CASE-ADAPTIVE EVIDENCE RELIABILITY FOR MULTIMODAL SURVIVAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal survival models combine complementary evidence from computed tomography (CT), radiomics, and clinical records, but the trustworthiness of that evidence varies across patients. Standard fusion rules infer modality weights from content alone and can therefore conflate semantic relevance with data quality, missingness, and prediction-derived certainty. We introduce *Case-Adaptive Evidence Reliability* (CAER), a modular framework that separates these two roles. Each modality produces an independently supervised discrete-time survival distribution. A reliability channel—which does not access the evaluated patient's outcome at inference—combines modality availability, an explicit observable-quality proxy, and distributional certainty. Reliability then acts as a bounded residual to the content gate; exact masking assigns zero weight to unavailable evidence and preserves a patient-level audit trail. On the locked split of the 422-patient NSCLC-Radiomics cohort, CAER obtains the best values among six statistical, tree-based, and neural survival baselines for Harrell and Uno concordance, 1/2/3-year time-dependent AUC, and integrated Brier score. Controlled ablations favor the complete relevance–reliability decomposition over content-only fusion and simpler reliability controls. The results support explicit case-level reliability as a useful inductive bias for multimodal survival prediction.

1 INTRODUCTION

Accurate survival prediction requires models to integrate heterogeneous evidence describing both disease phenotype and patient condition. In oncology, computed tomography (CT) captures tumor anatomy and spatial heterogeneity, radiomics converts medical images into quantitative shape, intensity, and texture descriptors, and clinical records summarize demographics, disease stage, histology, and treatment context (Aerts et al., 2014). Recent large-scale studies have further shown the value of combining imaging and structured clinical evidence for cancer outcome prediction (Jee et al., 2024; Keyl et al., 2025). Nevertheless, multimodal evidence is not equally trustworthy for every patient. CT regions may be incomplete or affected by acquisition and reconstruction variability, radiomic features inherit errors from image preprocessing and segmentation, and clinical variables may be missing or inconsistently recorded. As a result, multimodal survival prediction requires more than complementary representation learning: it must also determine how much each observed modality should be trusted for each individual patient.

Most multimodal fusion methods assign modality weights directly from learned representations, attention scores, or prediction confidence. Such a weight is therefore expected to encode two different properties simultaneously: *content relevance*, which reflects how informative a modality is for the prediction task, and *evidence reliability*, which reflects the quality and stability of the observation available for a particular patient. These properties are related but not equivalent. A modality may be highly predictive at the population level while being incomplete or degraded in an individual case; conversely, high-quality evidence may still produce an uncertain prediction for an intrinsically ambiguous patient. This distinction is particularly important in survival analysis, where censoring and time-dependent risk complicate the interpretation of predictive confidence. Existing work has studied dynamic fusion under low-quality multimodal inputs (Zhang et al., 2023), adaptation to missing modalities (Ramazanova et al., 2025), robustness to complex multimodal noise (Guo & Jin, 2025), and inference-time modality selection (Du et al., 2026). However, these approaches do not directly provide an explicit and auditable reliability mechanism for censored survival prediction. A

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

suitable mechanism should operate without access to the evaluated patient's outcome, distinguish missingness from low-quality presence, respond predictably to changes in evidence quality, and expose how reliability changes the final prediction.

We introduce CAER, a Case-Adaptive Evidence Reliability framework for multimodal survival prediction. CAER first processes CT, radiomic, and clinical evidence using independent modality encoders and discrete-time survival heads, so that every available branch produces an individually meaningful survival distribution (Yu et al., 2011; Lee et al., 2018; Gensheimer & Narasimhan, 2019). For each patient–modality pair, the framework then constructs an operational reliability score from modality availability, an observable quality proxy, and prediction-derived certainty. Observable quality acts as the primary reliability anchor, while certainty provides a smaller model-derived correction. CAER explicitly separates this reliability pathway from the learned content pathway: the content score determines how relevant a modality is to the current prediction, whereas reliability enters as a bounded residual that determines how strongly its evidence should be trusted. Masked normalization assigns exactly zero weight to unavailable modalities and redistributes fusion mass among the remaining evidence. The resulting prediction is accompanied by modality-specific survival distributions, quality indicators, reliability values, content scores, and fusion weights, providing a complete case-level audit trail.

Our contributions are:

- We formulate multimodal survival fusion as two coupled but distinct decisions: learning task-dependent content relevance and estimating patient-specific evidence reliability.
- We introduce an inference-time outcome-independent reliability estimator that combines modality availability, observable evidence quality, and prediction-derived certainty without accessing the evaluated patient's survival time or event status.
- We develop a reliability-residual fusion rule with independently supervised unimodal predictions, exact missing-modality masking, directional responses to reliability changes, bounded influence on the fusion gate, and auditable patient-level weights.
- We evaluate CAER on the NSCLC-Radiomics cohort under a fixed patient-level protocol against six established statistical, ensemble, and neural survival models. CAER achieves the strongest held-out performance across concordance, time-dependent AUC, and integrated Brier score, while controlled ablations support the complete relevance–reliability decomposition.

2 PRELIMINARIES

This section introduces the survival and multimodal-fusion formulations on which our method is built. We focus on the equations needed to distinguish conventional content-based fusion from the explicit reliability mechanism developed in Section 3.

2.1 SURVIVAL PREDICTION UNDER CENSORING

Let t_i and $\delta_i \in \{0, 1\}$ denote the observed follow-up time and event indicator for patient i. Classical Cox models represent relative risk by $r_i = f(x_i)$ and optimize the negative partial log-likelihood

$$\mathcal{L}_{\text{Cox}} = -\sum_{i:\delta_i=1} \left[r_i - \log \sum_{j \in \mathcal{R}(t_i)} \exp(r_j) \right], \tag{1}$$

where $\mathcal{R}(t_i)$ is the risk set immediately before t_i (Britain), 1928). Linear Cox regression, CoxNet, and DeepSurv differ primarily in how $f(\cdot)$ is parameterized or regularized (Simon et al., 2011; Katzman et al., 2018).

Our framework instead uses a discrete-time survival representation (Yu et al., 2011; Lee et al., 2018; Gensheimer & Narasimhan, 2019). After partitioning time into K intervals, a model produces interval hazards $h_{i,k} = \sigma(o_{i,k})$ and survival

$$S_i(t_k) = \prod_{j \le k} (1 - h_{i,j}). \tag{2}$$

If $y_{i,k}$ marks an event in interval k and $r_{i,k}$ indicates that patient i remains at risk, the censor-aware negative log-likelihood is

$$\ell_{\text{surv}}(o_i) = -\sum_{k=1}^{K} r_{i,k} \left[y_{i,k} \log h_{i,k} + (1 - y_{i,k}) \log(1 - h_{i,k}) \right]. \tag{3}$$

This formulation yields a complete survival curve and supports horizon-specific discrimination and probability-based evaluation.

2.2 Content-based multimodal late fusion

For modality $m \in \mathcal{M}$, a standard late-fusion model first computes

$$z_i^m = f_m(x_i^m), \qquad o_i^m = g_m(z_i^m), \tag{4}$$

then assigns a content score $b_i^m = s_m(z_i^m)$. With availability $a_i^m \in \{0, 1\}$, conventional masked attention uses

$$\beta_i^m = \frac{a_i^m \exp(b_i^m/\tau)}{\sum_{n \in \mathcal{M}} a_i^n \exp(b_i^n/\tau)}, \qquad o_i^{\text{late}} = \sum_{m \in \mathcal{M}} \beta_i^m o_i^m. \tag{5}$$

Here b_i^m is learned from modality content. It can express predictive relevance, but it does not explicitly distinguish relevance from acquisition quality, missingness, or output uncertainty. Consequently, a high attention weight is not necessarily evidence that the corresponding observation is trustworthy.

2.3 Quality-aware dynamic fusion

Quality-aware fusion augments the gate with a quality or uncertainty signal q_i^m (Zhang et al., 2023). A generic form is

$$\widetilde{\alpha}_i^m = \frac{a_i^m \exp((b_i^m + \phi(q_i^m))/\tau)}{\sum_{n \in \mathcal{M}} a_i^n \exp((b_i^n + \phi(q_i^n))/\tau)}, \tag{6}$$

where $\phi(\cdot)$ may be learned jointly with the predictor. Recent work handles multimodal noise, missing modalities, or test-time modality selection through adaptation, self-distillation, or dynamic selection (Guo & Jin, 2025; Ramazanova et al., 2025; Du et al., 2026). These results motivate case-adaptive integration, but a jointly learned quality correction can remain difficult to interpret in censored survival prediction. Our narrower objective is an operational reliability variable whose inputs are available without the evaluated patient's outcome, whose effect on the gate is explicit, and whose intermediate values can be inspected.

2.4 Reliability as constrained evidence reweighting

Let b_i^m denote content relevance and let $R_i^m \in [0, 1]$ denote a separately constructed operational reliability score. We seek an effective gate logit

$$g_i^m = b_i^m + \gamma R_i^m, \tag{7}$$

where $\gamma \geq 0$ controls the reliability correction. The content pathway remains free to learn semantic relevance, whereas reliability can only provide a bounded residual. Setting $\gamma = 0$ recovers ordinary late fusion; when content scores are tied, reliability orders their weights. Missing evidence is handled outside Eq. 7 by an exact availability mask, so absence cannot be mistaken for low-quality presence.

This formulation also separates prediction-derived certainty from reliability. Output concentration is observable at inference, but it is not evidence quality: a high-quality observation may yield a diffuse distribution for an ambiguous patient, whereas a poor observation may yield an overconfident prediction. We therefore use certainty only as a small correction to an observable-quality anchor.

3 Method

3.1 Problem formulation

Let $i \in \{1, \ldots, N\}$ index patients and $m \in \mathcal{M} = \{\text{CT}, \text{Rad}, \text{Clin}\}$ index modalities. Patient i has modality input x_i^m, availability $a_i^m \in \{0, 1\}$, observed time t_i, and event indicator δ_i. A training-only

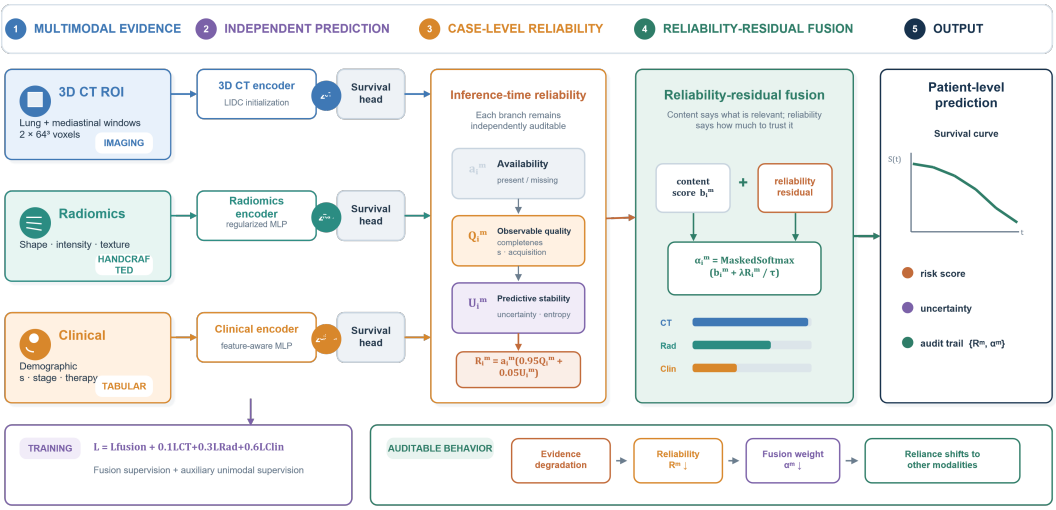


Figure 1: Overview of CAER. CT, radiomic, and clinical evidence are processed by independent survival branches. For each patient and modality, an inference-time outcome-independent reliability module combines availability, observable quality, and prediction-derived certainty. Reliability acts as a residual to a content gate; masked normalization yields patient-adaptive weights, a survival curve, and an audit trail.

transformation T_m, fitted without validation or test patients, produces the branch input $\widetilde{x}_i^m$. The modality encoder ϕ_m maps this input to a representation z_i^m, and a branch-specific linear head produces K interval logits:

$$\widetilde{x}_i^m = T_m(x_i^m), \qquad z_i^m = \phi_m(\widetilde{x}_i^m; \theta_m),$$
$$o_i^m = W_m z_i^m + c_m. \tag{8}$$

The interval hazards and branch survival curve are then

$$h_{i,k}^m = \sigma(o_{i,k}^m), \qquad S_i^m(t_k) = \prod_{j=1}^{k} \left(1 - h_{i,j}^m\right). \tag{9}$$

When $a_i^m = 0$, the branch output is excluded by an exact mask rather than interpreted as low-valued evidence. The prediction problem is therefore to learn a patient-level distribution $S_i(t_k)$ from a variable observed subset $\mathcal{M}_i = \{m : a_i^m = 1\}$ while retaining the independently meaningful branch distributions $S_i^m(t_k)$.

We separate two quantities that ordinary attention tends to conflate. A learned score b_i^m measures how relevant the branch representation is to the current prediction, whereas R_i^m measures how much operational trust the observed evidence warrants. The model outputs the fused curve S_i together with modality-level branch curves, quality, certainty, reliability, and fusion weights. This accompanying trace makes the final prediction auditable without overloading the problem definition with a second composite equation. Unless stated otherwise, CAER denotes the locked instantiation with 95% observable quality and 5% prediction-derived certainty; this numeric mixture is a configuration, not a separate model name.

3.2 Independent Modality Evidence

The CT input is a frozen 128-dimensional nodule embedding extracted from paired lung- and mediastinal-window 64^3 regions by a three-dimensional encoder pretrained on LIDC-IDRI without LUNG1 survival labels (Armato III et al., 2011). The radiomic branch processes 64 fixed shape, first-order, histogram, and peritumoral descriptors; the clinical branch processes 32 encoded demographic, stage, and histology variables. A separate two-layer projector maps each input vector to a 128-dimensional survival representation. Thus all prognosis models operate on fixed, leakage-safe modality features, while the three CAER branches retain distinct projectors and survival heads.

The content path is deliberately separate from reliability. For each branch, a two-layer scorer maps its representation to a scalar relevance logit,

$$b_i^m = (v_m)^\top \tanh(A_m z_i^m + d_m) + e_m. \tag{10}$$

Independent survival heads make every branch directly evaluable and prevent fusion from relying on a predictor that becomes meaningful only through cross-modal interaction. Consequently, b_i^m may encode patient-specific semantic complementarity, while the reliability path below is reserved for inspectable evidence condition.

3.3 Outcome-free case-level reliability

Reliability is computed without the evaluated patient's t_i or δ_i at inference. The branch parameters are nevertheless learned from training outcomes; "outcome independent" refers only to information available when a new patient is evaluated. For each patient–modality pair, we use three ingredients.

Availability. a_i^m indicates whether modality m can contribute. Missing evidence is masked before normalization rather than imputed into a potentially misleading pseudo-observation.

Observable quality. $Q_i^m \in [0, 1]$ is a deterministic proxy computed before survival modeling. Let r_i be the fraction of the tumor mask retained by the fixed ROI and d_i indicate that the mask touches an ROI boundary. For CT and radiomics,

$$Q_i^{\mathrm{CT}} = Q_i^{\mathrm{Rad}} = a_i^{\mathrm{ROI}} r_i (1 - 0.25 d_i). \tag{11}$$

Thus a fully retained, non-boundary ROI receives quality one; a boundary-touching ROI incurs a fixed 0.75 factor; and an unavailable ROI receives zero. For clinical evidence, let $\mathcal{V}$ contain age, sex, clinical T/N/M stage, overall stage, and histology. We use

$$Q_i^{\mathrm{Clin}} = \frac{1}{7} \sum_{v \in \mathcal{V}} \mathbb{I}[v \text{ is observed}]. \tag{12}$$

Missing clinical fields are encoded explicitly and lower Q_i^{Clin} rather than being treated as measured zeros. These scores are operational quality proxies tied to the available preprocessing metadata; they are not clinical judgments of scan or record quality.

Prediction-derived certainty. Branch hazards define a $(K+1)$-class event-time distribution p_i^m: $p_{i,k}^m = S_i^m(t_{k-1}) h_{i,k}^m$ for $k \le K$, with the remaining survival mass assigned to class $K+1$. We compute maximum mass, normalized entropy, and a bounded logit-energy proxy,

$$\mathrm{Conf}_i^m = \max_k p_{i,k}^m, \qquad \mathrm{Ent}_i^m = -\frac{\sum_k p_{i,k}^m \log p_{i,k}^m}{\log(K+1)}, \tag{13}$$

$$\mathrm{FE}_i^m = \frac{1}{K \log 2} \sum_{k=1}^{K} \mathrm{softplus}(-|o_{i,k}^m|). \tag{14}$$

Their average defines a prediction-derived certainty proxy

$$P_i^m = \frac{1}{3} \left[\mathrm{Conf}_i^m + (1 - \mathrm{Ent}_i^m) + (1 - \mathrm{FE}_i^m) \right]. \tag{15}$$

This quantity measures concentration of the predicted event-time distribution, not stability under repeated acquisition or perturbation. Because certainty can be miscalibrated, it is assigned only a small role. The locked CAER reliability is

$$R_i^m = a_i^m \left(0.95 Q_i^m + 0.05 P_i^m \right). \tag{16}$$

The dominant observable term anchors trust in inspectable metadata, while the smaller certainty correction responds to the branch distribution. The mixture was selected on development data and fixed before locked evaluation.

3.4 RELIABILITY-RESIDUAL FUSION

A learned content scorer produces $b_i^m = s_m(z_i^m)$, which captures the semantic relevance of modality m for patient i. Following the implementation, the residual temperature τ_R scales only the reliability term. We define

$$g_i^m = b_i^m + \frac{\lambda}{\tau_R} R_i^m, \tag{17}$$

$$\alpha_i^m = \frac{a_i^m \exp(g_i^m)}{\sum_{n \in \mathcal{M}} a_i^n \exp(g_i^n)}, \tag{18}$$

$$o_i^{\text{fusion}} = \sum_{m \in \mathcal{M}} \alpha_i^m o_i^m, \qquad S_i(t_k) = \prod_{j \le k} \left[1 - \sigma(o_{i,j}^{\text{fusion}})\right]. \tag{19}$$

The availability factor sets $\alpha_i^m = 0$ when $a_i^m = 0$ and renormalizes over observed modalities. The scale λ controls correction magnitude, while τ_R controls sensitivity to reliability without rescaling the content logits. If observable quality or prediction-derived certainty decreases, R_i^m decreases and transfers fusion mass to other available modalities. This path is part of the prediction mechanism rather than a post-hoc explanation.

3.5 BEHAVIORAL PROPERTIES AND INFERENCE PROCEDURE

The residual construction gives a deterministic directional response. For an available modality m, differentiating the masked softmax while holding content scores fixed yields

$$\frac{\partial \alpha_i^m}{\partial R_i^m} = \frac{\lambda}{\tau_R} \alpha_i^m (1 - \alpha_i^m) \ge 0, \tag{20}$$

$$\frac{\partial \alpha_i^n}{\partial R_i^m} = -\frac{\lambda}{\tau_R} \alpha_i^n \alpha_i^m \le 0, \quad n \ne m. \tag{21}$$

Thus increasing the reliability of one observed modality cannot decrease its own weight and transfers mass away from the remaining modalities. Because $R_i^m \in [0, 1]$, the correction is bounded by λ/τ_R. The gate is invariant to a common shift of available content logits, and the mask keeps missing modalities at exactly zero weight. Appendix C proves the corresponding bounded-odds, exact-masking, and perturbation-stability properties.

Inference is identical for complete and incomplete cases. Each available feature vector is projected and passed through its independently supervised survival head; Eqs. 11–15 produce Q_i^m and P_i^m; Eq. 16 gives reliability; and Eq. 18 combines reliability with content. The stored branch logits, scores, and weights reconstruct the final survival curve.

3.6 CENSOR-AWARE OPTIMIZATION

Every fused or unimodal survival term instantiates the censor-aware likelihood in Eq. 3. The selected model combines fused supervision with explicit unimodal supervision:

$$\mathcal{L} = \mathcal{L}_{\text{fusion}} + 0.1\mathcal{L}_{\text{CT}} + 0.3\mathcal{L}_{\text{Rad}} + 0.6\mathcal{L}_{\text{Clin}}. \tag{22}$$

The weights maintain predictive pressure on every branch while retaining fusion as the primary output. They were selected on development data and fixed before locked evaluation. The reported model uses no cross-modal distillation, so the comparison focuses on independently supervised evidence and reliability-residual fusion.

4 EXPERIMENTS

4.1 COHORT, PREPROCESSING, AND PROTOCOL

We evaluate on the public NSCLC-Radiomics (LUNG1) cohort, which contains pretreatment CT, tumor delineations, clinical variables, and overall-survival outcomes for 422 patients (Aerts et al., 2014). The endpoint is right-censored overall survival measured from treatment start. Splits are

fixed at the patient level: 269 patients for training, 68 for validation, and 85 for held-out evaluation. No patient contributes to more than one split, and the held-out outcomes are not used for feature selection, early stopping, or hyperparameter choice.

CT volumes are converted to Hounsfield units, resampled to 1 mm isotropic spacing, and cropped to fixed 64^3 regions. Lung and mediastinal windows are passed through the independently trained LIDC-IDRI encoder once, and its frozen 128-dimensional embedding is used by all prognosis models; no LUNG1 outcome enters pretraining. Radiomic and clinical vectors are standardized using training-set statistics only. Missing modalities are zero-filled only after availability indicators have been constructed, so imputed zeros remain distinguishable from observations.

All empirical transformations are estimated on training patients, and early stopping and model selection use only validation patients. The held-out partition is opened only after the architecture, reliability mixture, loss weights, and checkpoints are fixed. Reliability never receives the evaluated patient's survival outcome.

4.2 Baselines and implementation

We compare against six established survival models: elastic-net Cox regression (CoxNet) (Simon et al., 2011), random survival forest (RSF) (Ishwaran et al., 2008), DeepSurv (Katzman et al., 2018), MTLR (Yu et al., 2011), gradient boosting survival analysis (GBSA) (Hothorn et al., 2006), and DeepHit-Single (Lee et al., 2018). Each baseline receives the same concatenation of the frozen CT embedding, radiomic vector, clinical vector, and three availability indicators used to construct CAER inputs. Unavailable modality blocks are set to zero after masking, and all fitted preprocessing uses training patients only. The baselines therefore differ in survival estimator and fusion structure, not in access to patient evidence. Neural and stochastic models use five random seeds; seed-specific survival predictions are averaged before evaluation. Classical estimators and common metric routines are implemented with scikit-survival (Pölsterl, 2020).

CAER uses 12 time intervals, 128-dimensional modality representations, reliability temperature $\tau_R = 0.5$, residual scale $\lambda = 0.5$, and the 0.95/0.05 mixture in Eq. 16. The locked configuration uses neither cross-modal distillation nor modality dropout. We optimize with AdamW, learning rate 10^{-4}, weight decay 10^{-4}, batch size 16, and at most 200 epochs. Early stopping uses validation loss with patience 30; the held-out set is not used for model selection.

Every method uses the same outcome coding, patient partition, evaluation horizons, training-derived preprocessing, and common metric implementation. Neural models follow the same validation-only selection rule.

4.3 Metrics

We report Harrell's concordance index (Harrell et al., 1982), Uno's inverse-probability-of-censoring weighted concordance index (Uno et al., 2011), cumulative/dynamic AUC at 1, 2, and 3 years, and integrated Brier score (IBS) (Graf et al., 1999). Concordance and time-dependent AUC measure discrimination; IBS evaluates the probabilistic accuracy of the complete survival curve. Higher is better for concordance and AUC, while lower is better for IBS.

4.4 Main comparison with established survival models

Table 1 shows that CAER attains the best value on all six metrics. Relative to the strongest non-CAER baseline for each measure, it improves Harrell C by 0.0467 and Uno C by 0.0461. The benefit persists across prediction horizons: AUC improves by 0.0184 at one year, 0.0552 at two years, and 0.0623 at three years. IBS decreases from 0.2315 for the best baseline to 0.2147. The joint gain in ranking, horizon-specific discrimination, and probabilistic accuracy indicates that reliability-residual fusion improves the survival distribution rather than optimizing a single operating point.

The one-year AUC gain is smaller than the two- and three-year gains, while concordance summarizes ordering over follow-up and IBS evaluates the complete probability trajectory. Their joint direction makes the result less dependent on a single horizon. CoxNet is the strongest non-CAER comparator in concordance and RSF in IBS; the nonlinear neural baselines do not consistently exceed them. This

Table 1: Held-out comparison on the common patient-level evaluation split. Five trained seeds are combined at prediction level. Bold indicates the best value. Higher is better for concordance and AUC; lower is better for IBS.

Method	Harrell C ↑	Uno C ↑	AUC 1y ↑	AUC 2y ↑	AUC 3y ↑	IBS ↓
CoxNet	0.5639	0.5632	0.5584	0.6380	0.6819	0.2433
Random survival forest	0.5549	0.5550	0.5058	0.5980	0.6688	0.2315
DeepSurv	0.5208	0.5190	0.4658	0.5449	0.5698	0.2477
MTLR	0.5123	0.5097	0.5020	0.5521	0.5121	0.2784
GBSA	0.5392	0.5407	0.4840	0.5755	0.6265	0.2757
DeepHit-Single	0.5437	0.5412	0.5415	0.5717	0.5143	0.3372
CAER (ours)	**0.6106**	**0.6093**	**0.5768**	**0.6932**	**0.7442**	**0.2147**

Table 2: Ablation of the fusion mechanism under the shared multimodal protocol.

Variant	Harrell C ↑	Uno C ↑	IBS ↓
Late fusion	0.5745	0.5734	0.2170
Equal residual	0.5846	0.5844	0.2175
Learned reliability	0.5860	0.5846	0.2174
Observable quality only	0.5883	0.5870	0.2167
CAER (ours)	**0.6106**	**0.6093**	**0.2147**

pattern is consistent with the hypothesis that evidence combination, rather than model capacity alone, is consequential in this cohort.

4.5 Component ablation

Table 2 changes the fusion rule while retaining the same modality features and branch architecture. Equal residual weighting raises concordance but slightly worsens IBS, so an extra residual path alone is insufficient. Learned reliability yields only a small further change. Observable quality alone reaches 0.5883 Harrell C and 0.2167 IBS, showing that the explicit proxy is informative. Adding the 5% certainty correction yields 0.6106 Harrell C and 0.2147 IBS. Relative to late fusion, complete CAER improves Harrell C by 0.0361 and Uno C by 0.0359 and reduces IBS by 0.0023; relative to quality only, the corresponding changes are $+0.0223$, $+0.0223$, and -0.0020. These comparisons support the intended decomposition without treating output certainty as a substitute for observable quality.

4.6 Auditable case-adaptive behavior

The model exports $(o_i^m, Q_i^m, P_i^m, R_i^m, b_i^m, \alpha_i^m)$ for every available modality. These quantities separate high relevance with low operational reliability, high reliability with low relevance, and absence with $\alpha_i^m = 0$. Because the residual enters before normalization, the trace links each quality proxy and branch distribution to the final fusion weight and survival curve.

The rule neither reconstructs missing inputs nor requires outcome-conditioned reliability labels or a separate inference policy. For a patient, the fused prediction can be reconstructed from stored branch logits and weights, unavailable modalities can be verified to receive zero mass, and content and reliability contributions can be separated. The mechanism is therefore falsifiable: under fixed content, a declared quality degradation that fails to reduce the corresponding R_i^m and α_i^m violates Eq. 21.

5 Discussion

The central empirical finding is that multimodal survival prediction benefits from separating relevance from operational trust. A content gate identifies complementary information, whereas CAER supplies availability, observable quality, and prediction-derived certainty through an explicit bounded residual. The architecture therefore remains close to conventional late fusion while giving its weights a more

specific semantics: each weight combines learned relevance with an inspectable, case-level reliability adjustment.

This separation provides a useful inductive bias for modest clinical cohorts. A fully learned gate can absorb correlations between modality identity, missingness, and outcome without revealing which mechanism drives its decision. CAER instead constrains one part of the gate to an inference-time outcome-independent score and leaves semantic relevance to the content pathway. Independent branch supervision keeps each survival output usable on its own, reducing the number of roles assigned to any single latent score.

The result is consistent with quality-aware multimodal learning under variable evidence quality (Zhang et al., 2023) and incomplete-modality methods that adapt evidence at test time (Ramazanova et al., 2025; Du et al., 2026). CAER instead targets censored survival distributions while preserving an inference-time outcome-independent reliability trace, and its branch representations can be upgraded without changing the reliability interface.

The larger AUC gains at two and three years, together with the lower IBS, are consistent with a distribution-level benefit rather than improvement at one isolated operating point. We interpret this pattern as evidence that balancing complementary modalities matters across the predicted survival trajectory; establishing a horizon-specific causal mechanism would require additional cohorts and interventions.

The method is modular: a stronger backbone or alternative survival head can replace a branch if it exposes a valid distribution and the required metadata. Rather than a single opaque score, CAER outputs a compact evidence ledger—availability, operational quality, prediction-derived certainty, reliability, and fusion weight—for quality control and prospective monitoring.

The residual form is important beyond interpretability. Directly multiplying a content score by a quality estimate can suppress semantically useful evidence when the quality proxy is imperfect, whereas replacing content attention with reliability would discard patient-specific complementarity. In CAER, content remains the primary routing signal and reliability supplies a bounded correction whose maximum influence is controlled by λ/τ_R. This design also supports a simple counterfactual audit: holding branch representations and logits fixed while varying Q_i^m changes only the reliability residual, so the resulting change in α_i^m can be attributed to the declared evidence condition rather than to an unobserved change in semantic features.

Scope and limitations. The present evidence comes from one public cohort and an 85-patient held-out split. The quality terms are deliberately simple operational proxies: CT and radiomics share ROI-retention metadata, and the clinical score measures field completeness rather than record correctness. Moreover, P_i^m measures concentration of the model output, not test–retest stability or calibrated clinical confidence. The results therefore validate the proposed mechanism under the stated protocol, but prospective and multi-center studies are needed before clinical use.

6 Conclusion

We introduced CAER, a case-adaptive evidence reliability framework for multimodal survival prediction. The method maintains independently supervised CT, radiomic, and clinical branches; combines availability, operational quality, and prediction-derived certainty without using the evaluated patient's outcome at inference; and injects the resulting reliability as a residual into content-based late fusion. On NSCLC-Radiomics, CAER outperforms six established survival baselines across concordance, time-dependent AUC, and integrated Brier score, while controlled ablations favor the complete decomposition. These results support explicit case-level reliability as an effective and auditable mechanism for multimodal prognosis.

REPRODUCIBILITY STATEMENT.

We provide fixed patient-level splits, multiple random seeds, validation-only early stopping, and a locked held-out test set. Detailed preprocessing, model configuration, optimization settings, baseline implementations, metric definitions, and ablation protocols are provided in the appendix and supplementary materials.

ETHICS STATEMENT.

This study uses de-identified public research data for research-only prognostic modeling. Patient-level separation is enforced to prevent data leakage. The proposed model is not intended for direct clinical decision-making or patient identification.

AI USE STATEMENT

Generative AI tools were used to assist literature organization, manuscript drafting and language editing, LaTeX preparation, code implementation, and result presentation. Procedural scientific-agent resources were used during local document preparation (Kassis et al., 2026). The authors independently designed the research methodology, conducted the experiments, verified numerical results, reviewed all AI-assisted content, and take responsibility for the final manuscript.

REFERENCES

Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):4006, 2014.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Royal Statistical Society (Great Britain). Journal of the royal statistical society: Series b (methodological). Royal Statistical Society., 1928.

Siyi Du, Xinzhe Luo, Declan O'regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations*, volume 2026, pp. 116887–116913, 2026.

Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Zirun Guo and Tao Jin. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In *International Conference on Learning Representations*, volume 2025, pp. 2215–2232, 2025.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008.

Justin Jee, Christopher Fong, Karl Pichotta, Thinh Ngoc Tran, Anisha Luthra, Michele Waters, Chenlian Fu, Mirella Altoe, Si-Yang Liu, Steven B Maron, et al. Automated real-world data integration improves cancer outcome prediction. *Nature*, 636(8043):728–736, 2024.

Timothy Kassis, Vinayak Agarwal, Yuhuan He, Darshil Patel, and Aubrey M Brueckner. Scientific agent skills: A library of procedural knowledge for research agents. *arXiv preprint arXiv:2609.00065*, 2026.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18:24, 2018. doi: 10.1186/s12874-018-0482-1.

Julius Keyl, Philipp Keyl, Grégoire Montavon, René Hosch, Alexander Brehmer, Liliana Mochmann, Philipp Jurmeister, Gabriel Dernbach, Moon Kim, Sven Koitka, et al. Decoding pan-cancer treatment outcomes using multimodal real-world data and explainable artificial intelligence. *Nature Cancer*, 6(2):307–322, 2025.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *International Conference on Learning Representations*, volume 2025, pp. 82755–82774, 2025.

Noah Simon, Jerome H Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39:1–13, 2011.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in neural information processing systems*, 24, 2011.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

A Cohort Construction and Preprocessing

A.1 Patient-level partition

The NSCLC-Radiomics cohort contains 422 patients. We use a fixed partition of 269 training, 68 validation, and 85 held-out patients. All CT volumes, tumor regions, derived radiomic descriptors, and clinical records belonging to one patient remain in the same partition. The training split is used to estimate preprocessing statistics and fit model parameters; the validation split is used for early stopping and model selection; the held-out split is opened only after the complete method and hyperparameters have been selected. No held-out outcome is used to construct reliability, determine thresholds, or choose a checkpoint.

A.2 MODALITY PREPROCESSING

CT. DICOM pixels are converted to Hounsfield units, resampled to 1 mm isotropic spacing, and cropped to a fixed $64 \times 64 \times 64$ tumor-centered region. Lung and mediastinal windows are generated from the same HU volume and stacked as two channels. An LIDC-IDRI nodule encoder trained without LUNG1 survival outcomes is applied once to produce the frozen 128-dimensional CT embedding used by the prognosis branch.

Radiomics. The radiomic input contains shape, first-order intensity, and texture descriptors derived from the tumor region. Feature preprocessing is fit on the training partition and then applied unchanged to validation and held-out patients.

Clinical variables. Clinical covariates summarize age, sex, clinical T/N/M stage, overall stage, and histology. Continuous transformations and missing-value handling are determined from the training partition. Missingness is retained through explicit availability or validity indicators rather than being interpreted as a measured zero.

Leakage controls. Patient identifiers, partition membership, and survival outcomes are joined only through the fixed manifest. LIDC-IDRI pretraining never uses LUNG1 survival labels. The reliability estimator receives modality availability, observable quality, and branch-derived certainty statistics, but never receives t_i or δ_i at inference time.

B ARCHITECTURE AND RELIABILITY COMPONENTS

Table 3: Modality-specific branches used by CAER. Each branch produces a 128-dimensional representation and an independently supervised survival output.

Modality	Input	Encoder	Branch output
CT	Frozen 128-D LIDC embedding	Two-layer MLP projector	K hazard logits
Radiomics	64 fixed descriptors	Two-layer MLP projector	K hazard logits
Clinical	32 encoded variables	Two-layer MLP projector	K hazard logits

B.1 INDEPENDENT BRANCH SUPERVISION

Each modality branch minimizes Eq. 3 whenever its input and label are valid. This prevents a branch from becoming meaningful only through cross-modal interaction. The branch logits are therefore usable for unimodal evaluation, single-modality fallback, prediction-derived certainty, and patient-level auditing.

B.2 RELIABILITY VARIABLES

For every patient–modality pair, the evaluator records the following quantities:

- a_i^m: a binary availability indicator applied before normalization;
- Q_i^m: the operational evidence-quality proxy defined in Eqs. 11–12;
- P_i^m: the prediction-derived certainty in Eq. 15;
- R_i^m: the final inference-time outcome-independent reliability in Eq. 16;
- b_i^m and α_i^m: the content score and final masked fusion weight.

All reliability inputs are available at inference time. The selected CAER configuration assigns 95% of the reliability mixture to observable quality and 5% to prediction-derived certainty. The residual is added to, rather than substituted for, the content score. This retains the distinction between semantic relevance and operational evidence trust.

B.3 MISSING-MODALITY BEHAVIOR

The masked softmax in Eq. 18 enforces $\alpha_i^m = 0$ whenever $a_i^m = 0$. Remaining weights are renormalized to sum to one. If exactly one modality is available, its weight becomes one and the corresponding unimodal survival output is returned. The locked CAER configuration uses the observed availability patterns directly and does not apply additional modality dropout.

C THEORETICAL PROPERTIES OF RELIABILITY-RESIDUAL FUSION

This section formalizes the behavior used to motivate the reliability-residual gate. The analysis is conditional on a patient, a nonempty set of available modalities $\mathcal{A}$, fixed content logits $\{b_m : m \in \mathcal{A}\}$, and fixed branch predictions. Let $\kappa = \lambda/\tau_R \geq 0$ and suppress the patient index. For each $m \in \mathcal{A}$,

$$\alpha_m(R) = \frac{\exp(b_m + \kappa R_m)}{\sum_{n\in\mathcal{A}} \exp(b_n + \kappa R_n)}. \tag{23}$$

The statements concern the fusion mechanism itself; they do not assume that an estimated quality proxy equals an unobserved clinical notion of evidence correctness.

Proposition 1 (directional reliability response). For every available modality m, its fusion weight is nondecreasing in its own reliability, while every other available weight is nonincreasing:

$$\frac{\partial \alpha_m}{\partial R_m} = \kappa \alpha_m(1 - \alpha_m) \geq 0, \qquad \frac{\partial \alpha_n}{\partial R_m} = -\kappa \alpha_n \alpha_m \leq 0 \quad (n \neq m). \tag{24}$$

Proof. Equation 23 is a softmax applied to $b + \kappa R$. Its Jacobian satisfies $\partial \alpha_n / \partial (b_m + \kappa R_m) = \alpha_n(\mathbb{I}[n = m] - \alpha_m)$. Multiplication by κ gives Eq. 24; the signs follow from $\kappa \geq 0$ and $0 \leq \alpha_m \leq 1$. □

Proposition 2 (bounded change in pairwise routing odds). Let $\beta_m = \exp(b_m)/\sum_{n\in\mathcal{A}} \exp(b_n)$ be the content-only gate. For any available pair m, n,

$$\log \frac{\alpha_m/\alpha_n}{\beta_m/\beta_n} = \kappa(R_m - R_n), \qquad \left|\log \frac{\alpha_m/\alpha_n}{\beta_m/\beta_n}\right| \leq \kappa. \tag{25}$$

Proof. The common softmax denominators cancel, giving $\log(\alpha_m/\alpha_n) = b_m - b_n + \kappa(R_m - R_n)$. Subtracting the content-only log odds yields the equality. Since $R_m, R_n \in [0, 1]$, $|R_m - R_n| \leq 1$, which proves the bound. □

Proposition 2 distinguishes a residual correction from a reliability-only router. Reliability can systematically alter pairwise routing odds, but its maximum multiplicative effect is controlled by $\exp(\kappa)$ in either direction. The content pathway therefore cannot be displaced without a correspondingly large, explicit choice of κ.

Proposition 3 (exact missing-modality consistency). If $a_m = 0$, then $\alpha_m = 0$. For any two available modalities m, n, their weight ratio is independent of all unavailable branches:

$$\frac{\alpha_m}{\alpha_n} = \exp[(b_m - b_n) + \kappa(R_m - R_n)]\,. \tag{26}$$

If $\mathcal{A} = \{m\}$, then $\alpha_m = 1$.

Proof. The availability multiplier in Eq. 18 makes every unavailable numerator exactly zero. Cancelling the common denominator for two available modalities gives Eq. 26; hence adding or removing zero-masked branches cannot change their ratio. A singleton denominator equals its only numerator, so the remaining weight is one. □

Proposition 4 (stability to reliability perturbations). For a fixed availability set and fixed content logits, the gate is Lipschitz in reliability under the ℓ_1 norm:

$$\|\alpha(R) - \alpha(R')\|_1 \leq \frac{\kappa}{2}\|R - R'\|_1. \tag{27}$$

Consequently, for fixed branch logits at interval k,

$$\left|o_k^{\text{fusion}}(R) - o_k^{\text{fusion}}(R')\right| \le \frac{\kappa}{2} \max_{m \in \mathcal{A}} |o_k^m| \, \|R - R'\|_1. \tag{28}$$

Proof. For a fixed reliability coordinate R_m, the absolute column sum of the softmax Jacobian is

$$\sum_n \left| \frac{\partial \alpha_n}{\partial R_m} \right| = 2\kappa \alpha_m (1 - \alpha_m) \le \frac{\kappa}{2}, \tag{29}$$

because $u(1-u) \le 1/4$ for $u \in [0,1]$. Thus the induced Jacobian 1-norm is at most $\kappa/2$ everywhere. Integrating the Jacobian along the line segment from R to R' proves Eq. 27. Finally, $|\sum_m (\alpha_m(R) - \alpha_m(R'))o_k^m| \le \max_m |o_k^m| \, \|\alpha(R) - \alpha(R')\|_1$, which gives Eq. 28. □

For the reported configuration, $\lambda = \tau_R = 0.5$ and therefore $\kappa = 1$. The theoretical results guarantee directional, bounded, and missingness-consistent routing for the declared reliability values. They complement rather than replace the empirical evaluation of whether the chosen observable proxies are informative on the cohort.

D Optimization and Model Selection

Table 4: Configuration of the reported CAER model.

Setting	Value	Setting	Value
Time intervals K	12	Representation size	128
Optimizer	AdamW	Learning rate	10^{-4}
Weight decay	10^{-4}	Batch size	16
Maximum epochs	200	Early-stop patience	30
Reliability temperature τ_R	0.5	Residual scale λ	0.5
Quality/certainty mixture	0.95/0.05	Modality dropout	0.00
Loss weights	$1/0.1/0.3/0.6$	Distillation	disabled

The loss-weight row lists fusion, CT, radiomic, and clinical terms, respectively. The five random seeds are 20260729–20260733. The discrete-time boundaries are 0, 90, 180, 270, 365, 548, 730, 913, 1095, 1460, 1825, 2555, and 5475 days. At-risk masks are generated from the observed follow-up time and censoring indicator. Neural models use validation-loss early stopping; the checkpoint with the best validation criterion is retained for that seed. Five seed-specific predictions are combined at prediction level for the final held-out evaluation.

E Baseline and Ablation Definitions

E.1 External survival baselines

CoxNet uses elastic-net regularization of the proportional-hazards objective in Eq. 1. Random survival forest and GBSA provide nonlinear tree ensembles. DeepSurv replaces the linear Cox score with a neural risk function. MTLR and DeepHit-Single learn discrete-time survival distributions. All baselines receive the same patient partition, endpoint, feature availability, training-only preprocessing, evaluation horizons, and censoring estimator as CAER.

E.2 Internal fusion ablations

The ablations alter the reliability and fusion rule while preserving the underlying modality branches, patient partitions, optimization protocol, and evaluation metrics. This isolates the contribution of the explicit reliability pathway from changes in encoder capacity or data access.

Table 5: Interpretation of the fusion ablations reported in Table 2.

Variant	Definition
Late fusion	Content gate from Eq. 5; no explicit reliability residual.
Equal residual	The same residual contribution is assigned to every available modality.
Learned reliability	Reliability is produced by a learned estimator without the fixed CAER decomposition.
Observable quality only	$R_i^m = a_i^m Q_i^m$; prediction-derived certainty is omitted.
CAER	$R_i^m = a_i^m(0.95Q_i^m + 0.05P_i^m)$ followed by Eq. 18.

F Evaluation Metrics

Let η_i denote predicted scalar risk and let $\mathcal{P}$ be the set of comparable patient pairs. Harrell concordance can be written as

$$C_{\text{Harrell}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbb{I}(\eta_i > \eta_j), \tag{30}$$

with standard handling of tied risks. Uno C uses inverse-probability-of-censoring weights to reduce sensitivity to censoring. For a prediction horizon t, cumulative/dynamic AUC compares patients experiencing an event by t with patients known to survive beyond t.

The censoring-adjusted Brier score evaluates the predicted survival probability:

$$\text{BS}(t) = \frac{1}{N} \sum_{i=1}^{N} w_i(t) \left[\mathbb{I}(T_i > t) - \widehat{S}_i(t)\right]^2, \tag{31}$$

where $w_i(t)$ denotes the inverse-probability-of-censoring weight. IBS integrates $\text{BS}(t)$ over the prespecified evaluation interval. Harrell C, Uno C, and AUC are maximized; IBS is minimized. All methods use the same horizons and censoring estimator.

G Reproducibility and Audit Outputs

Every seed-specific run records the resolved configuration, software environment, split identifier, preprocessing parameters, best epoch, checkpoint, and evaluation outputs. The held-out evaluation stores one row per patient together with observed time, event indicator, predicted survival curve, and scalar risk. For CAER, the output additionally contains modality availability, observable quality, prediction-derived certainty, reliability, content score, fusion weight, and unimodal logits. These fields permit reconstruction of the fusion decision for an individual patient without rerunning training.

The intended reproduction order is: (1) validate the fixed patient manifest; (2) fit all preprocessing transformations on the training split; (3) train each seed using validation-only early stopping; (4) freeze selected checkpoints; (5) generate patient-level survival predictions; and (6) evaluate all methods with the common metric implementation. Any unavailable or failed run remains explicit in the run manifest rather than being silently omitted.