Leveraging Generative AI for Interpretable Clinical Decision Making Through Causal Graphs

Mehmet Eren Ahsen PhD[1], Rand Kittani[2], Travis Gerke ScD[3], Laya Krishnan[2], Sean Rogan[3], and Erick R. Scott MD MHS[3†]

[1] *Gies College of Business,* [2] *Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Champaign, IL, 61820, USA*
[3] *cStructure, La Jolla, CA, 92037, USA*
[†] *E-mail: erick@cstructure.io*

Clinical AI systems' lack of interpretability limits their adoption in evidence-based medicine. To address this challenge, we propose a computational framework that harnesses generative AI's medical knowledge to create interpretable structural causal models (SCMs) for clinical decision support, quality improvement evaluation, and population health management. We evaluated our approach through a case study using data from the Midwest Healthcare Conference Causal Diagram Challenge, where we compared transformer-based large language models against human performance on a complex causal reasoning task: estimating COVID-19 treatment effects through target trial emulation. Both groups designed SCMs to evaluate glucocorticoid treatment effects on 28-day mortality using real-world data from more than 2,000 hospitalized patients, benchmarked against published RECOVERY randomized controlled trial results. The best performing SCMs achieved bootstrap coverage rates exceeding 90% for two of three severity strata. Both human and AI models demonstrated equivalent clinical plausibility (n=3 expert reviewers) and similar statistical performance, though both struggled with critical disease severity. Ablation experiments comparing SCM-based approaches against traditional potential outcomes methods revealed SCMs achieved 76-98% coverage versus 1-37% for traditional methods. These results suggest that structural causal models can effectively bridge the interpretability gap in clinical AI by providing essential scaffolding for reliable causal inference and enabling meaningful human-AI collaboration while preserving methodological rigor essential for evidence-based medicine.

Keywords: Causal Inference, Large Language Models, Artificial Intelligence, Clinical Informatics, Structural Causal Models, AI-Supported Interactive Evaluation

1. Introduction

Clinical artificial intelligence poses interpretability and explainability challenges.[1] Although transformer-based large-language models demonstrate impressive performance metrics,[2] their black-box nature delays meaningful and safe integration into clinical decision-making workflows.[3] Regulatory health agencies require explainable AI systems when the device is intended

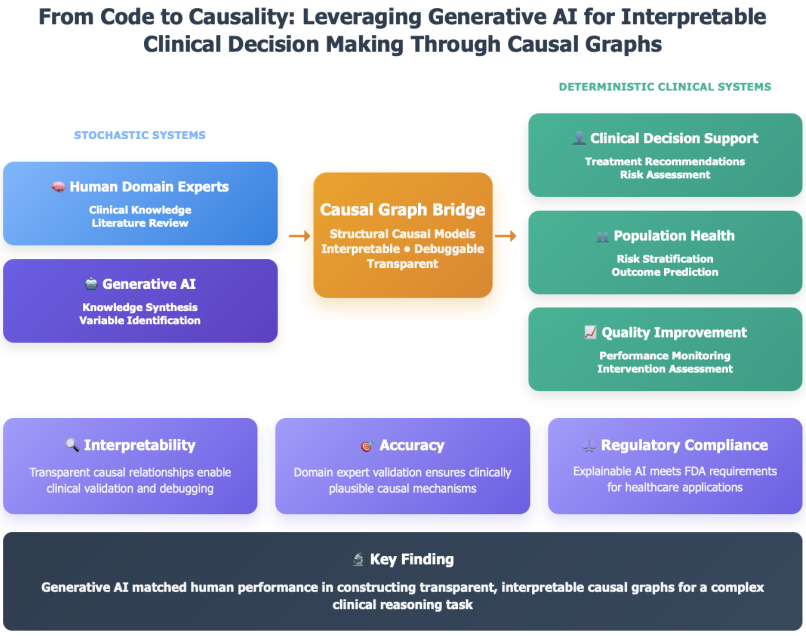


Fig. 1. Graphical Abstract

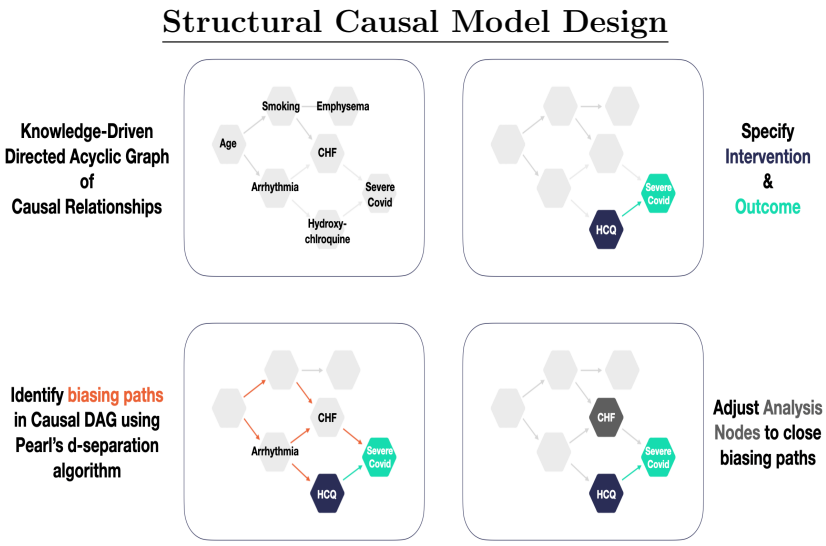
to diagnose, treat, cure, mitigate, or prevent disease.[4] However, most current AI approaches sacrifice interpretability for accuracy and rely upon post-hoc explanations that fail to reliably provide causal reasoning underlying their recommendations.[5–7]

Generative AI has demonstrated competency solving a limited set of programming tasks by converting probabilistic text generation into deterministic executable systems.[8] The key insight is in converting uncertain outputs into structured and interpretable representations that can be debugged, validated, and iteratively refined by humans.

The same principle can be applied to clinical reasoning. Generative AI's medical domain knowledge can be converted into interpretable structural causal models (SCMs) that provide transparent, debuggable, and quantitative representations of clinical knowledge (**Figure 1**). We evaluate this framework by comparing generative AI and human performance in designing structural causal models. The evaluation uses results from the Midwest Healthcare Conference Causal Diagram Challenge, where human participants without formal causal inference training and with varying levels of domain knowledge designed SCMs to estimate the safety and efficacy of COVID-19 treatment strategies (glucocorticoids in three disease severity contexts) using real-world electronic health record (EHR) data. This study provides empirical evidence for how structural causal models can bridge the gap between model intelligence and human reasoning in clinical medicine.

Collaborative Target Trial Emulation: Framework, Process, and Performance

Causal Roadmap Steps	Challenge Implementation
Step 1: Causal Question, Model, Estimand Define causal question and construct DAG	**Target Trial Emulation** Steroid effects on 28-day survival, stratified by severity
Step 2: Observed Data Assess data quality	**SDY1662 Dataset** 2,074 patients, person-time transformation
Step 3: Identifiability Evaluate assumptions	**cStructure Platform** Automated d-separation analysis
Step 4: Statistical Estimand Translate to statistical quantity	**Risk Ratio** 28-day survival with severity stratification
Step 5: Model Selection Choose statistical methods	**Pooled Logistic Regression** G-formula
Step 6: Sensitivity Analysis Assess robustness	**Bootstrap Validation** 500 samples per severity level
Step 7: Compare Designs Evaluate alternatives	**Human vs. AI** Collaborative comparison



Dimension	Transformer Models	Structural Causal Models
Breadth of Knowledge	+++	N/A
Speed	+++	–
Creativity	++	N/A
Natural Language Processing	++	–
Factual Accuracy	+	+++
Transparency	—	+++
Reproducibility	—	++
Quantitative Reasoning	—	++

While transformer-based LLMs excel in knowledge breadth and speed, SCMs provide superior accuracy, transparency, and reproducibility essential for clinical applications. Complementary strengths suggest hybrid approaches leveraging LLM knowledge within structural causal frameworks.

Fig. 2. **Collaborative Target Trial Emulation Framework and LLM-SCM Integration. Left:** Implementation of the Causal Roadmap framework in the UIUC Challenge demonstrating systematic translation from clinical questions to validated causal models. **Top right:** DAG construction process showing progression from domain knowledge specification to automated validation. **Bottom right:** Comparison of Large Language Models and Structural Causal Models across key dimensions, highlighting complementary strengths that enable hybrid approaches combining AI knowledge generation with rigorous causal inference.

2. Background and Related Work

2.1. *Causal Inference in Clinical Research*

The Causal Roadmap provides a systematic approach for generating high-quality real-world evidence through explicit specification of causal questions, identification assumptions, estimands, and statistical estimators, (**Figure 2, left panel**).[9] Target trial emulation employs this framework to design observational studies that carefully emulate randomized controlled trials (RCTs), enabling rigorous causal inference from real-world data.[10] Despite methodological advances, causal inference remains underutilized in clinical practice. Current tools require extensive statistical programming expertise, creating barriers for clinical researchers who have

essential domain knowledge—precisely those best positioned to identify confounders, evaluate clinical plausibility, and validate causal assumptions.

2.2. *Structural Causal Models as an Interpretable AI Framework*

Structural causal models (SCMs) represent causal relationships as directed acyclic graphs (DAGs) where nodes denote variables and edges encode causal dependencies between them.[11] The construction of SCMs follows a systematic process that transforms clinical knowledge into formal causal structures (**Figure 2, top-right panel**). Domain experts specify variables relevant to the clinical question, define directional causal relationships based on biological mechanisms and temporal precedence, and use Pearl's d-separation algorithm to identify biasing pathways that must be controlled through statistical adjustment to obtain valid causal estimates..[12]

Unlike black-box machine learning models, SCMs make causal assumptions explicit through visual representations that clinical experts can evaluate, modify, and validate against biological knowledge derived from randomized and observational studies.[13] This 'glass-box' approach enables target trial emulation for therapeutic safety and efficacy evaluation, clinical decision support for precision medicine, performance monitoring of quality improvement programs, and outcome prediction for population health management.[14] This causal graph approach addresses fundamental limitations of current clinical AI systems by providing interpretable, debuggable representations of causal reasoning. SCMs enable iterative refinement of causal assumptions, transparent communication of modeling decisions, and systematic evaluation of model validity against established clinical knowledge.[15]

2.3. *Human-AI Collaboration in Clinical Decision Making*

Effective clinical AI requires seamless integration of algorithmic capabilities with human expertise, as purely automated approaches often fail to capture nuanced clinical reasoning while manual approaches cannot scale to complex medical data.[16,17] Current systems typically provide predictions without exposing underlying reasoning processes, limiting meaningful human oversight—a critical gap given regulatory requirements for transparent, interpretable AI outputs that clinicians can evaluate and validate.[4]

Large language models show promise for clinical documentation and treatment planning but lack mechanisms for incorporating causal relationships or clinical knowledge in systematic, verifiable ways.[18] Complex causal reasoning tasks further challenge these models through long context windows and the 'lost-in-the-middle' problem.[19] However, the success of generative AI in programming tasks suggests a pathway: transforming probabilistic language generation into structured, interpretable representations that clinical experts can refine (**Figure 2, bottom-right panel**). We present evidence that structural causal models provide the necessary framework for this transformation, bridging model intelligence with human clinical expertise.

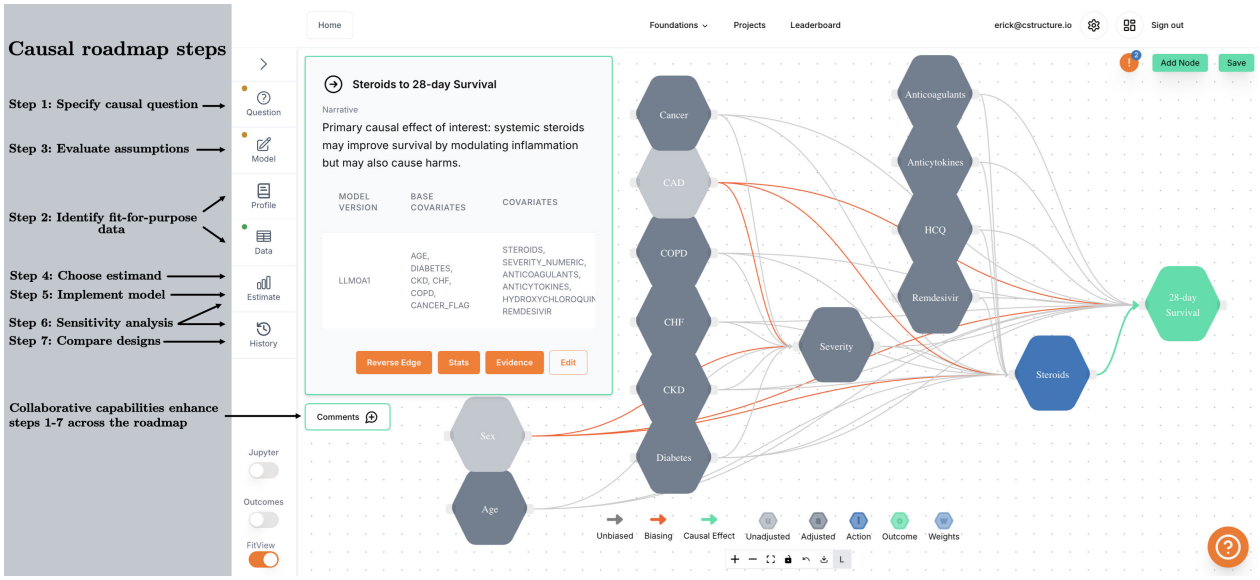


Fig. 3. Building knowledge-based Structural Causal Models in cStructure. Creating Edges: Participants could create directed edges between nodes to represent causal relationships. For example, they might add an edge from "Age" to "28-day Survival" to indicate that age directly affects survival outcomes. Adding Narratives: For each node and edge, users could provide detailed narratives explaining the node and its causal relationship with other nodes. These narratives were crucial for explaining the reasoning behind each causal connection and for evaluation of the model's plausibility. Users were encouraged to organize their graphs logically, typically with a left-to-right flow reflecting temporal precedence, where causes precede effects. This SCM along with the narrative text was generated by o3. User workflows in the cStructure platform are designed to follow the causal roadmap.

3. Methods

3.1. *Study Design and Participants*

We organized the UIUC Midwest Healthcare Conference Challenge to evaluate target trial emulation performance by participants without formal causal inference training. **Conference participants** were recruited without background restrictions and designed structural causal models via a collaborative interface to replicate RECOVERY trial effects of COVID-19 treatment using observational data, with the RECOVERY Trial serving as the gold-standard benchmark due to its randomized controlled design.[20] **Data access:** Participants received comprehensive SDY1662 metadata (data dictionaries, variable definitions, statistical profiles) but no patient-level data, with all analyses occurring via secure API calls through the cStructure interface. **AI comparison:** One year later, we conducted comparative analysis using four state-of-the-art generative AI models (OpenAI o3, DeepSeek R1, Gemini Pro 2.5, Claude Opus).

3.2. *Dataset and Target Trial Framework*

Dataset: We utilized ImmPort's SDY1662 dataset containing >2,000 hospitalized COVID-19 patients from Mount Sinai Health System (March-May 2020), transformed into person-time format (∼65,000 person-day observations) for pooled logistic regression.[21] **Severity classi-**

fication: Disease severity was classified using oxygen support requirements: mild-moderate (room air or low-flow oxygen, $SpO_2 \geq 94\%$ on $\leq$ 4L O_2), severe (high-flow oxygen, non-invasive/mechanical ventilation without vasopressors), and critical (mechanical ventilation with vasopressors or end-organ dysfunction: creatinine clearance < 30 mL/min or ALT $> 5\times$ ULN).[22] Missing data was handled using last observation carried forward; variables >50% missing were excluded. **Mini-cohorts:** To minimize immortal time bias, we created daily mini-cohorts evaluating treatment-eligible patients and assigning them to treatment strategies reflecting observed exposure, with patients appearing multiple times under different assignments and censoring upon strategy deviation.[23,24] For example, a patient receiving steroids on hospital day 2 would be classified as 'never treat' in mini-cohorts 0-1 but '10 days of steroids' in mini-cohort 2. **Target trial:** Participants estimated causal effects of glucocorticoids on 28-day all-cause mortality stratified by severity, applying RECOVERY eligibility criteria: hospitalized patients ≥18 years with COVID-19, no contraindications, sufficient follow-up, comparing glucocorticoid versus standard care. Data available at ImmPort.org under accession SDY1662.

3.3. *Model Construction and Platform Implementation*

cStructure platform: The platform provides a browser-based collaborative environment with interactive DAG editing, real-time d-separation analysis for confounding pathway identification, and JSON model serialization transmitted via REST API to secure analysis servers, with patient data remaining on secure servers while participants constructed models using metadata only (**Figure 3**). Real-time d-separation algorithms (TypeScript) identified open backdoor paths between treatment and outcome, providing automated feedback on confounding control. **Human-generated SCMs:** Participants constructed DAGs specifying nodes from dataset features, directional edges representing causal relationships, scientific rationale documentation, and node classifications (Action/Outcome/Adjusted/Unadjusted), with automated validation including d-separation analysis, temporal ordering checks, and convergence assessment. **Transformer-generated SCMs:** AI models received identical materials as human participants: challenge announcement, SDY1662 data dictionary, cStructure platform description, and GraphML specification format, with standardized prompts requesting causal diagrams evaluating steroid safety/efficacy on 28-day survival (**Supplementary Materials**). **SCM ablation study:** To test whether causal graphs provide unique benefits beyond LLM reasoning, we conducted controlled experiments with multiple models (OpenAI o3, Claude Opus, Gemini Pro 2.5, DeepSeekV1) receiving Potential Outcomes tasks requesting traditional causal inference methods (IPTW, AIPW, g-computation) without graph structure, using identical data dictionaries, bootstrap requirements (n=500), and target trial descriptions, isolating the effect of structured causal reasoning while holding constant statistical objectives, data access, model capabilities, and evaluation framework.

3.4. *Statistical Analysis and Evaluation*

Causal Plausibility: Three independent and blinded clinical reviewers evaluated the causal plausibility of the directed edges from Human (n=36) and Transformer-generated SCMS (n=86) using an ordinal scale (0=no plausible causal pathway, 1=low/weak rationale, 2=mod-

erate/plausible mechanism, 3=high/strong evidence). A Bayesian hierarchical model was used to compare plausibility scores between the two groups. **G-formula estimation:** Causal effects were estimated using the parametric g-formula (PyGFormula v1.1.6) with 500 bootstrap samples per severity stratum (seed=123), selected to overcome convergence issues encountered during the challenge, with bootstrap sampling at patient level to preserve within-patient correlations and risk ratios comparing predicted 28-day survival probabilities under treatment versus standard care. **Coverage analysis:** Bootstrapped confidence interval coverage against RECOVERY trial results using 500 bootstrap samples per severity stratum, ranking models by bootstrap estimates falling outside RECOVERY 95% CIs across all strata (n=1,500 total), with two-way binomial tests comparing each model's coverage against the Age base model. **Sign error analysis:** Directional accuracy where sign error equaled bootstrap estimate with opposite sign of severity-specific RECOVERY point estimate, ranking models by total sign errors across 1,500 iterations, with two-way binomial tests generated using coverage analysis procedures. **Software:** Statistical analysis used Python 3.13.3, Bayesian hierarchical regression (bambi v0.15.0, pymc v5.22.0), two-way binomial tests (scipy v1.15.2) were adjusted for multiple comparisons (n=7) with visualization via Matplotlib/Seaborn

4. Results

We evaluated causal diagrams, constructed by human participants and transformer-based language models, as transparent and interpretable bridges between domain expertise and quantitative clinical reasoning. Human participants from the Midwest Healthcare Conference Challenge (n=3 teams) and four transformer models (OpenAI o3, DeepSeek R1, Gemini Pro 2.5, Claude Opus) independently constructed causal diagrams to estimate COVID-19 glucocorticoid treatment effects on 28-day mortality using real-world data from more than 2,000 hospitalized patients. All models employed target trial emulation methodology with similar data access and challenge specifications. Performance was assessed through bootstrap coverage analysis against published RECOVERY trial benchmarks (n=500 iterations per disease severity stratum), sign error analysis for directional accuracy, and expert review of the causal plausibility of each directed edge.

Table 1: Performance comparison of transformer and human models showing causal plausibility scores, bootstrap outlier counts, and coverage percentages across disease severity strata.

Model	Type	Rank	Nodes	Edges	Adj. Set Final	Plaus. Score	Total Outliers	Coverage (%)
o3	LLM	1	15	35	13	2.25(0.68)	563	62.4
Team 1	Human	2	15	20	5	2.38(0.58)	576	61.6
GeminiPro2.5	LLM	3	13	22	4	2.59(0.36)	580	61.3
Age	Base	4	3	3	1	3	604	59.7
Team 2	Human	5	13	16	5	2.42(0.40)	610	59.3
Team 3	Human	6	6	5	4	2.27(0.86)	612	59.2
DeepSeekR1	LLM	7	5	7	2	2.52(0.42)	612	59.2
ClaudeOpus	LLM	8	19	22	9	2.29(0.74)	640	57.3

Legend: Adj. Set Final = Adjustment Set Final, Plaus. Score = Average Causal Plausibility Score (standard deviation)

4.1. *Comparative Performance of Human and Transformer Models*

4.1.1. *Model Complexity*

Structural complexity varied across human and transformer-generated models, as indicated by node count (5–19), edge count (5–35), and adjustment set node count (2–13) (**Table 1**). A key difference was that the human teams iteratively refined their graph topology and adjustment status to ensure statistical convergence. In contrast, Human-in-the-Loop (HITL) modifications were limited to changing adjustment status and enforcing the one-feature-per-node requirement, with no edges being added or removed. Specifically, the Claude Opus model required: 1) removing one feature from two nodes (ferritin_log2 was removed from Inflammatory Markers leaving CRP_log2, and Asthma was removed from Chronic lung disease node leaving COPD), and 2) unadjusting CAD, Chronic lung disease, Cancer Flag, and Sex. The OpenAI o3 model required unadjusting the CAD and Sex nodes (**Figure 3**), and the DeepSeek R1 model required unadjusting the CRP_log2 node to remove a mediator. Notably, the GeminiPro 2.5 model required no HITL modifications **(Supplementary Materials)**.

4.1.2. *Causal Plausibility*

To evaluate the quality of the causal relationships posited by each structural causal model (SCM), we calculated an average Plausibility Score (0 = implausible to 3 = strong biological basis and clinical evidence) for all directed edges across human-generated (n = 36) and LLM-generated (n = 86) models. These scores were derived from a blinded, expert review by three clinical domain experts, who assessed each edge for biological and clinical plausibility using the 0-3 ordinal scale, **(Table 1)**. Plausibility Scores were nearly identical between Transformer models and human teams, with score distributions differing by 5% at each level (0: 4% vs 6%; 1: 12% vs 12%; 2: 27% vs 22%; 3: 57% vs 60%) and statistically equivalent means (2.37±0.62 vs 2.38±0.56; Cohen's D=0.02). Overall, both human and LLM-derived models demonstrated a similar capacity to construct relationships deemed plausible by a domain expert.

4.1.3. *Sign Error Analysis and Directional Accuracy*

We performed a sign error analysis to assess the directional accuracy of treatment effect estimates compared to RECOVERY trial benchmarks (**Figure 4, Top and Middle Panels**). A sign error was defined as a bootstrap point estimate having the opposite sign of the corresponding severity-specific RECOVERY trial estimate.

Sign error measurement revealed substantial performance variation. Across all severity levels ($n = 1,500$ bootstrap iterations), the two best-performing models were Transformers: Claude Opus (11.1% error rate) and o3 (11.2% error rate), both significantly outperforming the baseline ($p = 2 \times 10^{-8}$). Transformer-generated models generally showed superior directional accu-

racy compared to human models, particularly in the critical disease stratum.

Directional accuracy varied sharply by severity. In mild-moderate disease (Severity 1), o3 achieved the lowest error rate at 9.0%. Severe disease (Severity 2) showed exceptional accuracy, with all models exhibiting minimal error rates (0% – 0.6%). Conversely, critical disease (Severity 3) had the poorest directional accuracy, with all models exceeding a 19% error rate. Performance in this stratum ranged from 19.6% (Claude Opus) to 46.6% (Team 3), suggesting challenges likely due to sample size limitations, unmeasured confounding, and/or true differences in the causal effect of glucocorticoids in this subpopulation.

4.1.4. *Coverage Analysis*

Table 1 presents the comparative performance of structural causal models submitted by the three top-ranked participating teams (Team 1, Team 2, and Team 3), four transformer-based large language models (o3, Gemini Pro 2.5, DeepSeek R1, Claude Opus) one year after the conclusion of the challenge, and a simple base model (Age). The best-performing model overall was generated by OpenAI's o3 transformer, achieving rank 1 with 563 total bootstrap outliers corresponding to 62.4% coverage of RECOVERY trial confidence intervals (unadjusted two-sided binomial p-value = 0.032 compared to Age base model, n=1500) . The second-ranked model was human-generated, achieving similar performance with 576 outliers (61.6% coverage). It is also noteworthy, that the top-performing human team utilized a literature-based approach with a parsimonious adjustment set (n=5 adjusted nodes), whereas the top-performing transformer model generated the most complex causal graph (n=35 edges) with the largest adjustment set (n=13 adjusted nodes).

When comparing coverage performance against the Base model (Age as the only adjustment node), statistical significance varied by severity stratum (**Figure 4, Bottom Panel**). In mild-moderate disease (Severity 1, n=500), only the o3 model outperformed the Base model (p-value = 0.013). For severe disease (Severity 2, n=500), three models showed superior performance: Team 1, o3, and Gemini Pro 2.5 (all p-values < 0.0005), while Claude Opus performed significantly worse than the Base model (p-value < 0.001).

A critical finding was the universal poor performance in the critical disease severity stratum (Severity 3), where all models achieved 0% coverage (500/500 bootstrap outliers). This suggests fundamental challenges in modeling treatment effects with this dataset using the gformula, potentially due to limited sample size and/or unmeasured confounding in this subgroup. In contrast, both mild-moderate and severe disease strata showed reasonable coverage performance, with the best models achieving >97% coverage for severity 1 and 90% coverage for severity 2 disease.

This provides initial evidence that state-of-the-art AI systems can construct models with performance similar to human domain experts without formal causal inference training, though both approaches face limitations in complex clinical scenarios.

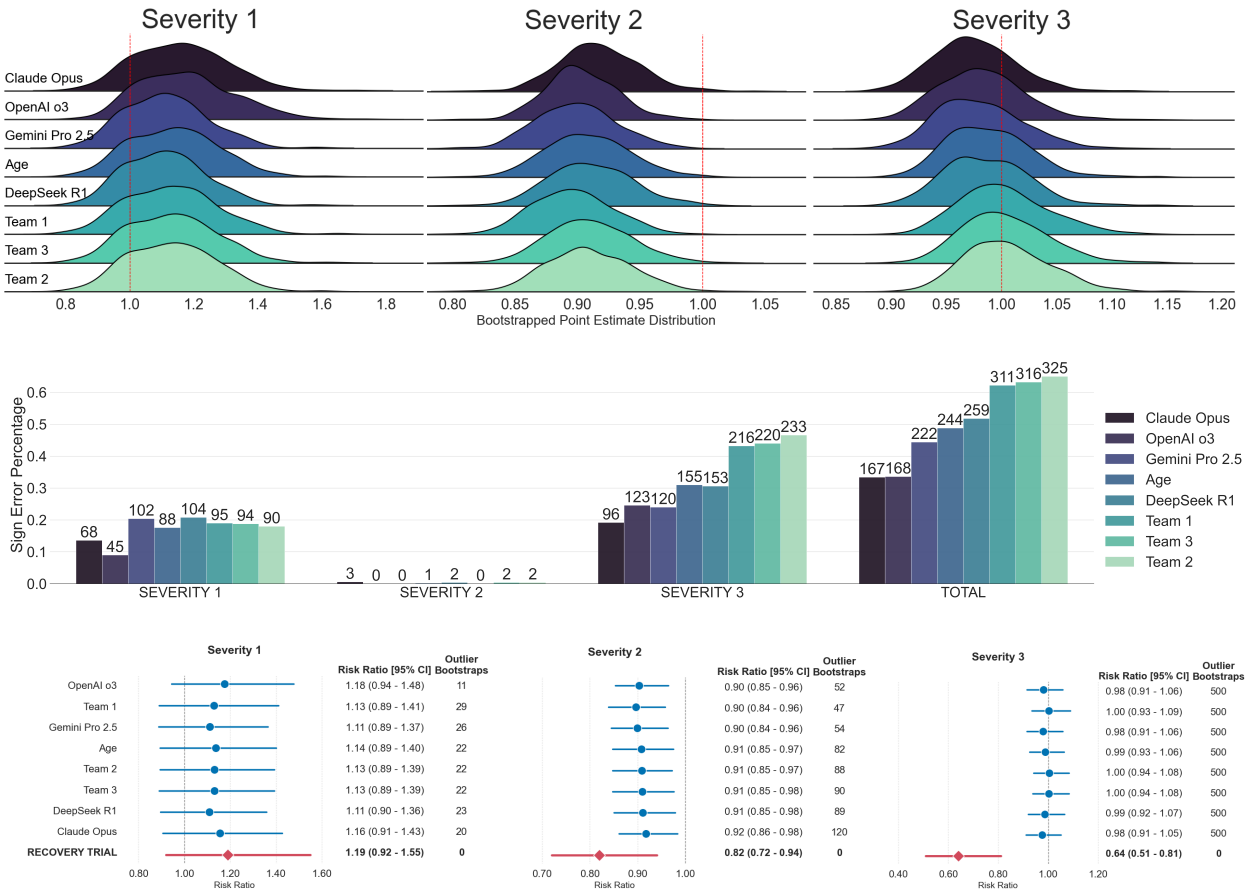


Fig. 4. **Top panel:** Bootstrapped Risk Ratio (RR) distributions across severity levels with ridge plots showing probability densities for each SCM. Red dashed line (RR = 1.0) indicates no effect; distributions right/left suggest increased/decreased 28-day mortality risk. Models ranked by total bootstrap sign errors. **Middle panel:** Sign error distribution by model and severity. Sign errors defined as bootstrap estimates with opposite sign from corresponding RECOVERY trial point estimates (500 iterations per severity, 1,500 total per model). **Bottom panel:** Bootstrap CI coverage versus RECOVERY benchmarks. Forest plot comparing coverage performance between human participants (blue) and AI models (red) across severity strata. Points show proportion of 500 bootstrap estimates within published RECOVERY 95% CIs. Models ranked by total outliers (estimates outside RECOVERY 95% CIs); 500 samples per stratum, 1,500 total. Dashed line indicates RR = 1.

4.1.5. *SCM Ablation using Potential Outcomes*

The SCM ablation results show structural causal models consistently outperformed traditional potential outcomes implementations across all tested models. Traditional PO approaches yielded poor coverage rates of 1-30%, 2-37%, and 1-10% across severities 1-3, with DeepSeekV1 failing to generate any causal estimate despite repeated debugging attempts. The wide PO confidence intervals occasionally produced better sign error rates—notably Claude PO's 0.8% versus SCM's 13.6% in Severity 1—suggesting imprecise estimates sometimes captured correct directionality by chance rather than proper calibration. The PO approaches exhibited

systematic failures regardless of sophistication level, from basic IPTW to advanced cross-fitted ensemble methods, all mishandling the person-time data structure, treatment eligibility, and immortal time bias (see Supplementary Materials). These findings suggest causal graphs provide essential scaffolding for reliable inference in complex longitudinal settings.

Table 2: Performance comparison between Structural Causal Models (SCM) and Potential Outcomes causal inference.

Model	Method	Severity 1	Severity 2	Severity 3
		Inside CI% / Correct Direction%		
OpenAI o3	SCM	97.8% / 91.0%	89.6% / 100.0%	0.0% / 80.8%
	PO	0.8% / 0.2%	1.8% / 63.4%	10.2% / 60.0%
	Δ	**+97.0pp / +90.8pp**	**+87.8pp / +36.6pp**	**-10.2pp / +20.8pp**
Claude	SCM	96.0% / 86.4%	76.0% / 99.4%	0.0% / 66.6%
	PO	19.6% / 99.2%	36.6% / 63.8%	0.6% / 7.4%
	Δ	**+76.4pp / -12.8pp**	**+39.4pp / +35.6pp**	**-0.6pp / +59.2pp**
Gemini 2.5	SCM	94.8% / 82.4%	89.2% / 100.0%	0.0% / 55.6%
	PO	29.6% / 88.6%	10.2% / 100.0%	9.8% / 23.2%
	Δ	**+65.2pp / -6.2pp**	**+79.0pp / 0.0pp**	**-9.8pp / +32.4pp**

RECOVERY Benchmarks: Severity 1: RR=1.19 [0.92, 1.55] (harm) — Severity 2: RR=0.82 [0.72, 0.94] (benefit) — Severity 3: RR=0.64 [0.51, 0.81] (benefit).
Legend: SCM = Structural Causal Model, PO = Potential Outcomes, Δ = Improvement, pp = percentage points

5. Discussion

5.1. *Causal Reasoning Comparative Performance Analysis*

Our evaluation of state-of-the-art transformer models (OpenAI o3, DeepSeek R1, Gemini Pro 2.5, Claude Opus) against human participants from the Midwest Healthcare Conference Challenge revealed nuanced performance differences on a complex causal reasoning task (Target Trial Emulation). All participants successfully constructed causal models and implemented analytical pipelines, though with varying degrees of sophistication in handling the person-time data structure and time-varying confounding inherent in the COVID-19 treatment effect estimation task.

With respect to directional accuracy, transformer-generated models significantly lower rates of sign errors compared to the baseline model. Even in the critical disease stratum (Severity 3) where coverage proved challenging, transformer models maintained better directional consensus. Sign errors are particularly clinically relevant when population health management decisions must be made with limited quality evidence, as was the case during the early days of the Covid-19 pandemic. This superior directional performance suggests transformer models

may better capture underlying biological plausibility even when statistical precision remains elusive.

Coverage analysis revealed a complex performance landscape across severity strata. While the top-performing transformer model (o3) achieved similar overall coverage to the best human team (62.4% vs 61.6% bootstrap coverage against RECOVERY trial benchmarks), performance varied substantially by disease severity. In mild-moderate disease (Severity 1), only o3 significantly outperformed the age baseline ($p = 0.013$), while human teams showed no significant improvement. For severe disease (Severity 2), three models—Team 1, o3, and Gemini Pro 2.5—demonstrated highly significant superior performance (all $p < 0.0005$), while Claude Opus performed significantly worse than baseline ($p < 0.001$). Universal poor performance in the critical disease severity stratum (0% coverage across all models) suggests fundamental challenges in modeling this particular treatment effect that transcend both human and AI approaches, likely reflecting dataset limitations rather than methodological deficiencies. Specifically, we found that sparse confounder patterns paired with missingness observed in this real-world dataset posed convergence challenges and statistical modeling limitations for the participants. While this represents a common impediment to causal inference applications in practice, a larger and more complex dataset may have revealed greater distinctions across human and AI performance.

The SCM ablation experiments suggests that framework choice influences transformer performance on complex causal tasks. SCMs explicitly encode time-oriented causal relationships in their graph structure, while potential outcomes frameworks require practitioners to separately specify these elements when implementing estimators like pooled logistic regression or marginal structural models. The consistent struggles of PO implementations across all transformer models—particularly with person-time data and immortal time bias—hint that translating abstract potential outcomes notation into concrete analytical code demands additional implicit reasoning that current AI systems find challenging. This pattern supports our broader finding that SCMs facilitate human-AI collaboration by providing explicit, visual representations of causal assumptions that both parties can evaluate, rather than leaving critical temporal dynamics implicit in statistical code.

5.2. *Generative AI to Causal Graph Interface*

We developed a novel interface that translates generative AI medical knowledge into regulatory-aligned structural causal models,[14] addressing fundamental limitations in current clinical AI systems. By combining the knowledge generation capabilities of large language models with the transparency and rigor required for target trial emulation, quality improvement monitoring, and clinical decision support, this approach provides an interpretable workflow for healthcare causal inference tasks. The cStructure platform integrates collaborative visual interface design, automated d-separation evaluation, and privacy-preserving statistical analysis to enable rapid systematic validation of AI-generated causal assumptions against established clinical knowledge.

5.3. *Transformer-Based AI for Clinical Causal Reasoning*

While transformer-based generative AI provides rapid access to vast medical domain knowledge with improving accuracy, their probabilistic nature and inscrutable generation processes necessitate structured frameworks that enable human oversight and validation in safety-critical applications. Our results demonstrate that structural causal models can bridge AI capabilities and human expertise by exposing interpretable causal reasoning that can be evaluated, modified, and validated using real-world data. The improved directional accuracy of transformer models suggests complementary strengths to human-generated models, while also serving as accelerators for human-led causal reasoning—particularly for non-experts—when SCMs provide the structured conduit. Practical implementation required extracting LLM domain knowledge into visually accessible representations with explanatory text that enable rapid human-in-the-loop (HITL) evaluation and debugging. This transparent design allowed clinical experts to identify potential biases, verify confounder adjustment, and modify assumptions based on domain knowledge. While transformer models required HITL modifications for statistical convergence—primarily unadjusting nodes causing convergence failures—we deliberately restricted interventions to adjustment status only, preserving AI-generated topology. This constraint simulates realistic rapid review scenarios, contrasting with human teams who iteratively revised both topology and adjustment sets, yet provides a more realistic assessment of how practitioners might use AI-generated causal models in time-constrained clinical settings.

5.4. *Future Directions*

Future research should expand evaluation to diverse causal reasoning tasks across multiple clinical domains, therapeutic areas, and patient populations to establish the generalizability and utility of human-AI equivalence in causal model construction. Additional priorities include developing standardized evaluation frameworks for clinical causal reasoning, investigating hybrid human-AI collaborative approaches that leverage complementary strengths (e.g. Covid Causal Diagram DREAM Challenge), and exploring real-time clinical deployment of AI-generated causal models with appropriate safety monitoring and validation protocols. The framework's potential for regulatory applications warrants investigation through partnerships with health agencies to establish guidelines for AI-assisted causal inference in evidence generation.

Supplement: Supplemental methods, materials, tables, and figures are available at https://github.com/cstructure/interpretable-clinical-scm

Acknowledgements: We would like to thank the participants of the Midwest Healthcare Conference Challenge for their contributions. With a special thanks to: April Wu, Yinglei Wu, Wendy Zheng, Brian Hong, July Chen, Brian Ellis, and Rebecca Shi.

References

1. Gary S Collins, Karel GM Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, et al. Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385, 2024.
2. Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.
3. Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.
4. U.S. Food and Drug Administration and Health Canada and MHRA. Transparency for machine learning-enabled medical devices: Guiding principles. `https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles`, June 13 2024. Joint publication by FDA, Health Canada MHRA promoting transparency principles for ML medical devices.
5. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
6. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
7. Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.
8. Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 33(8):1–79, 2024.
9. Lauren E Dang, Susan Gruber, Hana Lee, Issa J Dahabreh, Elizabeth A Stuart, Brian D Williamson, Richard Wyss, Iván Díaz, Debashis Ghosh, Emre Kıcıman, et al. A causal roadmap for generating high-quality real-world evidence. *Journal of Clinical and Translational Science*, 7(1):e212, 2023.
10. Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.
11. Judea Pearl. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450):428–431, 2000.
12. Judea Pearl and Thomas Verma. The logic of representing dependencies by directed graphs. In *Proceedings of the sixth National conference on Artificial intelligence-Volume 1*, pages 374–379, 1987.
13. MA Hernán and J Robins. Causal inference: What if. boca raton: Chapman & hill/crc.(2020). *Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations*, 2020.
14. U.S. Food and Drug Administration. Real-world evidence: Considerations regarding non-interventional studies for drug and biological products. Draft guidance for industry, U.S. Food and Drug Administration, March 2024. Guidance document, draft issued Mar.2024; docket no. FDA-2023-D-5470.
15. Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A

Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 2025.
16. Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040, 2024.
17. Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642, 2025.
18. Jeevan Tewari, Benjamin W. Dahl, and Jeffrey J. Saucerman. Benchmarking of signaling networks generated by large language models. *bioRxiv*, 2025.
19. Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.
20. RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with covid-19. *New England journal of medicine*, 384(8):693–704, 2021.
21. Sanchita Bhattacharya, Patrick Dunn, Cristel G Thomas, Barry Smith, Henry Schaefer, Jieming Chen, Zicheng Hu, Kelly A Zalocusky, Ravi D Shankar, Shai S Shen-Orr, et al. Immport, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific data*, 5(1):1–9, 2018.
22. Diane Marie Del Valle, Seunghee Kim-Schulze, Hsin-Hui Huang, Noam D Beckmann, Sharon Nirenberg, Bo Wang, Yonit Lavin, Talia H Swartz, Deepu Madduri, Aryeh Stock, et al. An inflammatory cytokine signature predicts covid-19 severity and survival. *Nature medicine*, 26(10):1636–1643, 2020.
23. Miguel A Hernán, Brian C Sauer, Sonia Hernández-Díaz, Robert Platt, and Ian Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of clinical epidemiology*, 79:70–75, 2016.
24. Miguel A Hernán, Jonathan AC Sterne, Julian PT Higgins, Ian Shrier, and Sonia Hernández-Díaz. A structural description of biases that generate immortal time. *Epidemiology*, 36(1):107–114, 2025.