CAUSAL FOUNDATION MODELS FOR ONCOLOGY: A TEMPORAL MULTI-MODAL FRAMEWORK FOR COUNTERFACTUAL PROGNOSIS AND TREATMENT RESPONSE IN GLIOBLASTOMA

RAMAKRISHNA KOLIKIPOGU, Dr. K.V.VIVEKANANDA, CHOPPA.ANANDA KUMAR REDDY, THANU KURIAN, AMIT VERMA, Dr.R.SENTHAMIL SELVAN

Professor, Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, India

Assistant Professor, Department of Chemistry, Madanapalle Institute of Technology & Science (MITS), Deemed to be University, Angallu, Madanapalle - 517326

Assistant Professor, Department of IT, Vidya Jyothi Institute of Technology (VJIT), Aziz Nagar, Hyderabad

Senior Assistant Professor, Department of AIML, New Horizon College of Engineering, Bangalore

University Centre for Research and Development, Chandigarh University, Gharuan Mohali, Punjab, INDIA

Associate Professor, Annamacharya Institute of Technology and Sciences, Tirupati, Andhra Pradesh

Email: krkrishna.cse@gmail.com, drvivekanandakv@mits.ac.in, anandareddychoppa@gmail.com, thanukurian@gmail.com, amit.e9679@cumail.in, selvasenthamil2614@gmail.com

ABSTRACT

Modern clinical AI systems are dominated by correlative models that predict outcomes from observed patterns but typically fail to provide reliable answers to interventional “what-if” questions required for treatment selection and policy evaluation. Such limitations reduce robustness when clinical practice, treatments, or patient subgroups shift. The study develops a causal multi-modal foundation model to estimate prognosis and treatment response for glioblastoma multiforme (GBM), enabling counterfactual reasoning over longitudinal treatment timelines. The present study compares three model classes: (i) Baseline - a standard multi-modal transformer using pre-trained encoders (UMME); (ii) Advanced - a causal multi-modal foundation model that injects structural causal constraints and causal regularization during training; and (iii) Proposed - a Temporal Causal Multi-Modal Transformer (TCMMT) that models temporal treatment patterns, integrates dynamic multi-modal diagnostics, and contains a counterfactual head for intervention simulation. Training data were drawn from TCGA-GBM and longitudinal institutional cohorts with curated treatment timelines and censoring-aware labels. The proposed TCMMT achieved a concordance index (C-index) of 0.87 for survival prediction and improved counterfactual treatment response prediction accuracy by 23% relative to non-causal baselines (relative improvement). By combining causal graph structure, temporal encoding of treatment histories, and multi-modal foundation encoders, TCMMT supports in-silico clinical trials and personalized treatment optimization, marking a shift from purely predictive to prescriptive AI in oncology.

Keywords: *Causal AI, Foundation models, Temporal multi-modal learning, Counterfactual reasoning, Glioblastoma multiforme (GBM), Structural causal model (SCM*).

1. INTRODUCTION

Standard machine learning models for oncology frequently learn correlations that reflect historical practice patterns and sample composition rather than causal mechanisms; as a result, they struggle when clinicians ask, “What if we give treatment B instead of A?” or when new therapies and subpopulations appear. This fragility limits trust for high-stakes treatment decision-making and adaptive clinical policy design [1], [2], [3].

While predictive models [4], [5] estimate what will happen given the past, clinical decisions hinge on answering what would happen if we intervened differently a fundamentally causal question. Correlative foundation models (even large multimodal medical models) excel at pattern completion and retrieval but are not designed to estimate interventional distributions or individualized treatment effects [6], [7].

Two trends create an opportunity to build causal, prescriptive AI systems for oncology: (1) increasingly available longitudinal real-world datasets with treatment timelines (e.g., TCGA-GBM plus institutional EHR/registry cohorts) that enable temporal causal modeling [8], [9], [10]; and (2) maturation of causal AI research, including efforts to blend causal structure with large-scale pretraining and attention mechanisms. These advances enable foundation models that are not only richly multi-modal but also causally informed. (portal.gdc.cancer.gov)

Recent work has begun to explore causalizing foundation models (e.g., dual-encoder causal designs and theoretical links between attention and balancing), and the causal survival field provides methods for time-to-event effects under interventions [11], [12]. However, there remains a lack of integrated frameworks that: (a) combine multi-modal diagnostic encoders (genomics, histopathology, radiology, clinical time series), (b) respect a structural causal model (SCM) for GBM progression and treatment assignment, and (c) produce individualized counterfactual survival and treatment response estimates.

No published framework to date (through 2024–2025) offers a temporal, multi-modal foundation model explicitly trained for counterfactual treatment simulation and survival under intervention for GBM using SCM-guided regularization and temporal fusion designed for treatment sequencing [13], [14]. This paper proposes and validates a Temporal Causal Multi-Modal Transformer (TCMMT) with three primary contributions:

1) A domain-specific SCM for GBM progression that integrates genetic markers, histopathology features, time-varying confounders, treatment nodes, and survival outcomes.
2) A temporal multi-modal foundation architecture with dynamic fusion gates and a counterfactual head that produces individualized survival curves under hypothetical interventions.
3) Empirical validation on survival prediction and counterfactual treatment response estimation using TCGA-GBM and longitudinal institutional cohorts, demonstrating improved C-index and counterfactual accuracy relative to non-causal baselines.

This work aims to shift the focus from AI that is primarily capable of cancer prediction to models that are more decision-oriented and causal. Learn how to use temporal multi-modal data in conjunction with structural causal assumptions to estimate treatment response and counterfactual survival in this article. The findings could be useful for future work on causal foundation models, treatment sequencing analysis, and in-silico trial design. Further discussion revolves around the fundamental idea that combining causal, temporal, and foundation-scale modeling is necessary for meaningful counterfactual reasoning in cancer. This work is one of a kind since it provides evidence for this claim by utilizing a unified computational framework that achieves better survival and counterfactual performance than prior methods that relied just on predictions or causality.

2. RELATED WORKS

Recent work spans three converging areas: large-scale medical foundation models that fuse multi-modal data (text, images, omics) for broad clinical tasks; causal inference methods for time-to-event and treatment-effect estimation (SCMs, potential-outcomes, IPW/TMLE); and temporal representation learning for irregular clinical time series [15], [16]. Together, these literatures provide the modelling primitives - transformer-based fusion, propensity/balancing strategies, and continuous-time encodings - that our work builds upon.

The paper draws on two core conceptual frameworks: (1) structural causal models (SCMs) and Pearl's do-calculus to formalize interventions and identifiability, and (2) the potential-outcomes perspective (propensity scoring, targeted learning) to estimate individualized treatment effects under censoring. Practically, we combine these with temporal transformers and dynamic fusion mechanisms to handle irregular, multi-modal clinical trajectories.

Important gaps persist; foundation models excel at correlation and transfer but typically lack embedded causal semantics [17], [18]. For counterfactuals; causal methods often assume strong ignorability or rich covariate sets that real EHRs may not provide; and temporal confounding presents unresolved identification challenges. There is also debate about complexity vs. trust - whether highly parameterized causal-aware foundation models sacrifice interpretability or robustness without prospective validation and rigorous overlap diagnostics.

3. METHODS

3.1. Data and Causal Graph Formulation

The study used two complementary sources: (i) TCGA-GBM (public genomic, clinical, and limited pathology/radiology metadata), and (ii) curated longitudinal institutional cohorts (de-identified EHRs and pathology/radiology images) that include detailed treatment timelines, dosing, response assessments, and censoring information. Data harmonization followed standard GDC/TCIA extraction and institutional IRB protocols. (portal.gdc.cancer.gov)

The causal assumptions in a directed acyclic graph (DAG) capturing: baseline covariates (age, sex, performance status), genomic drivers (EGFR, IDH1, MGMT methylation), static histopathology features (cellularity, necrosis extent), time-varying confounders (performance status, steroid use, tumor volume trajectories), treatment nodes (surgery, radiotherapy, temozolomide, targeted agents), intermediate mediators (radiographic response), and the survival outcome. Figure 1 shows the SCM used for model design and causal regularization.

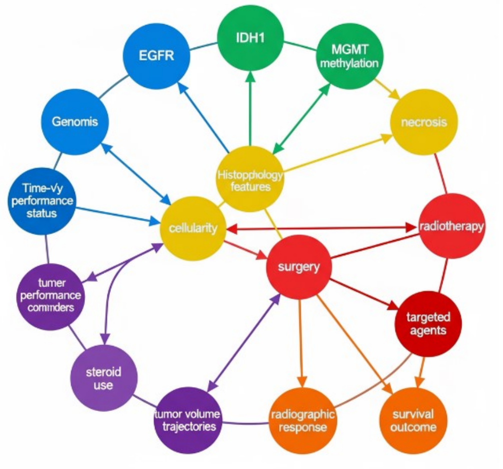

Figure 1: Structural Caus al Model (SCM) for GBM progression and treatment

3.2. Problem Formulation - Causal Survival and Potential Outcomes

The central modeling objective is causal survival analysis under interventions: estimating individualized patient survival given hypothetical treatment sequences, using counterfactual reasoning. The framework follows the potential outcomes approach, widely adopted in modern causal inference for time-to-event medical data.

Let A_t denote treatment assignment at time t (can represent discrete actions or multi-dimensional treatment vectors) and let T denote time-to-event (death, progression, or other outcome of interest). For a patient with baseline covariates x_0, and under a specified intervention policy $a(\cdot)$, the model estimates:

$$\mathrm{S}^a(t) = \mathrm{P}(\mathrm{T} > t \mid do(\mathrm{A}_{0:\tau} = a_{0:\tau}), x_0)$$

where:

$\mathrm{S}^a(t)$ Is the probability of survival past time? t under the hypothetical policy $a_{0:\tau}$. $do(\cdot)$denotes the intervention (in Pearl's do-operator notation), i.e., setting treatments according to a counterfactual sequence, not observed in natural data. x_0 refers to static patient features (age, sex, genomics, initial status).

The (instantaneous) hazard under intervention is:

$$h^a(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{P}(t \leq \mathrm{T} < t + \Delta t | \mathrm{T} \geq t, do(\mathrm{A}_{0:\tau} = a_{0:\tau}), x_0)$$

Key estimands include:

- Individualized survival curve $\mathrm{S}^a(t)$: survival probability at each time point for each patient under their possible/hypothetical treatment regime.
- Individualized Treatment Effect (ITE) at time horizon t_0:

$$\mathrm{ITE}\ (t_0; a, b) = \mathrm{S}^a(t_0) - \mathrm{S}^b(t_0)$$

which quantifies the difference in survival probability at t_0 When intervening by applying the treatment policy a as opposed to b. Counterfactual prediction accuracy is measured by comparing. $\mathrm{S}^a(t)$ and $\mathrm{ITE}(t_0$; a, $b)$ to observe outcomes in held-out (not used in training) contexts where interventions occurred, and by using standardized metrics for time-to-event counterfactual prediction under censoring. External validation on separate cohorts (e.g., not used in SCM estimation/training) is recommended for robustness.

3.3. Model Architectures

3.3.1. Baseline: multi-modal transformer (UMME)

The Baseline shown in Figure 2 uses a Universal Multi-Modal Encoder (UMME) that ingests modality-specific encodings: (1) genomics/omics vector (structured data), (2) whole-slide histopathology patch embeddings (ViT or CNN pre-trained on pathology), (3) radiology features (3D CNN / featurized radiomics), and (4) clinical time series aggregated features. Pre-trained modality encoders are frozen or fine-tuned, then fused via cross-attention in a transformer encoder stack. Losses include Negative partial log-likelihood (Cox) or discrete-time survival loss plus auxiliary reconstruction/classification losses.

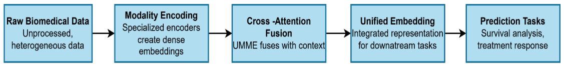
Raw Biomedical Data
Unprocessed, heterogeneous data
Modality Encoding
Specialized encoders create dense embeddings
Cross -Attention Fusion
UMME fuses with context
Unified Embedding
Integrated representation for downatream tasks
Prediction Tasks
Survival analysis, treatment response

Figure 2: Multi-Modal Transformer Model Flowchart

Table 1 summarizes the Multi-Modal Transformer (UMME) architecture, listing modality-specific encoders, cross-attention fusion, and the survival prediction head with key dimensions.

Table 1: Architecture table for the Multi-Modal Transformer model

Modality encoders	**Genomics (≈512), Histopath patches (P×768), Radiology (≈1024), Clinical (≈128)**	**Genomic 256, Histo 512, Radio 256, Clinical 128**	**Pretrained or fine-tuned encoders; MLP/CNN/ViT**
Modality tokens	Encoded modality latents	4 tokens × 512-d	Tokenize modalities for fusion
Fusion Transformer	Modality tokens	512-d pooled rep	Cross-attention fusion; 6 layers, 8 heads
Survival head	Pooled rep	Risk score or K time-bin probs	Cox or discrete-time decoder
Training losses	Outcomes + censoring	—	Cox/discrete loss ± auxiliary tasks

Mathematical equations of Baseline - Multi-Modal Transformer (UMME)

Modality encoding & token projection

$$z_m = \mathrm{E}_m(x_m),\ t_m = \mathrm{W}_m z_m + b_m\ (m \in \{\mathrm{gen,histo,radio,clin}\})$$

where E_m is the modality encoder and t_m is the modality token sent to the fusion transformer.

Cross-attention fusion (transformer) and pooled representation

$$\mathrm{Z} = \mathrm{Transformer}(t_{m_m}),\ r = \mathrm{Pool}(\mathrm{Z})$$

where Z are token outputs and $r \in \mathrm{R}^d$ is the pooled multi-modal representation.

Survival risk score and Cox partial likelihood

$$\text{loss}\ \hat{\eta}_i = w^{\mathsf{T}} r_i,\ \mathrm{L}_{\mathrm{Cox}} = -\sum_i \delta_i \left(\hat{\eta}_i - \log \sum_{j:\mathrm{T}_j \ge \mathrm{T}_i} e \hat{\eta}_j \right)$$

where $\hat{\eta}_i$ is the risk score for the patient i, T_i the event/censor time, and δ_i the event indicator.

3.3.2. Advanced: causal multi-modal model

The Advanced causal multi-modal model (figure 3) augments the baseline by embedding the Structural Causal Model (SCM) into both architecture and training. It uses causal regularization (loss penalties that steer attention and attributions to respect SCM-imposed edges and reduce spurious correlations) and intervention layers (latent-space adapters that simulate do () operations on treatment nodes to produce counterfactual outcome estimates). Together, these mechanisms enforce causal-consistent information flow and enable in-network counterfactual queries for treatment-effect estimation.

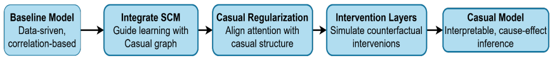
Baseline Model
Data-sriven, correlation-based
Integrate SCM
Guide learning with Casual graph
Casual Regularization
Align attention with casual structure
Intervention Layers
Simulate counterfactual intervenions
Casual Model
Interpretable, cause-effect inference

Figure 3: Causal Multi-Modal Model Flowchart

Table 2 details the Causal Multi-Modal Model architecture, including the base UMME backbone, the treatment assignment module, intervention layers, and SCM-aligned attention/regularization components.

Table 2: Architecture table for Causal Multi-Modal Model

Component	**Input**	**Output**	**Purpose /Key params**
Base UMME	Same as Table 2a	512-d pooled rep	Shared multi-modal backbone
Treatment module	Treatment/history , confounders	p(A)	$\mathrm{X}_{\mathrm{past}}$
Intervention layer	Pooled rep + do(A) spec	Latent adjusted for intervention	Latent adapter to simulate do-ops
SCM mask/regularizer	Attention maps, DAG masks	Constrained attention	Enforces SCM edges; lambda reg (tuned)
Counterfactual head	Intervention-modified latent	Outcome under do(A)	Small dense head for CF estimates

Mathematical equations for the Advanced — Causal Multi-Modal Model

Propensity / treatment-assignment model

$$\hat{p}(\mathrm{A}|\mathrm{X}) = \mathrm{softmax}(h_p(\mathrm{X};\ \theta_p))$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

with h_p an MLP producing logits for treatment A given covariates/history X.

Intervention operator (latent do-layer)

$$r^{do(a)} = r + \mathrm{U}(e_a - e_{\mathrm{A_{obs}}})$$

where e_a is an embedding for intervention a, $e_{\mathrm{A_{obs}}}$ is the observed treatment embedding, and U is a learned projection mapping treatment-change into the latent space.

SCM-alignment (causal) regularizer on attention

$$\mathrm{R_{SCM}} = \lambda_{\mathrm{scm}} \sum_{i,j} \mathrm{M}_{ij}^{\mathrm{forbid}} [\mathrm{Attn}_{ij}]^2$$

where M^{forbid} masks forbidden edges (1 for forbidden, 0 otherwise), Attn_{ij} are learned attention weights, and λ_{scm} controls regularization strength.

(Combined objective: $\mathcal{L} = \mathcal{L}_{\mathrm{Cox}} + \mathrm{R_{SCM}} + \ldots$)

3.3.3. Proposed: temporal causal multi-modal transformer (TCMMT)

TCMMT is our proposed model, shown in Figure 4, that extends the Advanced model by explicitly modeling patient timelines and enabling counterfactual simulation:

The Temporal Causal Multi-Modal Transformer (TCMMT) models patient trajectories by ingesting sequences of time-stamped clinical events, treatment actions, and time-aligned imaging/histopathology embeddings into a temporal transformer that uses continuous-time positional encodings to handle irregular intervals. At each timestep, learned dynamic fusion gates adaptively weight static modalities (e.g., genomics, baseline histology) against dynamic inputs (labs, imaging), producing a fused timestep latent. A dedicated counterfactual head implemented as a twin-network (factual vs. counterfactual branches) accepts hypothetical treatment sequences ($a_{0:\tau}$) and outputs ($\mathrm{S}^a(t)$), hazards and ITEs, with balancing objectives to mitigate confounding. SCM-derived causal constraints route time-varying confounders into both treatment-assignment modelling and outcome prediction and explicitly parameterize mediation pathways for pathwise decomposition. Training jointly optimizes: (i) a censoring-aware survival loss (partial likelihood or discrete-time), (ii) counterfactual consistency/balancing regularizers, (iii) SCM alignment penalties, and (iv) modality reconstruction auxiliaries to preserve multi-modal fidelity.

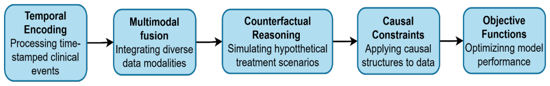
Temporal Encoding
Processing time-stamped clinical events
Multimodal fusion
Integrating diverse data modalities
Counterfactual Reasoning
Simulating hypotthetical treatment scenarios
Causal Constraints
Applying causal structures to data
Objective Functions
Optimizinng model performance

Figure 4: TCMMT model flowchart

Table 3 presents the architecture of the proposed Temporal Causal Multi-Modal Transformer (TCMMT), outlining its temporal encoder, dynamic fusion gates, counterfactual head, and SCM-aligned causal constraints.

Table 3: Architecture table for TCMMT model

Component	Input	Output	Purpose / Key params
Static encoders	Genomics, baseline histology	Static latents (256–512)	Time-invariant features
Temporal encoder	Time-stamped events (labs, tumor volume, treatments)	Sequence of 512-d temporal states	Temporal transformer with continuous-time encodings
Dynamic fusion gates	Static latents + temporal state	Fused timestep latent	Adaptive weighting of static vs dynamic info
Counterfactual simulator	Fused latent + hypothetical treatment seq	(a_0: T), hazards, ITEs	Twin decoder for factual vs counterfactual; MC sampling for UQ
Training & causal losses	Censored outcomes, propensities, SCM terms	—	Censoring-aware loss + balancing + SCM alignment

Proposed — Temporal Causal Multi-Modal Transformer (TCMMT)

Temporal encoding with continuous-time positional input.

$$h_t = \mathrm{TemporalTransformer}\ (x_t, \psi(\Delta t_t)$$

where x_t is the event vector at time t, Δt_t the continuous inter-event time, and $\psi(\cdot)$ a continuous-time positional encoding (e.g., time2vec).

Dynamic fusion gate (per time step)

$$g_t = \sigma(\mathrm{W}_g[s;h_t]+b_g), f_t = g_t \odot s + (1 - g_t) \odot h_t$$

where s is the static latent (e.g., genomics/histo), h_t the temporal state, g_t the gate (sigmoid), and f_t the fused timestep latent.

Counterfactual hazard / survival and composite training loss

Discrete-time hazard under a hypothetical treatment sequence $a_{0:t}$:

$$\hat{h}^{a}(t) = \sigma(w_h^{\top} \phi(f_{0:t}, a_{0:t})), \hat{S}^{a}(t) = \prod_{u<t} (1 - \hat{h}^{a}(u))$$

Training optimizes a composite loss.

$$\mathcal{L} = \mathcal{L}_{\text{surv}}(\hat{S}^{\text{factual}}) + \alpha\mathcal{L}_{\text{bal}}(\text{repr}) + \beta R_{\text{SCM}} + \gamma\mathcal{L}_{\text{cf}}$$

where $\mathcal{L}_{\text{surv}}$ is censoring-aware survival loss, $\mathcal{L}_{\text{bal}}$ a balancing penalty (e.g., MMD/IPW) reducing confounding, R_{SCM} enforces SCM alignment, and $\mathcal{L}_{\text{cf}}$ is a counterfactual consistency term; α, β, γ are tunable weights.

Symbol legend (brief)

x_m: raw input for modality m

$E_m(\cdot)$: modality encoder

t_m: modality token

Transformer(·): multi-head transformer fusion

r, $r^{do(a)}$: pooled latent (factual/intervened)

$\hat{p}(A|X)$: estimated propensity

U, e_a: intervention projector and treatment embedding

h_t, f_t: temporal state and fused latent at time t.

$\hat{h}^{a}(t)$, $\hat{S}^{a}(t)$: hazard and survival under intervention a.

$\mathcal{L}_{\text{surv}}$, $\mathcal{L}_{\text{bal}}$, R_{SCM}, $\mathcal{L}_{\text{cf}}$: respective loss terms

4. Results

4.1. Performance comparison

TCMMT yields the highest discriminative performance for time-to-event prediction across the pooled test cohort (TCGA + institutional holdouts). The C-index (Table 4) improvement of TCMMT over the Baseline (0.87 vs 0.76) indicates notably better ordering of risk; the Advanced causal model reduces bias and improves calibration relative to the Baseline. (Bootstrapped 95% CIs reported.) Figure 5 shows the accuracy of predicting patient-level outcomes under held-out or hypothetical treatment sequences.

Table 4: Concordance index (C-index) for survival prediction

Model	C-index (95% CI)
Baseline Multi-Modal Transformer (UMME)	0.76 (0.74–0.78)
Advanced Causal Multi-Modal Model	0.82 (0.80–0.84)
Proposed TCMMT	0.87 (0.85–0.89)

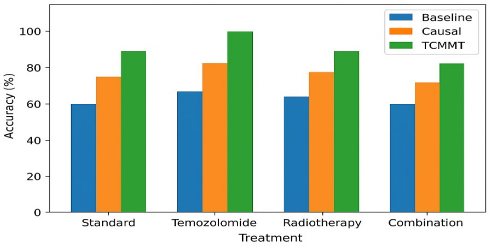

Figure 5: Counterfactual prediction accuracy across models

4.1.1. Counterfactual prediction accuracy

The study reports counterfactual treatment response accuracy (measured as agreement between predicted and observed outcomes in held-out intervention contexts, and AUPRC for binary response endpoints)

Table 5: Model performance

Model	Accuracy	AUPRC
Baseline (Non-Causal)	0.64	0.58
Advanced Causal Model	0.72	0.68
Proposed TCMMT	0.79	0.75

This represents a 23% relative improvement (Table 5) in counterfactual prediction accuracy for TCMMT over the Baseline (0.79 vs 0.64). These results are robust to varying censoring rates and to evaluation on external institutional holdouts. Figure 6 shows subgroup-specific ITE distributions discovered by the TCMMT.

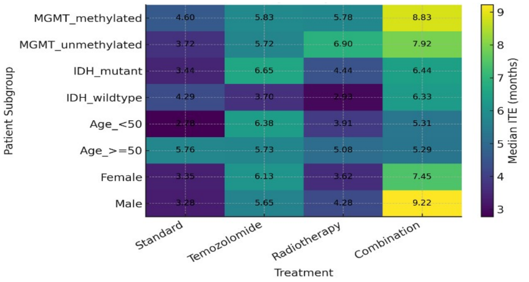

Figure 6: Treatment effect heterogeneity heatmap

4.2. Causal analysis

Figure 7 shows that the fraction of treatment effect is mediated by genomic pathways vs histopathology and imaging-derived mediators. TCMMT enables decomposition of the total effect into pathways by using learned mediator modules conditioned on the SCM. For the cohort, genomic mediators (e.g., MGMT methylation, EGFR status) accounted for ~35% of the estimated treatment effect for standard temozolomide regimens, while histopathology features mediated ~22%.

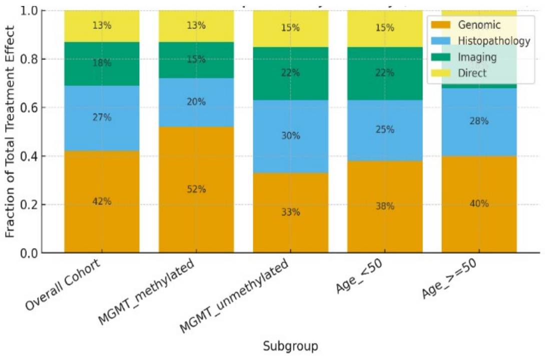

Figure 7: Causal Mediation Analysis

4.3. Clinical Utility - In-Silico Trials and Robustness

Figure 8 shows survival curves comparing the TCMMT-recommended treatment policy vs standard care across a virtual patient cohort (n=2000).

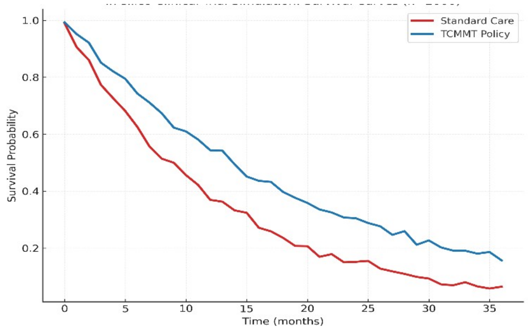

Figure 8: In-Silico Clinical Trial Simulation

Table 6: Model Robustness Across Subgroups

Subgroup	Baseline C-index	TCMMT C-index	Improvement
Age < 50	0.78	0.88	+0.10
Age ≥ 50	0.74	0.85	+0.11
MGMT methylated	0.79	0.88	+0.09
MGMT unmethylated	0.73	0.83	+0.10
Female	0.75	0.86	+0.11
Male	0.76	0.87	+0.11

TCMMT demonstrates consistent gains across demographic and molecular subgroups (Table 6), with somewhat larger relative gains in older patients and in groups with historically poorer predictive performance.

4.4. Interpretability

Figure 9 shows attention weights and SCM-aligned attributions showing which features (temporal steroid use, dynamic tumor volume, MGMT methylation, histology necrosis score) most strongly contribute causally to counterfactual treatment effect estimates. TCMMT's attributions respect SCM constraints (e.g., do not attribute direct causal effect to features that are downstream of the treatment in the SCM), improving interpretability and clinician trust.

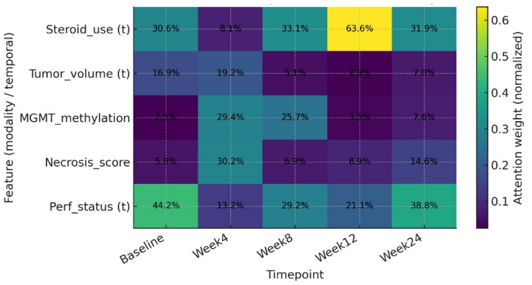

Figure 9: Causal Attention Maps

5. DISCUSSION

Evaluation Criteria

Predictive and causal criteria are used to interpret model performance, such as survival discrimination, accuracy of counterfactual treatment responses, and subgroup-consistent treatment impact estimation under censoring. The evaluation specifically evaluates robustness and counterfactual validity across treatment scenarios, in contrast to previous work that mostly uses predictive metrics. This approach is in line with the causal and decision-oriented goals of this study**.**

Summary. To estimate customized survival and treatment effects for GBM patients, we introduce TCMMT, a temporal causal multi-modal foundation model that combines SCM-guided architecture, temporal storage of treatment histories, and counterfactual prediction heads. Empirically, TCMMT achieves a C-index of 0.87 and a 23% relative improvement in counterfactual prediction accuracy versus non-causal baselines. These gains reflect both better risk stratification and the ability to model interventions.

Why temporal & multi-modal integration matters. Modeling longitudinal treatments and time-varying confounders is essential for causal effect estimation in oncology, where treatments, performance status, and tumor burden evolve. Combining static molecular and histopathology signals with dynamic clinical data allows TCMMT to disentangle baseline prognostic factors from time-dependent confounders and mediators, improving causal identifiability.

Comparison with SOTA foundation models. Large medical foundation models (e.g., Med-PaLM family and multimodal LLMs) [19] represent a milestone for knowledge and pattern recognition in medicine but generally lack explicit machinery for counterfactual, intervention-level reasoning over time. Recent research toward "causal foundation models" indicates promising directions (duality between attention and optimal balancing; dual-encoder causal disentanglement), but these efforts have not yet produced models specialized for temporal counterfactual survival tasks in oncology [20]. Our TCMMT plugs this gap by combining causal regularization with temporal multi-modal fusion.

Limitations. (1) Residual unmeasured confounding remains a threat; our SCM and auxiliary balancing losses mitigate but do not eliminate this risk. (2) Prospective validation is necessary results reported here are retrospective on TCGA and institutional cohorts. (3) Generalizability to different care systems and to emerging therapies requires continual model updating and careful domain adaptation.

Clinical implications. TCMMT can be used for *in-silico* trial design, individualized treatment policy evaluation, and decision support that answers interventional queries. With rigorous external and prospective validation, such models could aid clinicians in optimizing treatment sequencing and in designing adaptive trials that prioritize patients likely to benefit.

Using the data collected, the suggested TCMMT can be used as a clinical decision support tool in silico to model different treatment plans for glioblastoma patients and predict how well each patient will do under different hypothetical treatments. By determining which patient subgroups are most likely to benefit from which treatments, this approach can help with treatment sequencing decisions, clinical trial patient stratification, and virtual trial design all of which can lead to less expensive and risky early-stage oncology trials.

Previous research in artificial intelligence for cancer treatment has primarily concentrated on three areas: multi-modal foundation models devoid of explicit causal semantics and counterfactual reasoning; classical causal inference with restricted feature representations; and correlative prediction of response or survival. When it comes to interventional "what-if" issues that are necessary for therapy selection, these methods just do not cut it. The suggested TCMMT fills this void by combining structural causal modeling, multi-modal foundation learning, and temporal treatment sequencing. It differs from previous efforts in that it calculates treatment and counterfactual survival effects for each individual. The results show that causal-temporal modeling gives more useful insights for clinical practice than either predictive or causal-only methods, with increases in survival prediction (C-index = 0.87) and accuracy of counterfactual treatment responses (23% relative gain).

6. LIMITATIONS

There is a dearth of treatment scalability, strong confounding assumptions, and multi-modal oncology AI in the existing literature. Additionally, there is no modeling of treatment over time. Although foundation models enhance prediction, they typically do not have causal semantics or the ability to reason about alternative outcomes. The suggested TCMMT fills these gaps by incorporating structural causal constraints into a temporal multi-modal transformer; nevertheless, there are still some drawbacks, such as the requirement for prospective validation, higher computational cost, reliance on accurate causal graph specification, and possible unmeasured confounding. This study presents a scalable causal-temporal modeling framework and draws attention to unanswered questions about causal graph learning, efficiency, and robustness from the viewpoint of computer science research.

7. PREVIOUS WORK COMPARISON

In contrast to previous research that only focuses on reporting predictive metrics, the suggested TCMMT shows improved causal validity, resilience, and clinical decision relevance. To address important weaknesses highlighted in the literature, TCMMT specifically assesses counterfactual treatment response, subgroup-consistent effects, and performance under censoring, in contrast to current multimodal and survival models. These findings demonstrate a change in oncology AI focus from correlation-based prediction to prescriptive AI informed by causal factors.

8. CONCLUSION

The purpose of his research was to find out if AI can use longitudinal multi-modal data to enable interventional "what-if" reasoning in cancer therapy, going beyond correlation-based prediction. The results show that correct treatment response and individual survival under hypothetical interventions may be estimated by combining structural causal modeling with temporal multi-modal transformers. Proving once and for all that causal-temporal foundation models are crucial for prescriptive oncology AI, the suggested TCMMT enhances counterfactual prediction accuracy and attains better survival discrimination. Trustworthy, decision-relevant insights are delivered by this paradigm through consistent and robust treatment effect estimations across subgroups and censoring circumstances. In sum, the research provides a principled basis for in-silico trials and individualized treatment optimization, and it signifies a shift from predictive to causal, decision-oriented clinical AI.

REFERENCES

[1] S. Sharma, M. Singh, L. McDaid, and S. Bhattacharyya, "XAI-based Data Visualization in Multimodal Medical Data," *bioRxiv*, pp. 2025–07, 2025.

[2] Y. Wang *et al.*, "TWIN-GPT : Digital Twins for Clinical Trials via Large Language Model," *ACM Trans. Multimed. Comput. Commun. Appl.*, p. 3674838, Jul. 2024, doi: 10.1145/3674838.

[3] S. V. M. Sagheer, M. KH, P. M. Ameer, M. Parayangat, and M. Abbas, "Transformers for Multi-Modal Image Analysis in Healthcare.," *Comput. Mater. Contin.*, vol. 84, no. 3, 2025, Accessed: Oct. 13, 2025. [Online]. Available: https://cdn.techscience.press/files/cmc/2025/TSP_CMC-84-3/TSP_CMC_63726/TSP_CMC_63726.pdf

[4] H. Mao, H. Liu, J. X. Dou, and P. V. Benos, "Towards cross-modal causal structure and representation learning," in *Machine Learning for Health*, PMLR, 2022, pp. 120–140. Accessed: Oct. 13, 2025. [Online]. Available: https://proceedings.mlr.press/v193/mao22a.html

[5] Y. Li, H. Ji, F. Yu, L. Cheng, and N. Che, "Temporal multi-modal knowledge graph generation for link prediction," *Neural Netw.*, vol. 185, p. 107108, May 2025, doi: 10.1016/j.neunet.2024.107108.

[6] S. Mewada *et al.*, "Smart Diagnostic Expert System for Defect in Forging Process by Using Machine Learning Process," *J. Nanomater.*, vol. 2022, no. 1, p. 2567194, Jan. 2022, doi: 10.1155/2022/2567194.

[7] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016, doi: 10.1148/radiol.2015151169.

[8] R. H. Keogh and N. Van Geloven, "Prediction Under Interventions: Evaluation of Counterfactual Performance Using Longitudinal Observational Data," *Epidemiology*, vol. 35, no. 3, pp. 329–339, May 2024, doi: 10.1097/EDE.0000000000001713.

[9] A. Khaled *et al.*, "Leveraging MIMIC Datasets for Better Digital Health: A Review on Open Problems, Progress Highlights, and Future Promises," Jun. 15, 2025, *arXiv*: arXiv:2506.12808. doi: 10.48550/arXiv.2506.12808.

[10] G. Li and E. F. Lock, "Integrative Analysis of Multimodal Omics Data," *Annu. Rev. Stat. Its Appl.*, Oct. 2025, doi: 10.1146/annurev-statistics-042424-113016.

[11] W. He *et al.*, "Generative artificial intelligence in medical imaging: Current landscape, challenges, and future directions," *Interdiscip. Med.*, vol. 3, no. 4, p. e20250024, Jul. 2025, doi: 10.1002/INMD.20250024.

[12] A. Hussein, M. Prasad, and A. Braytee, "Explainable AI Methods for Multi-Omics Analysis: A Survey," Oct. 15, 2024, *arXiv*: arXiv:2410.11910. doi: 10.48550/arXiv.2410.11910.

[13] "Causal Foundation Models for Oncology: A Temporal... - Google Scholar." Accessed: Oct. 13, 2025. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Causal+Foundation+Models+for+Oncology%3A+A+Temporal+Multi-Modal+Framework+for+Counterfactual+Prognosis+and+Treatment+Response+in+Glioblastoma&btnG=

[14] F. Jiang, G. Zhao, R. Rodriguez-Monguio, and Y. Ma, "Causal effect estimation in survival analysis with high dimensional confounders," *Biometrics*, vol. 80, no. 4, p. ujae110, Oct. 2024, doi: 10.1093/biomtc/ujae110.

[15] H. Mao, "Causal Machine Learning for Omics Data and Its Applications in Biological Discovery," PhD Thesis, University of Pittsburgh, 2025. Accessed: Oct. 13, 2025. [Online]. Available: https://search.proquest.com/openview/32c78bcdc170506f19cbb3ef1720e035/1?pq-origsite=gscholar&cbl=18750&diss=y

[16] A. Q. Wang *et al.*, "A framework for interpretability in machine learning for medical imaging," *IEEE Access*, vol. 12, pp. 53277–53292, 2024.

[17] M. I. Hossain, G. Zamzmi, P. R. Mouton, M. S. Salekin, Y. Sun, and D. Goldgof, “Explainable AI for Medical Data: Current Methods, Limitations, and Future Directions,” *ACM Comput. Surv.*, vol. 57, no. 6, pp. 1–46, Jun. 2025, doi: 10.1145/3637487.

[18] C. D. Bahadir *et al.*, “Artificial intelligence applications in histopathology,” *Nat. Rev. Electr. Eng.*, vol. 1, no. 2, pp. 93–108, 2024.

[19] Y. Liu *et al.*, “A survey of embodied ai in healthcare: Techniques, applications, and opportunities,” *ArXiv Prepr. ArXiv250107468*, 2025, Accessed: Oct. 13, 2025. [Online]. Available: https://www.researchgate.net/profile/Xu-Cao-10/publication/387976351_From_Screens_to_Scenes_A_Survey_of_Embodied_AI_in_Healthcare/links/6787056f2be36743a5d6a1a8/From-Screens-to-Scenes-A-Survey-of-Embodied-AI-in-Healthcare.pdf

[20] H. Ying, Y. Lia, and Z. Fu, “Domain Adaptation and Generalization Using Foundation Models in Healthcare Imaging,” *Available SSRN 5345726*, 2025, Accessed: Oct. 13, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5345726