level evaluation, not static answer evaluation

Large language models change the scale of the problem because they can act as agents within simulated or real workflows. Mehandru and colleagues argued that LLM agents should be evaluated through their impact on clinical tasks and workflows rather than by benchmark question-answer performance alone [3]. Schmidgall and colleagues extended this concern through AgentClinic, a multimodal benchmark for tool-using clinical AI agents in simulated clinical environments, including patient interaction, multimodal data collection, incomplete information, and tool use [6]. This type of evaluation is closer to therapy because treatment decisions require sequential data acquisition and revision.

Hager and colleagues provide a necessary corrective to excessive enthusiasm about autonomous clinical decision-making by LLMs [4]. Their analysis indicates that current models remain limited in diagnostic accuracy across pathologies, guideline adherence, instruction-following, sensitivity to information order, and laboratory interpretation. These limitations are not abstract when the model influences therapy. A model that misreads creatinine, ignores anticoagulation risk, mishandles hypotension, or fails to revise a plan after new information creates action-level risk.

Recent evaluations of LLMs and agents in healthcare emphasize that assessment must account for data sources, manually designed clinical questions, workflow complexity, hallucination risk, tool use, and clinical context [5]. This supports the central claim of this review: therapeutic AI should be assessed as a sequence of clinically accountable actions. A correct answer to a board-style question does not demonstrate safe prescription behavior. A safe therapeutic agent must understand when a treatment is indicated, when it is unsafe, and when the plan is obsolete.

Table 2: DOI-indexed literature mapped to the therapeutic XAI problem.

Study	Main contribution to this article	Therapeutic XAI implication
Nagendran et al. [1]	AI recommendations significantly influenced prescription behavior; simple XAI did not exceed AI advice alone.	Prescription is a behavioral endpoint; explanation must be therapeutic, not only model-oriented.
Sivaraman et al. [2]	Clinicians negotiated AI treatment recommendations by following, rejecting, or delaying components.	Acceptance must be modeled as structured amendment, not a binary decision.
Mehandru et al. [3]	LLM agents should be evaluated in clinical workflows, not only on static benchmarks.	Therapeutic AI needs sequential assessment across action and revision.
Hager et al. [4]	LLMs are not ready for autonomous clinical decision-making and remain sensitive to clinical information structure.	Autonomous prescription is unsafe without oversight, validation, and audit.
Schmidgall et al. [6]	AgentClinic evaluates multimodal, tool-using clinical AI	Static question-answering benchmarks should not be

Study	Main contribution to this article	Therapeutic XAI implication
	agents under sequential simulated clinical interactions.	treated as sufficient for action-level clinical safety.
Collins et al. and Gallifant et al. [7,8]	TRIPOD+AI and TRIPOD-LLM provide reporting guidance for AI and LLM studies.	Therapeutic XAI studies must report data, task, model, user, workflow, and clinical applicability with transparency.

Reporting standards strengthen, but do not replace, therapeutic evaluation

The addition of reporting standards changes how clinical AI studies should be judged. TRIPOD+AI provides harmonized reporting guidance for clinical prediction models using regression or machine learning, including expanded reporting recommendations and an abstract checklist [7]. TRIPOD-LLM extends this logic to large language model studies and aims to improve reproducibility and clinical applicability in research that uses LLMs [8]. These statements do not solve therapeutic safety by themselves, but they expose an important standard: if the study does not report the clinical task, data source, intended use, user role, model behavior, and outcome in sufficient detail, readers cannot determine whether the system is ready to influence care.

For therapeutic XAI, the reporting problem is even sharper. A study may describe the model and explanation while omitting the actual order behavior. It may report agreement with AI advice without reporting dose, timing, monitoring, contraindication detection, or revision after new data. It may report diagnostic accuracy without reporting whether the action plan became safer. A therapeutic XAI study should therefore apply reporting standards while adding action-level items: prescription correctness, safety exclusions, monitoring completeness, and revision behavior.

This review uses reporting guidance not as a checklist for the current manuscript, but as evidence that clinical AI has moved beyond general claims of performance. The same movement is needed in XAI for therapy. The field needs transparent studies that show what the AI saw, what it recommended, what explanation it provided, what the clinician did, and whether the final therapeutic plan was clinically safer.

Human-AI collaboration protocols remain underdeveloped for therapy

Human-AI collaboration protocols have been studied in medical diagnosis, including different temporal arrangements of AI and human input [13]. These protocols are useful because they move beyond model accuracy and ask how the human and the system should interact. Therapeutic decision-making requires a comparable protocol, but with additional complexity. The physician is not only choosing a label. The physician is choosing an intervention under uncertainty, with consequences that require surveillance.

Cabitza and colleagues also describe pro-hoc explanations as support that offers alternative explanations for possible outcomes rather than a single post-hoc justification for a specific recommendation [14]. This design is relevant to therapeutic safety because treatment decisions often hinge on competing action paths. For instance, the question is not only whether a patient has bacterial pneumonia. The question is whether to initiate broad-spectrum antibiotics, narrow therapy, hold nephrotoxic agents, escalate respiratory support, or seek source control. A pro-hoc therapeutic explanation would compare action paths with their evidentiary trade-offs.

Stakeholder-driven work on XAI design indicates that clinicians need explanations adapted to user expertise, task complexity, timing of prediction, and case-specific demands [15]. This point is especially important in therapy. A senior intensivist may need a compact safety warning and revision trigger. A junior physician may need dose logic, contraindication checks, and monitoring prompts. A pharmacist may need interaction and renal dosing details. A therapeutic XAI system therefore requires role-sensitive explanation rather than a single explanatory template.

A Longitudinal Therapeutic XAI Framework:

The proposed Longitudinal Therapeutic XAI Framework treats explanation as a safety layer across the treatment cycle. It does not define XAI as an additional paragraph attached to a recommendation. It defines XAI as a structured account of why an intervention is indicated, how it should be executed, what could make it unsafe, what evidence should be monitored, and what event requires revision.

The framework has seven layers: indication, safety exclusion, dose and route, timing, monitoring, revision, and accountability. Each layer corresponds to a clinical failure mode. Indication prevents treatment without sufficient reason. Safety exclusion prevents contraindicated action. Dose and route prevent execution errors. Timing prevents delayed or premature action. Monitoring prevents blind continuation. Revision prevents obsolete treatment. Accountability prevents untraceable automation.

Imagem 1.Longitudinal Therapeutic XAI Framework

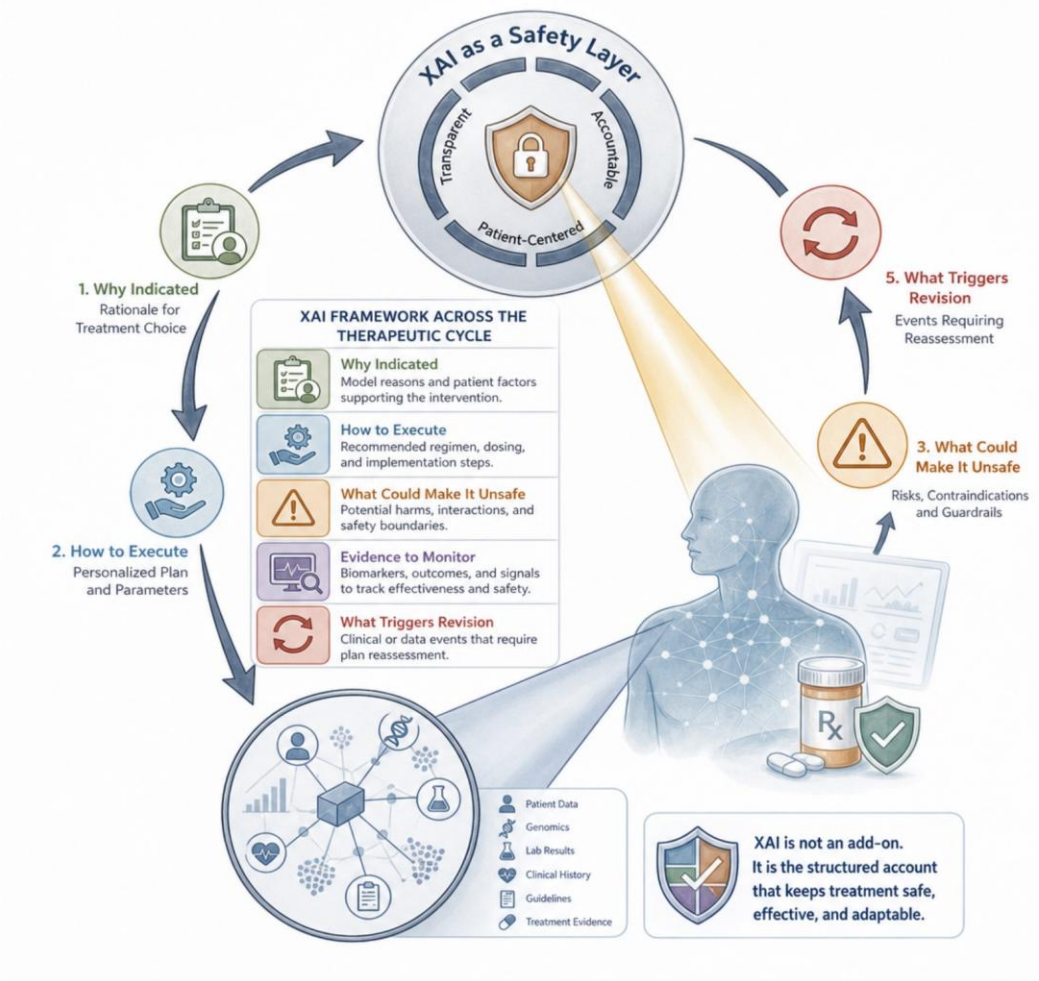

Table 3: Longitudinal Therapeutic XAI Framework.

Layer	Clinical question	Required explanation element	Safety failure addressed
1. Indication	Why treat this patient now?	Clinical features, threshold, guideline logic, diagnostic uncertainty.	Treatment without sufficient indication or delayed treatment despite high-risk findings.
2. Safety exclusion	What makes this treatment unsafe?	Contraindications, allergies, renal/hepatic function, interactions, bleeding risk, pregnancy, instability.	Harm from contraindicated or poorly screened intervention.
3. Dose and route	How should the intervention be executed?	Dose, route, interval, adjustment logic, maximum dose, administration constraints.	Wrong dose, wrong route, wrong interval, failure to adjust to organ function.
4. Timing	When should the action occur?	Immediate, time-critical, conditional, post-test, post-stabilization, or deferred execution.	Delayed sepsis therapy, premature anticoagulation, unnecessary escalation.
5. Monitoring	What will show benefit or harm?	Vital signs, laboratory tests, drug levels, cultures, imaging, adverse events, response interval.	Continuation without evidence of response or detection of toxicity.
6. Revision	When should the plan change?	Failure criteria, escalation/de-escalation triggers, stop rules, new data thresholds.	Obsolete plans maintained after patient status changes.
7. Accountability	Who accepted, changed, or rejected the advice?	Clinician amendment, reason for override, audit trail, uncertainty note.	Untraceable automation, unclear responsibility, weak post-event review.

Operational endpoints for therapeutic XAI:

Therapeutic XAI requires endpoints that match clinical action. Diagnostic accuracy alone is insufficient. A system could improve diagnostic naming while worsening therapeutic safety if it increases unnecessary treatment, omits contraindication screening, or reduces clinician monitoring. Conversely, a system could improve safety even when it does not change the final diagnosis, by prompting renal dose adjustment, earlier reassessment, or safer de-escalation.

The core endpoints should include treatment appropriateness, dose correctness, timing, safety exclusions, monitoring completeness, revision responsiveness, and clinician amendment quality. These endpoints are measurable in simulated cases, retrospective chart review, prospective silent-mode testing, and staged deployment. They also allow comparison across conditions: no AI, AI without explanation, static XAI, and longitudinal therapeutic XAI.

Table 4: Suggested outcome measures for future empirical testing.

Endpoint	Definition	Example measure
Therapeutic appropriateness	Whether the chosen intervention matches case severity, diagnosis, and guideline logic.	Correct initial antimicrobial class in sepsis vignette; appropriate anticoagulation decision in pulmonary embolism scenario.
Dose safety	Whether dose, route, interval, and adjustment match patient-specific risk.	Renal-adjusted antibiotic dose; safe anticoagulant dosing; avoidance of nephrotoxic combination.
Timing quality	Whether action occurs within the clinically relevant time window.	Time to antibiotic recommendation; delay before vasopressor escalation; timing of lactate reassessment.
Contraindication detection	Whether patient-specific safety barriers are recognized before action.	Allergy, bleeding risk, pregnancy, renal failure, hepatic dysfunction, QT prolongation, drug interaction.
Monitoring completeness	Whether the plan includes response and toxicity surveillance.	Repeat lactate, urine output, creatinine, cultures, anti-Xa level, potassium, mental status, oxygen requirement.
Revision responsiveness	Whether treatment changes when new evidence appears.	De-escalation after culture; stopping rule after adverse event; escalation after persistent shock.
Clinician negotiation quality	Whether acceptance, modification, delay, or rejection is clinically justified.	Structured override reason; documented uncertainty; request for additional evidence before execution.

Therapeutic explanation as action grammar:

A therapeutic explanation has a different grammar from a diagnostic explanation. Diagnostic explanation usually relates input findings to a disease label. Therapeutic explanation relates patient state to an intervention, then relates the intervention to risks, execution parameters, surveillance, and discontinuation. This grammar is closer to the way clinicians document orders and hand over patients. It is also closer to the way errors occur. A wrong diagnosis is one failure mode. A right diagnosis followed by an unsafe dose, missing contraindication, late escalation, absent monitoring, or failure to revise is another.

This action grammar can be described through six verbs: indicate, exclude, dose, time, monitor, and revise. A system that recommends antibiotics should indicate the suspected source, exclude allergy and renal constraints, specify regimen and dose, define urgency, request cultures and response markers, and state revision triggers. A system that recommends anticoagulation should indicate the thrombotic risk, exclude bleeding and procedural constraints, specify agent and dose, define initiation timing, monitor complications, and state when to stop, bridge, reverse, or escalate. These verbs convert explanation into a clinical order logic.

The distinction also changes how XAI failures should be classified. A model-centered explanation can be technically faithful to the model while clinically incomplete. It may truthfully report that creatinine, lactate, age, and oxygen saturation drove the prediction, yet still fail to explain how renal function changes the prescription. Therapeutic XAI needs clinical faithfulness as well as model faithfulness. Clinical faithfulness means that the explanation maps the recommendation onto the constraints that govern safe bedside action.

Table 5: Difference between diagnostic XAI and therapeutic XAI.

Dimension	Diagnostic XAI	Therapeutic XAI
Primary object	Disease label, probability, image class, risk score, or diagnostic category.	Action plan, prescription, timing, monitoring, and revision path.
Main user question	Why does the system think this diagnosis or risk category is likely?	Why is this intervention indicated, safe, timed correctly, and revisable in this patient?
Core failure mode	Wrong label, missed diagnosis, misleading confidence, incomplete differential.	Wrong treatment, unsafe dose, missed contraindication, absent monitoring, delayed revision.
Explanation requirement	Evidence supporting the diagnostic output and alternative diagnoses.	Indication, contraindication, dose, route, timing, monitoring, stop rules, escalation triggers.
Evaluation endpoint	Diagnostic accuracy, sensitivity, specificity, calibration, advice-taking.	Treatment appropriateness, dose safety, timing, monitoring completeness, revision responsiveness.

Mechanisms of prescription risk in AI-supported therapy:

Therapeutic AI creates risks that are not captured by diagnostic performance. The first is dose displacement. A physician may accept a model-suggested therapeutic direction but carry forward a dose that ignores renal function, body weight, age, hepatic failure, or drug interaction. The second is timing distortion. A system may make the correct recommendation but fail to communicate urgency, leading to late execution in sepsis or premature execution in a patient requiring confirmation before high-risk anticoagulation. The third is monitoring omission. The recommendation enters the chart, but the surveillance plan does not. This is common in high-risk medication use, where the harm emerges after the order, not at the moment of choosing the drug.

The fourth mechanism is inappropriate continuation. A plan that was initially reasonable becomes unsafe after culture results, bleeding, kidney injury, hemodynamic deterioration, or failure of response. Static recommendation systems are poorly suited to this reality unless they include revision triggers. The fifth mechanism is partial automation bias. The clinician rejects the obviously unsafe portion of advice but accepts a less visible unsafe portion, such as interval, duration, or monitoring interval. This reinforces the need to study therapy as a bundle of actions rather than as a single accept/reject choice.

The sixth mechanism is responsibility diffusion. If the AI suggests a plan, the physician modifies it, a nurse executes it, and a pharmacist later intervenes, the decision chain becomes distributed. Therapeutic XAI should therefore preserve the rationale for each modification. This is not merely medico-legal documentation. It is a learning mechanism: future system improvement depends on knowing where the clinician agreed, where the clinician changed the plan, and which change prevented harm.

Table 6: Prescription risk mechanisms and corresponding XAI safeguards.

Risk mechanism	Clinical manifestation	XAI safeguard
Dose displacement	Correct therapeutic direction with unsafe dose, interval, or route.	Dose logic tied to weight, renal function, hepatic function, age, and interaction checks.
Timing distortion	Right intervention executed too late or too early.	Urgency label and conditional timing statement linked to clinical threshold.
Monitoring omission	Treatment ordered without response or toxicity surveillance.	Required monitoring bundle with interval, parameter, and action threshold.
Inappropriate continuation	Plan remains active despite new data or adverse event.	Revision triggers, stop rules, and de-escalation prompts.
Partial automation bias	Clinician modifies visible part of advice but accepts hidden unsafe component.	Decomposed acceptance by indication, dose, timing, monitoring, and revision.
Responsibility diffusion	Unclear accountability across physician, pharmacist, nurse, and AI tool.	Human amendment log and audit trail for each accepted or changed component.

Clinical translation across high-risk treatment domains:

The framework is most valuable in clinical scenarios where treatment is time-sensitive and patient-specific safety constraints are common. Sepsis is an obvious example because therapy depends on early antimicrobial initiation, source suspicion, hemodynamic status, lactate, organ dysfunction, and reassessment. A diagnostic AI system that suggests sepsis but does not guide treatment timing, antimicrobial logic, cultures, fluid or vasopressor decision-making, and reassessment remains clinically incomplete.

Pulmonary embolism provides a different test. A system may correctly suggest the diagnosis, yet prescription safety depends on anticoagulation risk, renal function, hemodynamic status, pregnancy, bleeding risk, imaging confirmation, and thrombolysis criteria. The therapeutic explanation must distinguish diagnostic confidence from treatment threshold. It must also specify which data are missing before anticoagulation, what monitoring is required after initiation, and when escalation is warranted.

Diabetic ketoacidosis, acute heart failure, sedation in the intensive care unit, antimicrobial stewardship, and postoperative analgesia create similar demands. In each case, the safe plan is not exhausted by naming the disease. The explanation must carry the action from initial indication to monitoring and revision. A longitudinal design also prevents a common error in decision support: treating the first recommendation as final even after new information emerges.

Table 7: Clinical translation of the framework across high-risk treatment domains.

Clinical domain	Therapeutic explanation requirement	Example safety question
Sepsis	Source, antimicrobial timing, cultures, hemodynamics, lactate reassessment, renal dose, de-escalation.	Is the recommended antimicrobial safe for renal function and local resistance context?
Pulmonary embolism	Anticoagulation threshold, bleeding risk, imaging status, hemodynamic risk, thrombolysis criteria, monitoring.	Is treatment justified before confirmation, and what bleeding risk modifies the agent or timing?
Diabetic ketoacidosis	Fluid sequence, potassium logic, insulin timing, glucose transition, anion gap monitoring.	Is potassium known before insulin, and what parameter allows transition from acute protocol?

Clinical domain	Therapeutic explanation requirement	Example safety question
Acute heart failure	Volume status, blood pressure, diuretic choice, renal surveillance, oxygen support, escalation threshold.	Does hypotension, renal dysfunction, or ischemia modify the diuretic and vasodilator plan?
Antimicrobial stewardship	Initial coverage, cultures, narrowing criteria, duration, source control, adverse event surveillance.	When should broad therapy be narrowed or stopped after microbiology and clinical response?

Design requirements for future empirical studies:

The framework should be tested with tasks that force separation between diagnosis and treatment. A weak study would ask whether a physician accepts an AI plan. A stronger study would require the participant to choose the diagnosis, select intervention, prescribe dose and route, identify contraindications, order monitoring, and revise the plan after new information appears. The clinical case should contain at least one safety constraint, such as renal impairment, allergy, bleeding risk, drug interaction, or new laboratory deterioration. Without such constraints, prescription safety cannot be meaningfully evaluated.

A feasible experimental design would randomize participants into four conditions: no AI, AI recommendation only, static explanation, and longitudinal therapeutic explanation. In the static explanation arm, the system would provide a rationale for the recommended treatment. In the longitudinal arm, the system would additionally require safety exclusion, dose logic, monitoring, and revision triggers. The main comparison would not be subjective trust; it would be action quality. Secondary outcomes could include time, confidence, cognitive load, perceived usefulness, and unnecessary treatment.

The participant sample should be stratified by clinical experience. Therapeutic reliance differs between medical students, residents, attending physicians, pharmacists, and nurses. A recommendation that is useful to a junior clinician may be intrusive to an expert. A dose explanation that is redundant for a pharmacist may be essential for a physician under time pressure. Future studies should therefore measure interaction between user expertise and explanation granularity.

The cases should also include both correct and imperfect AI advice. If the AI is always right, the study measures acceptance of a helpful tool, not safety. Therapeutic systems need testing under partially incorrect conditions: correct indication with wrong dose, correct dose with missing monitoring, correct drug with overlooked contraindication, plausible treatment with wrong timing, and correct initial plan with absent revision after new data. These cases expose whether the explanation supports clinical judgment or merely makes automation smoother.

Table 8: Proposed experimental structure for validating longitudinal therapeutic XAI.

Element	Recommended specification
Study arms	No AI; AI recommendation only; static XAI; longitudinal therapeutic XAI.
Participant groups	Residents, attending physicians, pharmacists, and nurses when team execution is relevant.
Case domains	Sepsis, pulmonary embolism, diabetic ketoacidosis, acute heart failure, renal dosing, antimicrobial stewardship, sedation.
AI correctness scenarios	Correct plan; correct indication with wrong dose; missing contraindication; absent monitoring; obsolete plan after new data.
Primary endpoints	Treatment appropriateness, dose safety, contraindication detection, monitoring completeness, revision responsiveness.
Secondary endpoints	Time to plan, confidence, perceived usefulness, cognitive load, unnecessary treatment, quality of override explanation.
Data capture	Order choices, dose values, monitoring orders, plan changes, explanation views, response times, free-text rationales, override reasons.

Minimum reporting items for therapeutic XAI studies:

Studies of therapeutic XAI should report more than model type, explanation format, and overall accuracy. They should state the clinical task, the action under consideration, the user role, the risk level, the data available at the time of advice, the degree of AI correctness, the explanation content, and the final human action. Without these details, reviewers cannot determine whether the system improved therapy or merely altered confidence.

The report should also distinguish order recommendation from order execution. A physician may click agreement with an AI suggestion without placing the corresponding order. A physician may place the order but omit monitoring. A physician may modify the order after a pharmacist intervention. Each of these states has different safety meaning. Therapeutic XAI studies should therefore record the final order set, not only the displayed recommendation or stated intention.

Another reporting requirement is missing data. Therapeutic recommendations often depend on absent information: allergy history, creatinine trend, pregnancy status, weight, anticoagulation history, electrocardiogram, cultures, or current medications. A system that advises despite missing data should report what was missing and how the explanation represented that absence. Missing-data transparency is part of therapeutic safety.

Table 9: Minimum reporting items for therapeutic XAI studies.

Reporting item	Reason for inclusion
Clinical action target	Specifies whether the system influenced diagnosis, drug choice, dose, monitoring, escalation, or revision.
User role and expertise	Determines whether explanation granularity matches medical students, residents, attendings, pharmacists, or nurses.
Available data at advice time	Prevents retrospective inflation of system performance using information unavailable at the bedside.
AI correctness state	Separates correct advice, partially correct advice, unsafe advice, and incomplete advice.
Explanation content	Allows assessment of whether the explanation addressed action, safety, monitoring, and revision.
Final order behavior	Distinguishes stated acceptance from actual order execution and monitoring.
Override reason	Turns clinician disagreement into data for safety improvement.
Revision behavior	Shows whether the plan changed when new data emerged.

Governance and integration into clinical systems:

A therapeutic XAI system should not be deployed as an isolated chatbot beside the electronic health record. Prescription safety depends on structured data, medication lists, allergies, laboratory trends, renal function, culture results, clinical notes, and current orders. A system that lacks access to this context will generate advice that sounds clinically plausible while missing the variables that determine safe execution. Integration should therefore prioritize structured medication and laboratory pipelines before free-text conversational features.

Governance should also define when the system is allowed to advise, when it must remain silent, and when it must escalate. For example, an anticoagulation recommendation with missing platelet count or renal function should not present itself with the same confidence as a recommendation with complete data. A sepsis recommendation without allergy history should display a safety gap. An antibiotic recommendation after culture results should trigger de-escalation logic rather than repeat the initial broad plan. These design choices transform XAI from explanation into operational control.

The need for assurance infrastructure has been emphasized in health AI governance, including the proposal of a network of assurance laboratories for evaluation, localized testing, ongoing monitoring, and reporting [12]. This logic applies directly to therapeutic XAI. Local validation is necessary because prescribing practices, formularies, laboratory ranges, protocols, and available medications differ across institutions. A therapeutic XAI framework should include local configuration: antimicrobial formulary, renal dosing references, high-alert medications, local sepsis protocol, anticoagulation policy, pharmacist review thresholds, and escalation pathways.

Finally, audit should be continuous. The system should record recommendation, explanation layer, missing data, clinician amendment, final order, monitoring plan, and subsequent revision. These records can identify systematic over-acceptance, frequent override reasons, unsafe suggestions, and workflow points where clinicians ignore or modify advice. Audit also prevents the false comfort of dashboard accuracy. What matters is not only what the model predicted; what matters is what happened to the patient-facing order.

Discussion:

This review argues that prescription safety should become a central endpoint of explainable clinical AI. The available evidence suggests that AI advice influences therapeutic behavior [1], that clinicians respond to treatment advice through negotiation rather than simple acceptance [2], and that LLM agents require evaluation inside clinical workflows [3,5,6]. Taken together, these findings expose a gap in current XAI design. Many explanations are still optimized to justify predictions. Treatment decisions require explanations that structure action.

The difference matters because therapeutic decisions are layered. A clinician does not decide only whether antibiotics are appropriate. The clinician chooses a regimen, timing, cultures, renal adjustment, reassessment interval, de-escalation criteria, and adverse-event surveillance. A clinician does not decide only whether anticoagulation is appropriate. The clinician weighs bleeding risk, renal function, indication strength, hemodynamic status, interaction, monitoring, and reversal strategy. A single feature-importance explanation cannot carry this load.

A second implication concerns accountability. If clinicians negotiate AI recommendations, systems must record the negotiation. A modified recommendation is not necessarily non-adherence. It may be the safest expression of expert judgment. The interface should capture why the physician accepted the indication but changed the dose, why

treatment was delayed pending a laboratory result, or why the recommendation was rejected because of a contraindication. This record is clinically useful, educationally useful, and essential for post-event audit.

A third implication concerns large language models. Their fluency makes them attractive for therapeutic explanation, but fluency is not safety. Hager and colleagues demonstrate limitations that are directly relevant to action: current LLMs struggle with instruction-following, clinical information order, guideline adherence, and laboratory interpretation [4]. These limitations require guardrails. Therapeutic XAI should integrate retrieval of authoritative sources, structured patient data, dose calculators, interaction checkers, uncertainty statements, and human confirmation before any action is executed.

The proposed framework also reframes the role of explanation. In diagnostic XAI, explanation often asks whether the physician understands why a model suggested a label. In therapeutic XAI, explanation asks whether the physician has enough structured information to execute, monitor, and revise a plan. This is a higher standard. It is also closer to clinical responsibility.

Future empirical studies should compare at least four conditions: usual care without AI, AI recommendation without explanation, static XAI, and longitudinal therapeutic XAI. The outcome set should include action-level endpoints rather than diagnosis alone. High-fidelity simulated cases are appropriate before live deployment, especially in sepsis, pulmonary embolism, diabetic ketoacidosis, acute heart failure, renal failure dosing, antimicrobial stewardship, and sedation. Participants should include physicians at different levels of training, pharmacists, nurses, and specialists when the therapeutic decision depends on team execution.

A final concern is workflow burden. Therapeutic explanations can become unsafe if they are too long, too generic, or too disruptive. The answer is adaptive explanation. A low-risk recommendation may need a compact rationale. A high-risk prescription should require safety checks and documented clinician judgment. A junior clinician may need dose logic and monitoring prompts. A senior clinician may need a concise exception alert. A pharmacist may need interaction and adjustment details. Role-specific explanation is not cosmetic. It is a safety requirement.

Limitations:

This review is conceptual and narrative. It does not produce pooled effect estimates. The studies included differ in clinical setting, participant expertise, AI task, explanation type, and outcome measure. The proposed framework therefore requires empirical testing before being treated as an implementation standard.

Another limitation is that much of the available evidence comes from simulation, controlled user studies, or benchmark evaluations. These designs are appropriate for early safety work but do not fully reproduce fatigue, legal responsibility, institutional protocols, time pressure, team communication, or electronic health record friction. Prospective clinical deployment should follow simulation, silent-mode testing, and governance review.

The review also focuses on literature with DOI identifiers. This strengthens traceability but excludes relevant technical reports, institutional guidelines, and regulatory documents that lack DOI metadata. Such documents should inform implementation, but they were not used as cited evidence in this manuscript.

Conclusion:

Clinical AI should not stop at diagnostic suggestion. The clinical value of AI increases when reasoning becomes safe action: choosing, dosing, timing, monitoring, revising, and documenting treatment. The literature on AI-assisted prescription, clinician negotiation of treatment advice, LLM agents, reporting standards, and human-AI collaboration indicates that therapeutic XAI needs its own evaluation framework.

The Longitudinal Therapeutic XAI Framework proposed here defines explanation as a safety layer across the treatment cycle. It requires indication logic, contraindication screening, dose and route explanation, timing, monitoring, revision triggers, and accountability. This shifts XAI from the justification of a model output to the governance of clinical action. A model that explains why it recommends a diagnosis has not yet explained how to care for the patient. Therapeutic XAI must close that gap.

Acknowledgments:

None.

Ethical approval:

Not applicable. This review did not involve human participants, animal subjects, private clinical records, or patient-level data.

Conflict of interest:

The author declares no competing interests.

Funding:

No external funding was received for this work.

Data availability:

No dataset was generated or analyzed. All cited material is publicly available through the referenced publications.

Use of artificial intelligence tools:

Artificial intelligence tools were used for language drafting assistance and formatting support. The author reviewed, revised, and assumes full responsibility for the scientific content, argument, references, and final manuscript.

References:

1. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. NPJ Digital Medicine. 2023;6:206. doi:10.1038/s41746-023-00955-z.

2. Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023. doi:10.1145/3544548.3581075.

3. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digital Medicine. 2024;7:84. doi:10.1038/s41746-024-01083-y.

4. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine. 2024;30:2613-2622. doi:10.1038/s41591-024-03097-1.

5. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models and agents in healthcare: key challenges in clinical applications. Intelligent Medicine. 2025;5(2):151-163. doi:10.1016/j.imed.2025.03.002.

6. Schmidgall S, Ziaei R, Harris C, et al. AgentClinic: a multimodal benchmark for tool-using clinical AI agents. NPJ Digital Medicine. 2026. doi:10.1038/s41746-026-02674-7.

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

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

9. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-180. doi:10.1038/s41586-023-06291-2.

10. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nature Medicine. 2025;31:943-950. doi:10.1038/s41591-024-03423-7.

11. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025;642:442-450. doi:10.1038/s41586-025-08866-7.

12. Shah NH, Halamka JD, Saria S, et al. A nationwide network of health AI assurance laboratories. JAMA. 2024;331(3):245-249. doi:10.1001/jama.2023.26930.

13. Cabitza F, Campagner A, Ronzio L, et al. Rams, hounds and white boxes: investigating human-AI collaboration protocols in medical diagnosis. Artificial Intelligence in Medicine. 2023;138:102506. doi:10.1016/j.artmed.2023.102506.

14. Cabitza F, Natali C, Famiglini L, Campagner A, Caccavella V, Gallazzi E. Never tell me the odds: investigating pro-hoc explanations in medical decision making. Artificial Intelligence in Medicine. 2024;150:102819. doi:10.1016/j.artmed.2024.102819.

15. Subramanian HV, Canfield C, Shank DB. Designing explainable AI to improve human-AI team performance: a medical stakeholder-driven scoping review. Artificial Intelligence in Medicine. 2024;149:102780. doi:10.1016/j.artmed.2024.102780.

16. Panigutti C, Beretta A, Giannotti F, Pedreschi D. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical decision support systems. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022. doi:10.1145/3491102.3502104.

17. Bussone A, Stumpf S, O'Sullivan D. The role of explanations on trust and reliance in clinical decision support systems. 2015 IEEE International Conference on Healthcare Informatics. 2015:160-169. doi:10.1109/ICHI.2015.26.

18. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. International Journal of Human-Computer Studies. 2023;169:102941. doi:10.1016/j.ijhcs.2022.102941.

19. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;3(11):e745-e750. doi:10.1016/S2589-7500(21)00208-9.

20. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digital Medicine. 2020;3:17. doi:10.1038/s41746-020-0221-y.

21. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews. Health Services Research and Managerial Epidemiology. 2024;11. doi:10.1177/23333928241234863.

22. Goodman KE, Yi PH, Morgan DJ. AI-generated clinical summaries require more than accuracy. JAMA. 2024;331(8):637-638. doi:10.1001/jama.2024.0555.

23. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digital Medicine. 2021;4:31. doi:10.1038/s41746-021-00385-9.