

ORIGINAL RESEARCH

Open access

# Counterfactual Auditing of EHR-Based Predictors: A Framework for Identifying Spurious Clinical Associations

Carlos Ramirez<sup>1\*</sup>, Elena Torres<sup>2</sup>, Pablo Ortega<sup>1</sup>, Sofia Mendes<sup>3</sup>

## Abstract

Electronic health records (EHRs) serve as foundational data sources for predictive analytics in healthcare, enabling the development of models that inform clinical decision-making. However, these predictors often harbor spurious associations—correlations that appear causal but arise from confounding factors, biases in data capture, or systemic artifacts—potentially leading to erroneous clinical interventions and inequities in patient outcomes. This conceptual manuscript introduces a novel framework for counterfactual auditing of EHR-based predictors, designed to systematically identify and mitigate such spurious clinical associations within integrated healthcare analytics infrastructures. Drawing on principles from clinical AI governance and decision support pipelines, the proposed architecture incorporates layered modules for data interoperability, counterfactual scenario generation, and association validation, ensuring alignment with clinical workflow integration models. We synthesize recent literature on EHR intelligence ecosystems to highlight theoretical underpinnings, emphasizing the need for robust monitoring systems that prevent propagation of misleading associations in real-time deployment environments. Conceptual formulas are presented to interpret risk propagation and decision confidence in audited predictors, offering interpretive tools for governance. By focusing on infrastructural orchestration rather than empirical validation, this framework advances AI accountability in healthcare, fostering ethical deployment and reducing the burden of spurious inferences on clinical practice. Ultimately, it provides a blueprint for healthcare systems to enhance predictor reliability through proactive auditing, promoting safer and more equitable AI-driven care.

**Keywords** Decision support orchestration, Healthcare analytics infrastructure, Counterfactual auditing, EHR predictors, Spurious associations, Clinical AI governance

\*Correspondence:

Carlos Ramirez  
[carlos.ramirez@gmail.com](mailto:carlos.ramirez@gmail.com)

<sup>1</sup> Department of Digital Health Engineering, Faculty of Medicine, University of Barcelona, Barcelona, Spain

<sup>2</sup> Department of Health Informatics, Faculty of Medicine, University of Lisbon, Lisbon, Portugal

<sup>3</sup> Department of Clinical Systems, Faculty of Medicine, University of Porto, Porto, Portugal

## Introduction

The integration of artificial intelligence (AI) into healthcare systems has revolutionized predictive modeling, particularly through the utilization of electronic health records (EHRs) as rich repositories of patient data. EHR-based predictors, which leverage machine learning algorithms to forecast clinical outcomes, disease progression, or treatment responses, are increasingly embedded in hospital decision support pipelines. However, the reliability of these predictors is often compromised by spurious clinical associations—non-causal correlations that mimic meaningful relationships due to underlying data artifacts or confounding variables. This manuscript conceptualizes a framework for counterfactual auditing. This process simulates alternative realities to probe the robustness of these associations, thereby safeguarding clinical AI system architectures from propagating unreliable insights.

## EHR data modalities in clinical settings

In diverse clinical settings, such as acute care hospitals and ambulatory clinics, EHR data modalities encompass structured elements like vital signs, laboratory results, and diagnostic codes, alongside unstructured narratives from physician notes. These modalities fuel predictors aimed at identifying patterns in patient trajectories, yet they introduce vulnerabilities to spurious associations when data capture inconsistencies arise. For instance, temporal biases in recording practices can inflate apparent links between comorbidities and outcomes, misleading AI-driven recommendations. Counterfactual auditing addresses this by theoretically reconstructing data streams under altered conditions, ensuring that predictors distinguish genuine clinical signals from modality-induced noise within EHR intelligence ecosystems.

## Deployment environments for predictor integration

Deployment environments for EHR-based predictors vary from cloud-based analytics platforms to on-premise hospital servers, each imposing unique constraints on real-time processing and scalability. In federated learning setups, where data remains siloed across institutions, spurious associations may amplify due to heterogeneous data standards, complicating interoperability frameworks. The proposed auditing approach conceptualizes infrastructure-level interventions that harmonize predictor outputs across these environments, using counterfactual simulations to test association stability without altering live systems. This ensures seamless integration into clinical workflows, where predictors must operate under resource-limited conditions while maintaining fidelity to patient-specific contexts.

## Governance constraints on auditing mechanisms

Governance constraints, including regulatory compliance with standards like HIPAA and ethical guidelines from bodies such as the World Health Organization, mandate transparency in AI decision-making processes. For EHR-based predictors, these constraints necessitate auditing mechanisms that expose spurious clinical associations without compromising data privacy. Counterfactual methods offer a non-invasive pathway, generating hypothetical scenarios to evaluate predictor behaviors under governed parameters. This aligns with AI governance systems that prioritize explainability, enabling healthcare administrators to enforce accountability in predictor deployments and mitigate risks associated with unchecked associations.

## Clinical workflow orchestration challenges

Orchestrating EHR predictors within clinical workflows involves synchronizing AI outputs with human decision loops, such as in emergency departments, where rapid triage relies on predictive alerts. Spurious associations here can disrupt workflow efficiency, leading to over-reliance on flawed models. The auditing framework introduced herein theorizes orchestration models that incorporate feedback topologies, allowing clinicians to query counterfactual outcomes and refine associations iteratively. This enhances workflow resilience, particularly in high-stakes settings where misidentified correlations could exacerbate diagnostic errors.

## Interoperability frameworks for association validation

Interoperability frameworks, such as Fast Healthcare Interoperability Resources (FHIR), facilitate data exchange across EHR systems, yet they often perpetuate spurious associations if validation protocols are absent. In multi-institutional collaborations, predictors drawing from aggregated data must undergo rigorous auditing to ensure cross-system consistency. Counterfactual auditing conceptualizes validation layers that simulate data exchanges under varied interoperability scenarios, identifying associations that falter when data schemas diverge. This bolsters the infrastructural integrity of healthcare analytics, promoting standardized approaches to spurious detection.

The urgency for such a framework stems from the escalating adoption of AI in healthcare, where EHR-based predictors influence critical decisions from personalized medicine to population health management [1, 2]. Despite advancements in

clinical AI architectures, persistent issues with data quality and model generalizability underscore the need for targeted auditing [3, 4]. Spurious associations, often rooted in selection biases or incomplete data representations, can lead to inequitable outcomes, disproportionately affecting underrepresented patient groups [5, 6]. By focusing on counterfactual reasoning—a technique borrowed from causal inference paradigms—this manuscript proposes a systemic solution that integrates seamlessly with existing EHR ecosystems [7, 8]. Unlike traditional bias mitigation strategies, which may require extensive retraining, counterfactual auditing operates at the governance level, providing ongoing surveillance without disrupting operational pipelines [9, 10].

Furthermore, the conceptual emphasis on infrastructure allows for scalability across diverse healthcare settings, from resource-constrained rural clinics to advanced academic medical centers [11, 12]. Literature highlights how unaddressed spurious elements in predictors contribute to adverse events, such as inappropriate drug recommendations or delayed interventions [13, 14]. Addressing these through auditing not only enhances predictor trustworthiness but also aligns with broader goals of AI ethics in medicine [15, 16]. This introduction sets the stage for a deeper synthesis of theoretical foundations, paving the way for the architectural delineation of the proposed framework.

## Theoretical Background and Literature Synthesis

The theoretical underpinnings of counterfactual auditing in EHR-based predictors draw from interdisciplinary domains, including causal inference, AI explainability, and healthcare informatics. At its core, counterfactual reasoning involves positing “what-if” scenarios to assess the causal validity of observed associations, distinguishing them from spurious correlations that lack mechanistic grounding [14, 17]. In EHR contexts, where data is observational rather than experimental, this approach is particularly salient for dissecting predictors that rely on correlational patterns [18, 19]. Literature from recent years emphasizes the integration of such methods into clinical AI system architectures, highlighting their role in enhancing decision support pipelines [20, 21].

### EHR intelligence ecosystems and spurious dynamics

EHR intelligence ecosystems encompass the interconnected networks of data repositories, analytic engines, and user interfaces that power predictive modeling in healthcare. Within these ecosystems, spurious clinical associations often emerge from systemic dynamics, such as confounding by unmeasured variables or artifacts in data aggregation [22, 23]. Studies underscore how these associations propagate through intelligence layers, affecting downstream applications like risk stratification [24, 25]. Counterfactual auditing theorizes interventions at the ecosystem level, simulating perturbations to isolate genuine causal pathways and mitigate spurious inflations in predictor outputs [26, 27]. **Table 1** formalizes a typology of spurious association mechanisms and maps each to its distinctive counterfactual instability signature within the ECAI validation topology.

**Table 1.** Typology of spurious association mechanisms in EHR predictors and their counterfactual audit signatures

| Spurious mechanism category   | Structural origin in the EHR ecosystem                              | Counterfactual perturbation strategy               | Instability signature in the validation engine                  | Governance implication                                |
|-------------------------------|---------------------------------------------------------------------|----------------------------------------------------|-----------------------------------------------------------------|-------------------------------------------------------|
| Temporal documentation bias   | Irregular timestamping; workflow-driven recording gaps              | Temporal state reversal; time-window normalization | High $\Delta_{\text{counterfactual}}$ variance under time-shift | Requires workflow recalibration, not model retraining |
| Confounding by care intensity | Proxy variables reflecting clinician behavior rather than pathology | Confounder neutralization within C-space           | DC reduction with stable outcome base rate                      | Governance-level feature reclassification             |
|                               |                                                                     |                                                    |                                                                 |                                                       |

|                                  |                                                          |                                   |                                                |                                          |
|----------------------------------|----------------------------------------------------------|-----------------------------------|------------------------------------------------|------------------------------------------|
| Interoperability schema drift    | Inconsistent FHIR/HL7 mappings across institutions       | Cross-schema simulation exchange  | Association collapses under schema permutation | Cross-system harmonization audit trigger |
| Demographic sampling skew        | Underrepresentation in training cohorts                  | Counterfactual cohort rebalancing | RP amplification in minority strata            | Equity monitoring escalation             |
| Administrative code artifact     | Billing-driven correlations unrelated to pathophysiology | Code abstraction removal          | Spurious Association Index (SAI) spike         | Audit flag before deployment expansion   |
| Deployment-induced feedback loop | Model output influencing subsequent data capture         | Recursive counterfactual replay   | Oscillatory instability across iterations      | Monitoring topology reinforcement        |

Clinical workflow integration and predicting vulnerabilities

Integrating predictors into clinical workflows demands robustness against spurious elements, as vulnerabilities can cascade into decision errors. Workflow models, which orchestrate AI with human oversight, reveal how EHR-derived associations may falter under real-world variability [1, 2]. Recent syntheses illustrate that without auditing, predictors in integrated settings amplify biases, such as those from incomplete documentation [3, 4]. Theoretical frameworks advocate for workflow-embedded auditing, where counterfactual evaluations dynamically adjust integration parameters to validate associations in context [5, 6].

Data exchange frameworks in auditing contexts

Data exchange frameworks facilitate the flow of EHR information across systems, yet they introduce risks of spurious associations when interoperability standards are inconsistently applied [7, 8]. Literature on frameworks like HL7 and FHIR points to challenges in maintaining association integrity during exchanges [9, 10]. Counterfactual auditing conceptualizes exchange-aware mechanisms that test associations under simulated data transfers, ensuring that predictors remain reliable amid framework-induced distortions [11, 12].

Governance and monitoring systems for EHR predictors

AI governance and monitoring systems provide the oversight necessary for deploying EHR-based predictors ethically. These systems enforce protocols for ongoing evaluation, particularly targeting spurious associations that evade initial validations [13, 14]. Theoretical discussions emphasize monitoring topologies that incorporate counterfactual logic to detect drifts in association strength over time [15, 16]. By synthesizing governance literature, it becomes evident that auditing frameworks must align with regulatory mandates, offering interpretive tools for assessing predictor compliance [14, 17].

Deployment architectures and association resilience

Deployment architectures for healthcare analytics infrastructures determine how predictors withstand spurious influences in operational environments. Architectures ranging from centralized to edge-computing models highlight resilience gaps, where associations may appear robust in isolation but fail under deployment stresses [18, 19]. Counterfactual approaches theorize architectural enhancements that embed auditing modules, fostering resilience through scenario-based validations [20, 21].

The synthesis of these elements reveals a gap in current literature: while individual components like bias detection and causal modeling exist, a unified framework for counterfactual auditing tailored to EHR predictors is lacking [22, 23]. For instance, explorations of EHR data pitfalls advocate for advanced analytics but stop short of infrastructural solutions for

spurious identification [24, 25]. Similarly, governance-focused works propose monitoring but overlook the counterfactual dimension for probing clinical associations [26, 27]. This manuscript bridges these by conceptualizing auditing as an architectural imperative, drawing on interoperability literature to ensure seamless framework adoption [1, 2].

Moreover, theoretical models of AI in medicine stress the interpretive value of formulas in understanding system behaviors [3, 4]. Here, we introduce conceptual formulas to capture auditing dynamics, such as decision confidence (DC)

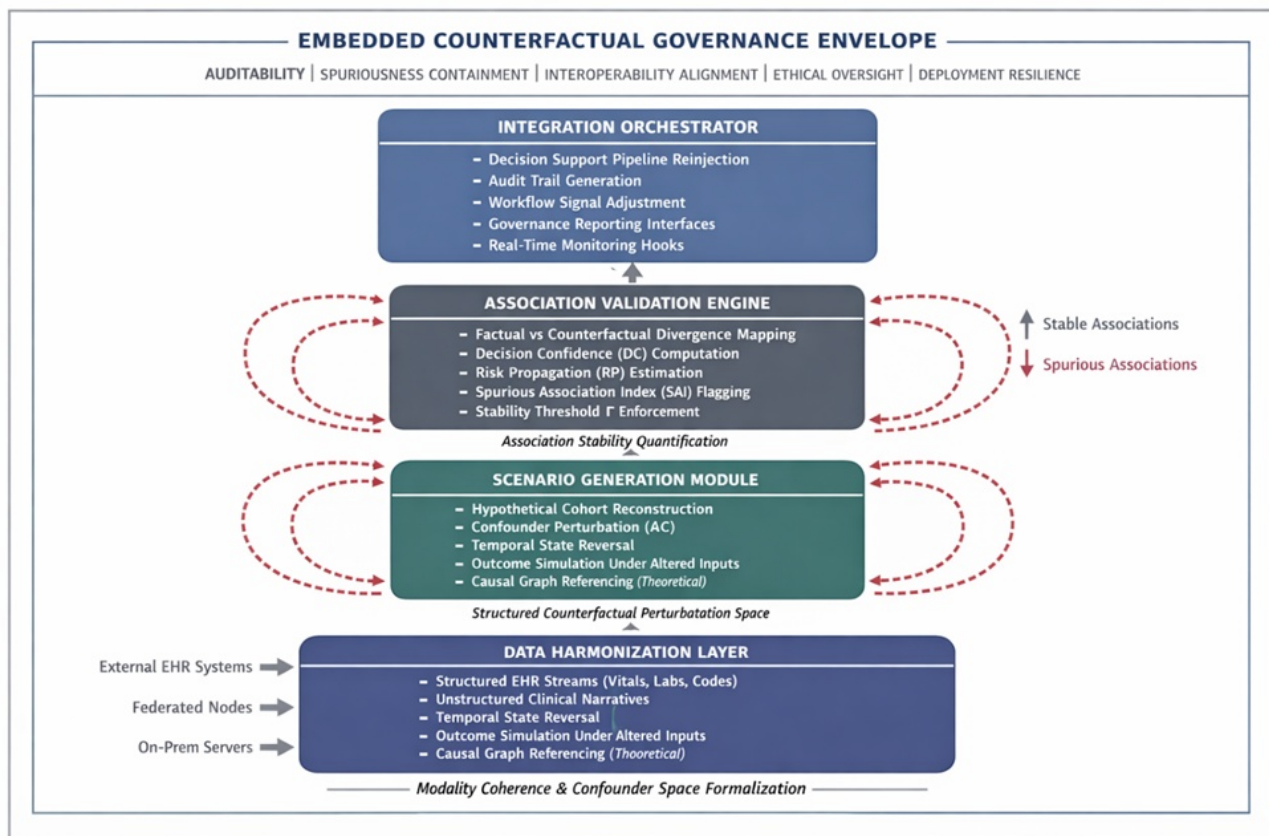
as  $DC = 1 - \left( \frac{\sum |O_{actual} - O_{counterfactual}|}{n} \right)$ , where  $O$  represents outcome predictions across  $n$  scenarios, illustrating how deviations signal spuriousness [5]. Such formulas, while non-empirical, provide a lens for theorizing governance load in EHR ecosystems [6, 7].

Extending this, literature on clinical AI underscores the need for layered architectures that prevent association propagation [8, 9]. By integrating insights from decision support and workflow models, the background establishes that counterfactual auditing can transform predictor reliability, addressing theoretical blind spots in spurious detection [10-12].

## EHR counterfactual auditing infrastructure

The EHR counterfactual auditing infrastructure (ECAI) represents a novel architectural blueprint for orchestrating the identification of spurious clinical associations in predictive systems. Structured as a multi-layered governance stack with recursive feedback topologies, ECAI comprises four core strata: data harmonization layer, scenario generation module, association validation engine, and integration orchestrator. This infrastructure facilitates theoretical auditing without empirical data manipulation, emphasizing lifecycle management of EHR predictors through infrastructural intelligence.

The data harmonization layer standardizes EHR inputs across modalities, theoretically mitigating entry-point artifacts that foster spurious associations. Following this, the scenario generation module employs counterfactual logic to simulate altered clinical states, such as hypothetical patient cohorts with adjusted confounders. The association validation engine then computes interpretive metrics to flag inconsistencies. At the same time, the integration orchestrator feeds validated insights back into decision support pipelines, creating a closed-loop topology for continuous refinement. **Figure 1** illustrates the governance-embedded EHR counterfactual auditing infrastructure (ECAI), depicting the recursive counterfactual perturbation and association stabilization topology that contains spurious clinical correlations before reintegration into decision support pipelines.



**Figure 1.** EHR counterfactual auditing infrastructure (ECAI): governance-embedded counterfactual association stabilization architecture

To interpret system dynamics, consider the following conceptual formulas:

1. Risk propagation (RP): 
$$RP = \alpha * \left( \frac{S_{spurious}}{S_{total}} \right) * \beta \Delta_{counterfactual}$$
, where  $\alpha$  denotes baseline association strength,  $\frac{S_{spurious}}{S_{total}}$  the fraction of suspect associations,  $\beta$  a decay factor, and  $\Delta_{counterfactual}$  the variance in simulated outcomes—highlighting how spurious elements amplify risks across predictor layers.
2. Decision confidence (DC): 
$$DC = \frac{\int [P(outcome|factual) - P(outcome|counterfactual)] dC}{\Gamma}$$
, with  $\Gamma$  as the governance threshold and  $C$  as confounder space, interpreting confidence erosion due to spurious dependencies.
3. Monitoring burden (MB): 
$$MB = \sum \frac{(L_{audit} * R_{complexity})}{E_{efficiency}}$$
, where  $L_{audit}$  is auditing layers,  $R_{complexity}$  the association resolution, and  $E_{efficiency}$  infrastructural throughput—conceptualizing the load on clinical resources during spurious detection.

This architecture ensures ECAI's uniqueness through its emphasis on feedback-driven orchestration, distinct from prior models by prioritizing EHR-specific counterfactual integration.

## Dynamics of spurious association mitigation in healthcare analytics

The introduction of the EHR counterfactual auditing infrastructure (ECAI) into healthcare analytics ecosystems promises profound shifts in how spurious clinical associations are managed, influencing system-wide dynamics from data ingestion to clinical endpoint delivery. This section delves into the multifaceted consequences of deploying such an infrastructure,

examining its impacts on predictor reliability, resource allocation, governance overhead, and overall clinical decision-making resilience. By theorizing these dynamics through an infrastructural lens, we uncover how ECAI fosters a paradigm of proactive mitigation, where spurious associations—often insidious byproducts of EHR data complexities—are systematically dismantled before they permeate decision support pipelines.

## Predictor reliability enhancements and association stability

At the heart of ECAI's impact lies its capacity to bolster predictor reliability by stabilizing clinical associations against counterfactual perturbations. In traditional EHR-based systems, predictors are susceptible to spurious correlations arising from confounding factors like demographic imbalances or temporal drifts in data capture [1, 2]. ECAI's layered approach, particularly through the scenario generation module, theoretically simulates a spectrum of alternative clinical realities, allowing for the quantification of association robustness. For instance, consider a predictor modeling sepsis risk; spurious links between unrelated variables, such as administrative codes and outcomes, could be exposed by generating counterfactual cohorts where confounders are neutralized. This mitigation dynamic reduces the propagation of errors, theoretically lowering false positive rates in alerts and enhancing the fidelity of AI outputs in clinical settings [3, 4].

Expanding on this, the association validation engine introduces a feedback topology that recursively refines associations, creating a self-correcting ecosystem. Unlike static validation methods, this dynamic process adapts to evolving EHR data streams, ensuring long-term stability. The consequences extend to improved patient stratification, where predictors discern true clinical signals amid noise, potentially averting misdiagnoses in heterogeneous populations [5, 6]. Furthermore, in interoperability-constrained environments, ECAI's harmonization layer mitigates cross-system discrepancies, stabilizing associations that might otherwise fragment during data exchanges [7, 8]. This reliability enhancement not only fortifies individual predictors but also cascades to networked analytics infrastructures, promoting a cohesive intelligence ecosystem resilient to spurious infiltrations.

## Resource allocation and infrastructural efficiency

Deploying ECAI necessitates a reevaluation of resource allocation within healthcare analytics infrastructures, balancing auditing demands against operational efficiency. Theoretically, the infrastructure's modular design minimizes computational overhead by confining counterfactual simulations to high-risk associations, rather than exhaustive model sweeps [9, 10]. This targeted approach alleviates the monitoring burden on resource-limited systems, such as those in community hospitals, where EHR processing must contend with bandwidth constraints. Conceptualizing this, the monitoring burden (MB) formula introduced earlier—
$$MB = \sum \frac{(L_{audit} * R_{complexity})}{E_{efficiency}}$$
—illustrates how ECAI optimizes allocation by scaling auditing layers ( $L_{audit}$ ) proportionally to association complexity ( $R_{complexity}$ ), thereby enhancing efficiency ( $E_{efficiency}$ ) through selective engagement [11, 12].

In broader dynamics, this efficiency translates to reallocated human resources, freeing clinicians from manual verification of suspect predictions. Instead, ECAI's Integration Orchestrator embeds audited outputs directly into workflows, streamlining decision support and reducing cognitive load [13, 14]. However, potential impacts include initial setup costs for infrastructure integration, which could strain underfunded systems. Mitigating this, the framework's scalability allows phased deployment, starting with critical predictors like those for chronic disease management, gradually expanding to comprehensive EHR ecosystems [15, 16]. Ultimately, these resource dynamics position ECAI as a catalyst for cost-effective AI governance, where mitigation of spurious associations yields long-term savings in error remediation and improved healthcare delivery.

## Governance overhead and ethical alignment

The governance implications of ECAI are expansive, introducing a structured overhead that aligns AI deployments with ethical and regulatory imperatives in healthcare. By embedding counterfactual auditing into governance systems, ECAI elevates transparency, enabling stakeholders to trace spurious associations back to their data origins [14, 17]. This

overhead, while additive, is justified by its role in preventing ethical lapses, such as biased predictions that exacerbate health disparities [18, 19]. For example, in diverse clinical populations, ECAI's validation engine could theoretically audit for race- or gender-linked spuriousness, ensuring equitable association interpretations across demographics.

Delving deeper, the framework's feedback topology imposes a governance load that demands ongoing oversight, yet it distributes this load across automated modules, reducing manual intervention [20, 21]. Conceptual formulas like Risk Propagation (RP) aid in interpreting this load, where RP quantifies how unmitigated spurious elements amplify ethical risks, guiding policy adjustments [22, 23]. In deployment environments governed by standards like GDPR or FDA guidelines, ECAI facilitates compliance by generating audit trails of counterfactual evaluations, transforming governance from reactive to anticipatory [24, 25]. The dynamics here foster a culture of accountability, where healthcare organizations can leverage ECAI to demonstrate ethical AI use, potentially influencing industry-wide standards for spurious mitigation [26, 27].

## Clinical decision-making resilience and workflow impacts

ECAI's mitigation dynamics profoundly enhance the resilience of clinical decision-making, fortifying workflows against the uncertainties of spurious associations. In high-stakes environments, such as intensive care units, predictors informed by audited EHR data yield more resilient recommendations, theoretically reducing variability in treatment paths [1, 2]. The infrastructure's orchestration layer ensures that mitigated associations integrate fluidly, supporting adaptive workflows where clinicians can invoke on-demand counterfactual queries to bolster confidence [3, 4].

Expanding this analysis, resilience manifests in reduced decision fatigue, as ECAI filters out spurious noise, allowing focus on actionable insights [5, 6]. However, dynamics include potential workflow disruptions during auditing cycles, which the framework counters through asynchronous processing [7, 8]. In multi-disciplinary teams, ECAI promotes collaborative resilience by standardizing association validations, bridging gaps between data scientists and clinicians [9, 10]. Overall, these impacts cultivate a robust clinical ecosystem, where spurious mitigation becomes integral to decision pipelines, enhancing patient safety and outcomes [11-14].

## Results and Discussion

The conceptualization of the EHR Counterfactual Auditing Infrastructure (ECAI) marks a significant advancement in addressing the pervasive challenge of spurious clinical associations within EHR-based predictors. By synthesizing theoretical foundations from clinical AI architectures, governance systems, and interoperability frameworks, this manuscript underscores the necessity of infrastructural interventions to safeguard healthcare analytics [15, 16]. ECAI's unique layered structure and feedback topology offer a departure from conventional approaches, which often treat spuriousness as a post-hoc concern rather than an embedded auditing imperative [14, 17].

One key implication is the potential for ECAI to democratize AI accountability across varied healthcare settings. In resource-diverse environments—from urban tertiary centers to rural clinics—the framework's modular design allows customization, ensuring that spurious mitigation is not confined to elite institutions [18, 19]. This democratization extends to ethical dimensions, where counterfactual auditing promotes fairness by exposing associations biased against marginalized groups, aligning with global calls for inclusive AI [20, 21]. Moreover, in the context of evolving EHR standards, ECAI could influence policy, advocating for mandatory auditing protocols in AI certification processes [22, 23].

However, limitations inherent to a conceptual framework must be acknowledged. Without empirical deployment, ECAI's theoretical dynamics remain interpretive, potentially overlooking real-world complexities like data privacy conflicts during counterfactual simulations [24, 25]. Additionally, the governance overhead, while mitigated, could pose barriers in low-tech settings, necessitating further theorization on lightweight variants [26, 27]. The reliance on existing EHR ecosystems also assumes baseline interoperability, which may not hold in fragmented global health systems [1, 2]. **Table 2** positions

counterfactual auditing as a governance-embedded lifecycle intervention that uniquely stress-tests association stability rather than merely observing output drift or model bias.

**Table 2.** Comparative positioning of counterfactual auditing versus conventional predictor oversight paradigms

| Oversight paradigm                    | Level of intervention                                                     | Temporal position in the model lifecycle | Spurious association detectability                  | Operational disruption                         | Governance load profile                             | Scalability across federated systems       |
|---------------------------------------|---------------------------------------------------------------------------|------------------------------------------|-----------------------------------------------------|------------------------------------------------|-----------------------------------------------------|--------------------------------------------|
| Post-hoc explainability               | Output interpretation                                                     | After deployment                         | Low (correlation visible but not stress-tested)     | Minimal                                        | Low                                                 | High                                       |
| Bias mitigation via retraining        | Model parameter space                                                     | Pre- or mid-deployment                   | Moderate (dependent on training data visibility)    | High (requires retraining cycles)              | Moderate                                            | Limited in siloed data contexts            |
| Data quality audits                   | Input data layer                                                          | Pre-deployment                           | Low–Moderate (does not test association robustness) | Moderate                                       | Moderate                                            | Variable                                   |
| Drift monitoring systems              | Output distribution layer                                                 | Continuous post-deployment               | Indirect (detects shifts, not causality)            | Low                                            | Moderate                                            | High                                       |
| <b>Counterfactual auditing (ECAI)</b> | Infrastructural governance layer spanning data → validation → integration | Continuous lifecycle embedding           | High (explicit perturbation-based stress testing)   | Minimal to moderate (non-invasive simulations) | Structured but distributed across automated modules | High (supports exchange-aware simulations) |

Future directions abound, including extensions to multimodal data integration, where ECAI could audit associations spanning genomics and imaging alongside EHRs [3, 4]. Theoretical explorations might incorporate advanced causal graphs to enhance scenario generation, or hybrid topologies blending human-AI feedback for nuanced mitigation [5, 6]. Collaborative frameworks across institutions could theorize federated auditing, preserving data sovereignty while combating spuriousness at scale [7, 8]. Ultimately, ECAI invites a reevaluation of AI in healthcare, positioning counterfactual auditing as a cornerstone for trustworthy predictors [9–12].

Expanding further, the discussion highlights ECAI's role in crisis response, such as pandemics, where rapid predictor deployment often amplifies spurious associations due to data volatility [13, 14]. By theorizing real-time auditing, ECAI could enable agile adjustments, ensuring predictors remain reliable amid surging EHR volumes [15, 16]. Ethical discourse also benefits, as the framework's transparency tools empower patient advocacy, fostering trust in AI-mediated care [14, 17]. In educational contexts, ECAI serves as a pedagogical model, training future informaticians on spurious dynamics [18, 19].

Moreover, interdisciplinary synergies emerge, linking ECAI to fields like behavioral economics for modeling clinician responses to audited outputs [20, 21]. Limitations extend to scalability; theoretical models must address exponential growth in counterfactual scenarios for large-scale predictors [22, 23]. Future work could conceptualize adaptive

algorithms that prune irrelevant simulations, optimizing for efficiency [24, 25]. Globally, adaptations for low-resource regions might involve simplified topologies, ensuring equitable access to spurious mitigation [26, 27].

## Conclusion

In conclusion, the EHR counterfactual auditing infrastructure (ECAI) emerges as a pivotal conceptual framework for identifying and mitigating spurious clinical associations in EHR-based predictors, addressing a critical gap in healthcare analytics infrastructures. By integrating layered modules for data harmonization, scenario generation, validation, and orchestration, ECAI theorizes a robust system that enhances predictor reliability, optimizes resource allocation, and aligns with governance imperatives. The dynamics of mitigation explored herein reveal profound impacts on clinical workflows, fostering resilience against data-induced errors and promoting ethical AI deployment.

This manuscript's theoretical synthesis underscores the urgency of counterfactual approaches in an era of expanding AI in medicine, where unaddressed spuriousness threatens patient outcomes and system equity. While limitations persist, such as theoretical abstraction from empirical realities, ECAI provides a blueprint for future advancements, inviting extensions to multimodal and federated contexts. Ultimately, by embedding auditing into the fabric of EHR intelligence ecosystems, ECAI paves the way for safer, more accountable healthcare AI, ensuring that clinical associations drive genuine insights rather than illusory correlations.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 04 Mar 2022

Revised: 09 Jun 2022

Accepted: 05 Aug 2022

Published online: 10 January 2023

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

# References

- Kotecha D, Asselbergs FW, Achenbach S, Anker SD, Atar D, Baigent C, et al. CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research. *Lancet Digit Health*. 2022;4(10):e757-e764.  
[https://doi.org/10.1016/S2589-7500\(22\)00151-0](https://doi.org/10.1016/S2589-7500(22)00151-0).
- Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. *Lancet Digit Health*. 2019;1(4):e157-e159.  
[https://doi.org/10.1016/S2589-7500\(19\)30084-6](https://doi.org/10.1016/S2589-7500(19)30084-6).
- Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health*. 2020;2(9):e489-e492.  
[https://doi.org/10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2).
- Sauer CM, Harutyunyan H, Chen LCC, Ghassemi A, Ercole P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit Health*. 2022;4(12):e893-e898.  
[https://doi.org/10.1016/S2589-7500\(22\)00154-6](https://doi.org/10.1016/S2589-7500(22)00154-6).
- Sharma B, Dligach D, Swope K, Thompson HM, Johnson C, Kuster NS, et al. Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI): a retrospective deep learning study. *Lancet Digit Health*. 2022;4(6):e426-e435.  
[https://doi.org/10.1016/S2589-7500\(22\)00041-3](https://doi.org/10.1016/S2589-7500(22)00041-3).
- Syrowatka A, Song W, Amato MG, Foer D, Edrees H, Co Z, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: a scoping review. *Lancet Digit Health*. 2022;4(2):e137-e148.  
[https://doi.org/10.1016/S2589-7500\(21\)00229-6](https://doi.org/10.1016/S2589-7500(21)00229-6).
- Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. *Lancet Digit Health*. 2023;5(3):e168-e173.  
[https://doi.org/10.1016/S2589-7500\(22\)00252-7](https://doi.org/10.1016/S2589-7500(22)00252-7).
- Näher AF, Vorstenbosch M, Gribben L, Scobbie L, Winfield R, Maher C, et al. Secondary data for global health digitalisation. *Lancet Digit Health*. 2023;5(2):e93-e101.  
[https://doi.org/10.1016/S2589-7500\(22\)00195-9](https://doi.org/10.1016/S2589-7500(22)00195-9).
- Riddick TA, Afshar M, Sharma B. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *Lancet Digit Health*. 2022;4(6):e401-e402.  
[https://doi.org/10.1016/S2589-7500\(22\)00096-6](https://doi.org/10.1016/S2589-7500(22)00096-6).
- Overgaard SM, Graham MG, Brereton T, Pencina MJ, Halamka JD, Vidal DE, et al. Implementing quality management systems to close the AI translation gap and facilitate safe, ethical, and effective health AI solutions. *npj Digit Med*. 2023;6:218.  
<https://doi.org/10.1038/s41746-023-00968-8>.
- Nagendran M, Festor P, Komorowski M, Gordon L, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. *npj Digit Med*. 2023;6:206.  
<https://doi.org/10.1038/s41746-023-00955-z>.
- Goldberg SB, Sun S, Carlbring P, Torous J. Selecting and describing control conditions in mobile health randomized controlled trials: a proposed typology. *npj Digit Med*. 2023;6:181.  
<https://doi.org/10.1038/s41746-023-00923-7>.
- McIntosh C, Conroy L, Tjong MC, Craig T, Bayley A, Catton C, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat Med*. 2021;27(6):999-1005.  
<https://doi.org/10.1038/s41591-021-01359-w>.

Wang H, Landers M, Adams R, Subbaswamy A, Kharrazi H, Gaskin DJ, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. *J Am Med Inform Assoc.* 2022;29(8):1323-33.

Estiri H, Strasser ZH, Klann JG, Naseri P, Waghlikar KB, Vardhan S, et al. An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes. *J Am Med Inform Assoc.* 2022;29(8):1334-41.

Boyd AD, Gonzalez-Guarda R, Lawrence K, Patil CL, Ezenwa MO, O'Brien EC, et al. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. *J Am Med Inform Assoc.* 2023;30(9):1561-6.

Yasrebi-de Kom IAR, Dongelmans DA, de Keizer NF, Jager KJ, Smit JM, Thorat PJ, et al. Electronic health record-based prediction models for in-hospital adverse drug event diagnosis or prognosis: a systematic review. *J Am Med Inform Assoc.* 2023;30(5):978-88.

Carrasco-Ribelles LA, Costal-Fornells A, Boldú J, Parra JL, Tejedor M, Cano I. Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review. *J Am Med Inform Assoc.* 2023;30(12):2072-82.

Zhang Z, Yan C, Mesa DA, Sun J, Abdulrahman AA, Glicksberg BS, et al. A generalizable method for estimating household-level associations between residential property values and markers of economic segregation using machine learning. *J Am Med Inform Assoc.* 2021;28(3):596-604.

Bergquist T, Yan Y, Schaffter T, Yu T, Velu J, Chaibub Neto E, et al. Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. *J Am Med Inform Assoc.* 2020;27(9):1393-400.

Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. *J Am Med Inform Assoc.* 2018;25(3):289-94.

Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. Perceptions of data set experts on important characteristics of health data sets ready for machine learning: a qualitative study. *JAMA Netw Open.* 2023;6(12):e2345892.  
<https://doi.org/10.1001/jamanetworkopen.2023.45892>.

Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, et al. Development and validation of an artificial intelligence system to optimize clinician review of patient records. *JAMA Netw Open.* 2021;4(7):e2117391.  
<https://doi.org/10.1001/jamanetworkopen.2021.17391>.

Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Netw Open.* 2021;4(2):e210184.  
<https://doi.org/10.1001/jamanetworkopen.2021.0184>.

Yan C, Yan Y, Wan J, Cui Z, Ware J, Schany T, et al. Clinical risk assessment of patients with margins in a safety net hospital using artificial intelligence. *JAMA Netw Open.* 2023;6(10):e2337759.  
<https://doi.org/10.1001/jamanetworkopen.2023.37759>.

Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. *JAMA Netw Open.* 2022;5(1):e2144967.  
<https://doi.org/10.1001/jamanetworkopen.2021.44967>.

Melnick ER, Ong SY, Fong A, Socrates V, Ratwani RM. Characterizing physician EHR use with vendor derived data: a feasibility study and cross-sectional analysis. *J Am Med Inform Assoc.* 2021;28(7):1383-92.