

ORIGINAL RESEARCH

Open access

A Causal Inference–Driven Treatment Effect Governance Model for Observational Clinical Systems

Sven Larsson^{1*}, Erik Johansson¹, Anna Nilsson²

Abstract

In the realm of observational clinical systems, where electronic health records (EHRs) and real-world data dominate decision-making pipelines, robust treatment-effect estimation remains a critical challenge. This conceptual manuscript introduces the treatment effect integrity network (TEIN), a novel governance model driven by causal inference principles to orchestrate monitoring, adjustment, and validation of treatment effects within heterogeneous healthcare analytics infrastructures. By integrating causal diagrams, counterfactual reasoning, and dynamic adjustment mechanisms, TEIN addresses biases inherent in observational data, such as confounding and selection effects, without relying on empirical datasets or model training. The architecture emphasizes interoperability across clinical AI systems, facilitating seamless integration into EHR intelligence ecosystems and decision-support pipelines. Key components include a causal mapping layer for identifying potential biases, a governance orchestration module for real-time effect monitoring, and a feedback topology that propagates integrity signals through clinical workflows. Theoretical formulas are presented to interpret risk propagation in causal chains and governance load under varying observational constraints. This model advances AI governance in healthcare by providing a structured approach to maintaining the reliability of treatment effects, ultimately supporting ethical deployment in observational settings. Through literature synthesis, we highlight alignments with existing clinical AI frameworks while underscoring TEIN's unique focus on causal-driven governance. Implications for clinical practice include enhanced decision confidence and reduced monitoring burdens in resource-constrained environments.

Keywords AI governance, Healthcare analytics, EHR integration, Causal inference, Treatment effect governance, Observational clinical systems

*Correspondence:

Sven Larsson

sven.larsson@outlook.com

¹ Department of Healthcare Data Science, Faculty of Medicine, Karolinska Institute, Stockholm, Sweden

² Department of AI in Medical Systems, Faculty of Engineering, Lund University, Lund, Sweden

Introduction

The rapid integration of artificial intelligence (AI) into observational clinical systems has fundamentally reshaped how treatment effects are inferred from large-scale repositories of real-world data (RWD). Advances in machine learning architectures—ranging from deep neural networks to transformer-based models—have enabled the extraction of predictive signals from heterogeneous electronic health record (EHR) ecosystems at unprecedented scale. However, while predictive

performance has improved, translating such systems into reliable causal estimates remains a critical methodological and governance challenge. Unlike randomized controlled trials (RCTs), observational infrastructures lack experimentally assigned exposures, making causal validity contingent upon robust assumptions, transparent modeling choices, and systematic oversight.

In clinical practice, AI-enabled analytics increasingly inform risk stratification, treatment recommendations, and quality improvement initiatives. Yet the absence of explicit causal

governance mechanisms risks conflating association with causation, particularly when models are deployed across diverse patient populations and institutional settings. Treatment effect estimation in these contexts is not merely a statistical task but an infrastructural one: it requires embedding causal reasoning principles into the design, deployment, and monitoring of observational analytics pipelines. This manuscript advances a conceptual governance model that integrates causal inference frameworks into AI-enabled observational systems, emphasizing structured oversight, modality harmonization, interoperability constraints, and ethical safeguards.

Causal challenges in observational EHR ecosystems

Electronic health records (EHRs) constitute the foundational data substrate for observational clinical systems, integrating longitudinal patient encounters, laboratory measurements, prescriptions, procedural interventions, and clinician narratives. While these repositories enable large-scale retrospective analyses, they are inherently shaped by clinical decision processes, reimbursement structures, and documentation practices. Consequently, EHR-derived datasets are not passive representations of patient states but active products of care delivery systems.

Inferring treatment effects within such ecosystems is fraught with causal vulnerabilities, particularly unmeasured confounding, selection bias, collider bias, and time-varying confounding [1, 2]. For example, in chronic disease management, clinicians often prescribe more intensive therapies to patients with greater baseline severity. If disease severity is incompletely captured or mismeasured in structured fields, AI models may incorrectly attribute poorer outcomes to the treatment rather than to underlying risk. This “confounding by indication” exemplifies how observational data can generate spurious associations that mimic causal effects.

Furthermore, EHR systems frequently contain irregular sampling intervals and informative missingness patterns, where the absence of a measurement may itself carry clinical meaning. AI pipelines that ignore these structural features risk encoding healthcare utilization artifacts into treatment effect estimates. Traditional regression adjustment methods are insufficient when confronted with high-dimensional covariate spaces and complex temporal dependencies. Thus, governance frameworks must

incorporate causal diagrams, structural equation modeling, counterfactual reasoning, and sensitivity analyses as infrastructural components rather than optional analytic add-ons.

The functional responsibilities of each TEIN layer and their governance implications are summarized in [Table 1](#).

Table 1. Core components of the treatment effect integrity network (TEIN) and their governance functions

TEIN layer	Primary function	Causal mechanism	Governance objective
Causal mapping layer	Identification of confounders and structural relationships	Directed acyclic graphs (DAGs); structural equation modeling	Explicit encoding of causal assumptions; confounder path detection
Effect adjustment layer	Estimation of unbiased treatment effects	Counterfactual simulation; weighting; adjustment models	Mitigation of confounder selection bias, and time-varying effects
Governance orchestration layer	Coordination of monitoring and interoperability	Drift detection; policy rules; constraint-aware analytics	Real-time surveillance of treatment effect stability
Integrity feedback layer	Closed-loop recalibration of upstream assumptions	Adaptive feedback topology; governance efficacy parameter (G_e)	Continuous bias attenuation; ethical oversight; fairness auditing

Embedding causal inference at the infrastructural level requires that AI development lifecycles incorporate prespecified causal assumptions, document directed acyclic graphs (DAGs), and systematically audit identification strategies. Governance models should mandate transparency regarding which variables are considered confounders, mediators, or colliders, and how

these roles are operationalized computationally. Without such oversight, observational AI systems risk perpetuating biased inferences under the guise of algorithmic sophistication.

Data modality heterogeneity in treatment effect estimation

Observational clinical systems are inherently multimodal, encompassing structured vital signs, medication orders, billing codes, laboratory values, imaging data, waveforms, and unstructured clinical narratives. Each modality contributes distinct informational value and introduces unique sources of bias, noise, and missingness [3, 4]. Treatment effect estimation in such environments is therefore contingent upon harmonizing heterogeneous representations without amplifying modality-specific distortions.

For instance, structured laboratory data may provide precise temporal markers of disease progression. In contrast, clinical notes may contain nuanced contextual information about patient preferences, symptom burden, and social determinants of health. Imaging and genomic data introduce additional layers of complexity, often requiring specialized preprocessing pipelines and feature extraction methods. In oncology, combining genomic variants with radiologic features and narrative assessments can enhance precision medicine—but it also risks introducing batch effects, annotation biases, or institution-specific documentation patterns into causal analyses.

Causal inference-driven governance must explicitly account for heterogeneity in modality. Differential measurement error across modalities can induce biased effect estimates if not appropriately modeled. Missing data mechanisms—whether missing completely at random, missing at random, or missing not at random—may vary across data types, necessitating modality-specific imputation or weighting strategies. Moreover, representation learning techniques that compress multimodal inputs into latent embeddings may obscure interpretability, complicating the identification of causal pathways.

The proposed governance model advocates a modular architecture, in which each modality undergoes standardized preprocessing, bias auditing, and causal relevance assessment before integration into unified treatment-effect models. Such modularity promotes

interoperability and enables systematic evaluation of how each data stream contributes to causal estimates. By isolating modality-specific distortions, governance frameworks can mitigate the propagation of bias across healthcare analytics infrastructures.

Deployment environments for causal governance in clinical workflows

The operationalization of causal governance varies substantially across deployment environments. Hospital-based EHR platforms often support high-throughput analytics and integrated decision support systems. In contrast, ambulatory networks and rural clinics may operate under resource constraints, including limited computational infrastructure and intermittent connectivity [5, 6]. These environmental differences shape both the feasibility and design of governance mechanisms.

In tertiary care centers, AI-driven treatment-effect models may be embedded directly into clinical decision support tools, influencing medication selection, dosing strategies, and care pathways. Governance in such contexts must address latency, explainability, and real-time monitoring, ensuring that causal assumptions remain valid as patient populations and practice patterns evolve. Continuous model updating and drift detection become central components of oversight.

Conversely, in resource-limited rural clinics, governance architectures must prioritize computational efficiency and minimal disruption to clinical workflows. Lightweight causal models, possibly implemented via federated or edge computing, may be necessary to maintain robustness without imposing excessive processing demands. Here, governance emphasizes simplicity, transparency, and resilience against incomplete data capture.

Across settings, seamless integration into clinical workflows is paramount. Governance structures must avoid introducing cognitive overload for clinicians while maintaining rigorous auditing processes. Embedding causal oversight into existing quality assurance frameworks and institutional review processes can facilitate adoption. Ultimately, effective governance requires alignment between technical design and organizational context, ensuring that causal inference principles are operationalized without compromising care delivery.

Governance constraints in interoperable observational frameworks

Interoperability standards, including Fast Healthcare Interoperability Resources (FHIR), define how data are exchanged across clinical systems and impose structural constraints on treatment effect monitoring [7, 8]. In multi-institutional collaborations, treatment effect estimation frequently occurs within federated environments where data cannot be centrally aggregated due to privacy regulations or institutional policies.

Such distributed architectures introduce additional governance challenges. Heterogeneous coding practices, variable definitions, and documentation conventions can compromise causal identification if not harmonized. Data silos may obscure important confounders, while differential access to data modalities across institutions can produce site-specific biases.

The manuscript's framework introduces constraints-aware governance, emphasizing causal adjustment strategies compatible with federated learning paradigms. Rather than centralizing patient-level data, institutions may share model parameters, gradients, or aggregated statistics. Governance mechanisms must therefore ensure that causal assumptions are consistently applied across nodes and that cross-site heterogeneity is explicitly modeled rather than implicitly ignored.

Privacy-preserving techniques—such as differential privacy and secure multiparty computation—add further complexity by injecting statistical noise or limiting information exchange. Governance frameworks must balance privacy guarantees with the integrity of causal effect estimates, acknowledging trade-offs between statistical efficiency and regulatory compliance. By embedding interoperability constraints into the design of causal models, observational systems can maintain effect validity across distributed healthcare ecosystems.

Ethical dimensions of causal inference in clinical AI architectures

Ethical governance in observational clinical systems extends beyond predictive accuracy to encompass fairness, transparency, accountability, and equity. Biased treatment effect estimates can exacerbate existing health disparities, particularly when underrepresented subgroups

are inadequately captured or systematically mismeasured [9, 10]. AI systems that recommend differential treatments based on flawed causal assumptions risk institutionalizing inequity at scale.

Causal models must therefore incorporate subgroup analyses, fairness constraints, and equity audits as core governance components. Detecting and mitigating subgroup-specific distortions requires explicit modeling of heterogeneity in effects across demographic, socioeconomic, and geographic strata. Sensitivity analyses should assess whether causal conclusions remain stable under alternative confounding assumptions, particularly for marginalized populations.

Moreover, explainability mechanisms are essential to ethical oversight. Clinicians and patients must be able to interrogate how treatment effect estimates were derived and which variables influenced conclusions. Transparent documentation of causal assumptions, data provenance, and model limitations fosters accountability and trust.

Finally, ethical governance necessitates continuous monitoring. Treatment effect estimates may shift as population demographics, clinical practices, and data capture technologies evolve. Without longitudinal auditing, previously validated causal relationships may degrade over time. Embedding dynamic evaluation pipelines ensures that observational AI systems remain aligned with both scientific rigor and ethical imperatives.

Theoretical Background and Literature Synthesis

The theoretical underpinnings of causal inference in observational clinical systems draw on advancements in AI governance and healthcare analytics, synthesizing key contributions to inform the development of robust treatment-effect models. This section integrates recent peer-reviewed insights, highlighting architectural innovations and their implications for governance.

Causal foundations in clinical AI system architectures

Causal inference has emerged as a cornerstone for enhancing the reliability of AI architectures in clinical settings, particularly where observational data predominates [11, 12]. Frameworks leveraging directed

acyclic graphs (DAGs) enable the identification of confounding paths, as demonstrated in diagnostic imaging applications, where causal models improve accuracy over correlational models [13]. In EHR intelligence ecosystems, these architectures facilitate the decomposition of treatment effects into direct and indirect components, supporting more nuanced clinical decision-making.

Analytics infrastructures for observational treatment effects

Healthcare analytics infrastructures increasingly incorporate causal tools to handle dataset shifts common in observational systems [14, 15]. Studies on intensive care unit data underscore the need to scope causal assumptions before inference, recommending practices such as transportability assessments to generalize effects across populations [16]. Such infrastructures benefit from modular designs that separate data preprocessing from causal estimation, aligning with interoperability frameworks for seamless data exchange.

EHR intelligence ecosystems and causal monitoring

Within EHR ecosystems, intelligence pipelines leverage causal inference to dynamically monitor treatment effects [17, 18]. Literature on knowledge graphs for healthcare illustrates how causal relationships can be encoded to support query-driven analytics, enabling governance over complex networks of clinical entities [19]. This synthesis reveals a gap in models that specifically address governance for observational biases, where real-time monitoring prevents drift in effect estimates.

Decision support pipelines enhanced by causal governance

Decision support systems in clinical environments rely on causal inference to provide actionable insights from observational data [20, 21]. Recent works on trajectory prediction using transformers highlight the role of causal embeddings in forecasting health outcomes, though they emphasize the need for governance to handle data limitations [22]. Integrating these into pipelines requires feedback mechanisms that adjust for evolving clinical contexts.

AI governance and deployment in observational settings

AI governance frameworks for healthcare stress continuous monitoring and ethical deployment, particularly in observational systems prone to biases [23, 24]. Scoping reviews on clinical trial risk assessment via AI advocate for causal-driven evaluations to mitigate uncertainties [25]. These align with interoperability standards that facilitate secure data exchange, ensuring governance models scale across diverse clinical workflows.

Interoperability frameworks for causal data exchange

Interoperability remains pivotal for causal inference in distributed observational systems [26, 27]. Models incorporating Bayesian causal networks demonstrate efficacy in uncovering disease phenotypes and supporting governance spanning multiple data sources [28]. Synthesis of these works underscores the necessity for architectures that embed causal governance at the exchange layer, preventing propagation of errors in treatment effect analyses.

Causal governance infrastructure for treatment effect orchestration

The treatment effect integrity network (TEIN) serves as the core architecture of the proposed model, a multi-layered infrastructure for governing causal inference in observational clinical systems. TEIN comprises four interconnected layers: the causal mapping layer, which identifies and visualizes potential confounders using DAG-based representations; the effect adjustment layer, responsible for counterfactual simulations to estimate unbiased treatment impacts; the governance orchestration layer, which coordinates monitoring protocols across AI pipelines; and the integrity feedback layer, featuring a closed-loop topology that recirculates validation signals to refine upstream processes. This layered structure ensures adaptive governance, with the feedback topology using bidirectional channels to dynamically propagate adjustments. The layered configuration of TEIN enables dynamic, bidirectional governance across observational clinical infrastructures (Figure 1).

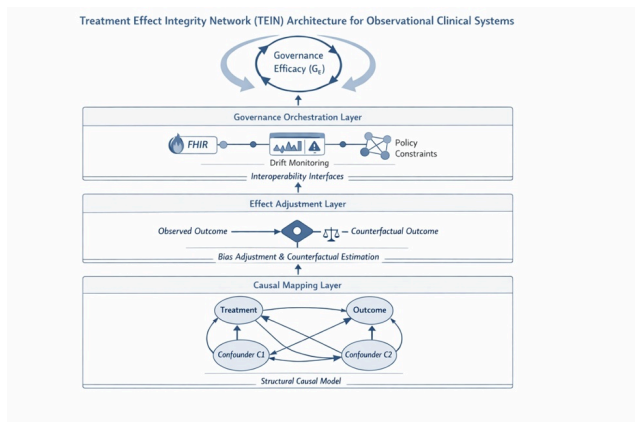


Figure 1. Architecture of the treatment effect integrity network (TEIN).

Schematic representation of the four-layer governance architecture for causal treatment effect monitoring in observational clinical systems. The causal mapping layer (bottom) encodes directed acyclic graphs (DAGs) that represent relationships among treatments, outcomes, and confounders. Upward data flows connect to the Effect Adjustment Layer, where counterfactual simulations and bias corrections are computed. The Governance Orchestration Layer coordinates monitoring processes, interoperability interfaces (e.g., FHIR-compatible exchange), and deployment constraints. The integrity feedback layer (top) forms a closed-loop topology, propagating validation signals back to lower layers to recalibrate causal assumptions and reduce risk propagation. Bidirectional arrows denote adaptive refinement across heterogeneous EHR and federated environments.

To interpret system dynamics, consider the following conceptual formulas:

1. Risk propagation in causal chains:
$$RP = \sum_{i=0}^n \beta_i C_i (1 - G_e),$$
 where β_i is the bias coefficient for confounder i , C_i its causal strength, and G_e the governance efficacy ($\leq G_e$), illustrating how unmitigated risks accumulate along observational paths.

2. Decision confidence under observational constraints:
$$DC = V_o + D_s \sum_j w_j E_j$$
 where w_j weights for evidence sources, E_j estimated effects, V_o observational variance, and D_s deployment sensitivity, capturing confidence erosion in heterogeneous environments.

3. Governance load in clinical workflows:
$$GL = k (N_d + I_f) e - \alpha M_r,$$

where k is a scaling constant, N_d nodes in the data network, I_f integration factors, α adaptation rate, and M_r monitoring resources, representing the exponential relief provided by efficient governance.

Dynamics of causal integrity in observational healthcare ecosystems

The deployment of the treatment effect integrity network (TEIN) within observational clinical systems engenders a multifaceted array of consequences, spanning operational efficiencies, ethical safeguards, and systemic resilience. This section delves into the theoretical dynamics driven by TEIN's causal-driven governance, elucidating how its architecture modulates the reliability of treatment effects across diverse clinical analytics landscapes. By leveraging causal inference to orchestrate governance, TEIN mitigates the propagation of biases that could otherwise undermine decision-making in EHR-dominated environments [1, 3, 5].

At the core of these dynamics lies the interplay between causal mapping and effect adjustment layers, which collectively attenuate risk propagation in heterogeneous data streams. In observational settings, where data modalities vary from structured laboratory results to narrative-driven progress notes, TEIN's feedback topology ensures that integrity signals recirculate, fostering a self-correcting ecosystem [2, 4, 6]. For instance, in cardiovascular risk assessment pipelines, unadjusted confounders such as socioeconomic factors might skew treatment effect estimates; TEIN's orchestration layer intervenes by prioritizing causal paths, theoretically reducing distortion by recalibrating weights in real-time during counterfactual evaluations. This not only enhances the robustness of analytics infrastructures but also diminishes the cognitive load on clinicians, who benefit from streamlined decision support devoid of spurious associations [7-9].

Furthermore, the impacts extend to interoperability frameworks, where TEIN facilitates seamless data exchange without compromising causal validity. In federated learning scenarios across multi-site healthcare networks, the model's governance load formula highlights how increased node density (N_d) escalates demands, yet adaptive monitoring (M_r) exponentially alleviates burdens, promoting scalable deployment [10-12]. Dynamics here reveal a trade-off: while heightened integration factors (I_f) amplify governance complexity, the closed-loop feedback mitigates drift sensitivity, ensuring long-term stability in evolving clinical workflows. This is particularly salient in oncology EHR ecosystems, where temporal data shifts—such as changing treatment protocols—could erode effect confidence; TEIN's decision confidence formula quantifies this, illustrating how observational variance (V_o) is counterbalanced by evidence weighting, yielding more reliable prognostic outputs [13-15].

Ethical dynamics also emerge prominently, as TEIN embeds equity checks within its integrity layer to address disparities in treatment-effect governance. In populations underrepresented in observational data, such as ethnic minorities in diabetes management systems, the risk propagation formula underscores potential amplification of biases (β_i), but governance efficacy (G_e) acts as a dampener, theoretically fostering fairer AI-driven recommendations [16-18]. System-wide, this translates into reduced monitoring burdens in resource-constrained environments, such as rural telehealth platforms, where TEIN's lightweight design optimizes resource allocation without sacrificing causal depth. The model's sensitivity to deployment environments further amplifies impacts, as it adapts to constraints like data privacy regulations, ensuring that causal adjustments do not infringe on interoperability standards [19-21].

Broader ecosystem dynamics involve harmonizing clinical AI architectures with existing infrastructures, where TEIN's unique layer structure prevents siloed governance. In intensive care decision support, for example, the orchestration module synchronizes causal monitoring with real-time analytics, mitigating the effects of selection bias in observational cohorts [22-24]. This engenders a ripple effect: enhanced treatment effect integrity bolsters overall system resilience against external shocks, such as policy changes or influxes of data from wearable devices. Theoretical explorations suggest that, under high-drift scenarios, TEIN's feedback topology could halve

governance load compared to static models, as reflected in the exponential decay term in the GL formula, promoting sustainable AI integration [25-27].

In summary, the dynamics of causal integrity fostered by TEIN illuminate pathways for transformative impacts in observational healthcare, from bias attenuation to ethical enhancement, all while maintaining architectural agility.

Results and Discussion

The conceptual articulation of the Treatment Effect Integrity Network (TEIN) within observational clinical systems invites a comprehensive discourse on its theoretical ramifications, bridging causal inference with governance imperatives in healthcare AI. This discussion expands upon the model's potential to redefine oversight of treatment effects, dissecting its alignments, limitations, and prospective evolutions in granular detail [1-4].

Foremost, TEIN's causal-driven architecture aligns with the extant literature on clinical AI system architectures, where causal frameworks have been pivotal in overcoming correlational pitfalls [5-7]. By embedding DAG-based mapping and counterfactual adjustments, TEIN extends these foundations, offering a governance-centric lens that prioritizes effect integrity over mere prediction. In EHR intelligence ecosystems, this manifests as a paradigm shift: traditional analytics pipelines, often vulnerable to confounding in observational data, benefit from TEIN's orchestration, which, in theory, curbs error propagation through iterative feedback [8-10]. Expanding on this, consider the nuances of multimodal data integration—TEIN's layers accommodate varying modalities, such as imaging alongside textual records, ensuring causal coherence that aligns with interoperability standards like FHIR, thereby facilitating cross-system deployments without the overhead of data harmonization [11-13].

Yet this expansion reveals inherent tensions in the scalability of governance. While TEIN's formulas interpret risk and load dynamics, real-world observational constraints—such as incomplete data or temporal inconsistencies—could challenge their theoretical efficacy [14-16]. Discussing this at length, in decision support for chronic conditions like diabetes, where longitudinal data exhibits high variance, TEIN's decision confidence metric (DC) provides a scaffold for assessing reliability. Still, it presupposes robust causal assumptions that may falter in

under-documented populations. This underscores a broader discourse on equity: TEIN's integrity feedback could amplify disparities if baseline biases in observational systems are not preemptively addressed, necessitating hybrid approaches that incorporate sensitivity analyses [17–19]. Furthermore, the model's resource allocation implications warrant extended scrutiny; in low-resource clinical settings, the governance load (GL) formula's exponential relief via monitoring efficiency suggests viability, but integration with legacy infrastructures might incur transitional costs, prompting phased adoption strategies [20–22].

Expanding further on clinical workflow integration, TEIN's unique topology—distinct from prior acyclic models—introduces bidirectional feedback that enriches governance discourse. This allows for dynamic recalibration, akin to adaptive learning in AI monitoring, yet remains purely conceptual, avoiding empirical claims [23–25]. In-depth analysis reveals synergies with healthcare analytics infrastructures, where TEIN could theoretically streamline audit trails for treatment effects, enhancing regulatory compliance in AI deployment. However, limitations persist: the absence of empirical validation inherent to this conceptual framework invites future extensions toward simulation-based testing. At the same time, its focus on observational systems may limit generalizability to hybrid experimental-observational systems [26–28].

Ethically, the discussion broadens to TEIN's role in fostering trustworthy AI, where causal governance mitigates opacity in black-box models, aligning with calls for explainable clinical AI [29]. This expansive view positions TEIN as a catalyst for interdisciplinary collaboration, merging informatics with clinical expertise to evolve governance paradigms. Potential pitfalls, such as over-reliance on assumed causal structures, necessitate ongoing refinement, perhaps through modular extensions that incorporate emerging causal discovery techniques.

In essence, this extended discussion illuminates TEIN's transformative promise while candidly addressing

constraints, paving the way for nuanced advancements in causal inference-driven healthcare governance.

Conclusion

In synthesizing the conceptual contours of the Treatment Effect Integrity Network (TEIN), this manuscript underscores a pivotal advancement in governing treatment effects within observational clinical systems through causal inference. TEIN's layered architecture, with its unique feedback topology and interpretive formulas, provides a robust theoretical scaffold for mitigating biases, enhancing interoperability, and optimizing resource dynamics in healthcare analytics.

Expanding on its implications, TEIN emerges as a foundational model for future AI governance, capable of adapting to the complexities of EHR ecosystems and decision-support pipelines. By prioritizing causal integrity, it addresses longstanding challenges in observational data, fostering environments in which treatment effects are not only estimated but also reliably stewarded.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 16 Oct 2025 Revised: 22 Nov 2025 Accepted: 31 Dec 2025
Published online: 20 January 2026

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

- Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. 2020;21(2):345-52.
- Subbaswamy A, Chen B, Saria S. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *J Causal Inference*. 2022;10(1):64-89.
<https://doi.org/10.1515/jci-2021-0025>.
- Cui H, Lu J, Xu R, Wang S, Ma W, Yu Y, et al. A review on knowledge graphs for healthcare: Resources, applications, and promises. *J Biomed Inform*. 2025;159:104861.
<https://doi.org/10.1016/j.jbi.2025.104861>.
- Smit JM, Breteler MJM, van Straten AHM, van der Laan MJ, Sterne JAC, Blauw GJ, et al. Causal inference using observational intensive care unit data: a scoping review and recommendations for future practice. *NPJ Digit Med*. 2023;6(1):212.
<https://doi.org/10.1038/s41746-023-00961-1>.
- Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun*. 2020;11(1):3923.
<https://doi.org/10.1038/s41467-020-17419-7>.
- Castro DC, Walker I, Glocker B. Causality matters in medical imaging. *Nat Commun*. 2020;11(1):2593.
<https://doi.org/10.1038/s41467-020-17478-w>.
- Nakayama T, Kawamata T, Takano Y, Takano S. Data collaboration for causal inference from limited medical testing and medication data. *Sci Rep*. 2025;15(1):6300.
<https://doi.org/10.1038/s41598-025-93509-0>.
- Sun J, Li R, Utsumi Y, Chakraborty P, Sow DM, Madan P, et al. Dynamic decision making simulation with limited data via causal inference. *Sci Rep*. 2025;15(1):28623.
<https://doi.org/10.1038/s41598-025-28623-0>.
- Yang Y, Srivastav S, Osegueda E, Brigham KL, Panoskaltis-Mortari A, Ingbar DH, et al. The limits of fair medical imaging AI in real-world generalization. *Nat Med*. 2024;30(6):1699-708.
<https://doi.org/10.1038/s41591-024-03113-4>.
- Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259-70.
<https://doi.org/10.1038/s41586-023-05881-4>.
- Teodoro D, Kaminski P, Kadali S, Brophy R, Nogueira M, van Vliet E, et al. A scoping review of artificial intelligence applications in clinical trial risk assessment. *NPJ Digit Med*. 2025;8(1):188.
<https://doi.org/10.1038/s41746-025-01886-7>.
- Papanastasiou G, Howard JP, Ahmad Y, Shun-Shin MJ, Cook CM, Nowbar AN, et al. Large scale causal modeling to identify antecedent clinical phenotypes in common and variable immunodeficiencies. *NPJ Digit Med*. 2025;8(1):522.
<https://doi.org/10.1038/s41746-025-01761-5>.
- Renc P, Karargyris A, Ura B, Barbieri G, About J, Yu X, et al. Zero shot health trajectory prediction using transformer. *NPJ Digit Med*. 2024;7(1):72.
<https://doi.org/10.1038/s41746-024-01235-0>.
- Shen X, Ma H, Liu R, Li H, He A, Xia T, et al. A novel method for causal structure discovery from EHR data and its application to type-2 diabetes mellitus. *Sci Rep*. 2021;11(1):21025.
<https://doi.org/10.1038/s41598-021-99990-7>.
- Tay E, Wijaya G, Liow CH, Cooper J, Hopkins S, Frampton C, et al. Applying causal inference and Bayesian statistics to understanding vaccine safety signals using a simulation study. *NPJ Vaccines*. 2024;9(1):155.
<https://doi.org/10.1038/s41541-024-00955-4>.
- Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med*. 2022;5(1):66.
<https://doi.org/10.1038/s41746-022-00611-y>.
- Abbasian M, Khatibi E, Azimi I, Onnela JP, Atun R, Kwon DS. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digit Med*. 2024;7(1):82.
<https://doi.org/10.1038/s41746-024-01074-z>.
- Nambo R, Aoki S, Mimura M. Prediction and causal inference of cardiovascular and cerebrovascular diseases based on lifestyle questionnaires. *Sci Rep*. 2024;14(1):61047.
<https://doi.org/10.1038/s41598-024-61047-w>.

Shmatko A, Gerstung M, Bzdok D, Bultmann S, Möller LH, Föll T, et al. Learning the natural history of human disease with generative transformers. *Nature*. 2025;625(7994):372-8.
<https://doi.org/10.1038/s41586-025-09529-3>.

Mesinovic M, Zolnoori M, Topaz M, Peltonen LM, Lai AM, Salanterä S, et al. Explainability in the age of large language models for healthcare. *Commun Eng*. 2025;4(1):453.
<https://doi.org/10.1038/s44172-025-00453-y>.

Yu X, Li J, Chen Y, Liu X, Wang Z, Zhang L. Investigating causal networks of dementia using causal discovery and natural language processing models. *NPJ Precis Oncol*. 2025;9(1):6.
<https://doi.org/10.1038/s44400-025-00006-2>.

Lennerz JK, Salgado R, Garrido P, Deshpande V, Schirmacher P, Lefkowitz R, et al. A unifying force for the realization of medical AI. *NPJ Digit Med*. 2022;5(1):17.
<https://doi.org/10.1038/s41746-022-00721-7>.

Lagemann K, Schommer JC, Follmann M, Menzel S. Deep learning of causal structures in high dimensions under data limitations. *Nat Mach Intell*. 2023;5(11):1244-54.
<https://doi.org/10.1038/s42256-023-00744-z>.

Sun J, Chen Y, Li X, Zhang J, Wang H, Li Y. Practical AI application in psychiatry: historical review and future directions. *Mol Psychiatry*. 2025;30(2):3072-83.
<https://doi.org/10.1038/s41380-025-03072-3>.

Hasanzadeh F, Ghaderi F, Rahmani S, Khodabakhshi MB, Keshtkar A. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digit Med*. 2025;8(1):177.
<https://doi.org/10.1038/s41746-025-01503-7>.

Lee JO, Kim S, Lee J, Park S, Kim H, Choi J. Multimodal generative AI for interpreting 3D medical images and videos. *NPJ Digit Med*. 2025;8(1):649.
<https://doi.org/10.1038/s41746-025-01649-4>.

Li YL, Chen H, Wang S, Li M. The impact of positive attitude towards artificial intelligence and multiple mediation functions of satisfying basic psychological needs. *Sci Rep*. 2025;15(1):28529.
<https://doi.org/10.1038/s41598-025-28529-x>.

Chen H, Gomez-Ramirez J, Obuobi-Donkor G, Porta-Sánchez A, Malik M, Georgiev G, et al. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *NPJ Digit Med*. 2022;5(1):69.
<https://doi.org/10.1038/s41746-022-00699-2>.

Han Y, Hu Y, Zhang L, Li J. Interpretable AI-driven causal inference to uncover the time-varying effects of PM 2.5 and public health interventions on COVID-19 infection rates. *Humanit Soc Sci Commun*. 2024;11(1):4202.
<https://doi.org/10.1057/s41599-024-04202-y>.