M2M-FSURV: MULTIMODAL DEEP FUNCTIONAL MODELING FOR COMPETING RISKS SURVIVAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern clinical datasets contain rich multimodal information, including clinical narratives, longitudinal physiological measurements, and baseline patient characteristics. At the same time, patients may experience multiple competing outcomes, such as discharge or in-hospital death. Integrating these heterogeneous modalities to model competing events over time remains a fundamental challenge in survival analysis. Existing classical and deep learning–based survival models remain limited in jointly modeling multimodal inputs, competing risks, and interpretable continuous-time event predictions. To address this limitation, we propose M2M-FSurv, a deep learning framework for multimodal competing-risks survival analysis that incorporates principles from functional data analysis to estimate smooth, continuous-time cause-specific density incidence functions. M2M-FSurv employs an attention-based shared encoder that integrates inputs and adjusts modality contributions across individuals, along with cause-specific prediction heads that estimate smooth, cause-specific event-time densities approximated using B-spline expansions. We evaluate the framework on the MIMIC-III and METABRIC datasets. Experimental results demonstrate that M2M-FSurv achieves improved predictive performance while providing interpretable insights into modality contributions and cause-specific event dynamics.

1 INTRODUCTION

Predicting when a patient will experience an outcome, and which outcome will occur, is a central problem in clinical time-to-event modeling. In intensive care, discharge alive and in-ICU death are competing events: the occurrence of one precludes the other during the same ICU stay. Predicting these outcomes requires integrating complementary information from clinical narratives, longitudinal physiological measurements, and baseline patient characteristics. These modalities differ in their structure and temporal resolution, creating challenges for learning patient representations and estimating event-specific probabilities over time. We study this problem under a 24-hour landmark design, using information available during the first 24 hours after ICU admission to predict subsequent outcomes among patients who remain alive and in the ICU at the landmark.

Classical regression approaches, including Cox proportional hazards and Fine–Gray models, provide established tools for survival and competing-risks analysis but commonly impose proportionality assumptions and specified covariate effects (Fine & Gray, 1999; Mao & Lin, 2017). Deep learning approaches accommodate nonlinear covariate relationships using different event-time representations. DeepHit predicts a joint distribution over event type and discretized time (Lee et al., 2018). SAMVAE further combines multimodal inputs with parametric continuous-time event distributions (Garrido et al., 2025). Building on these developments, we investigate whether locally supported B-spline representations of cause-specific event-time subdensities, coupled with multimodal fusion, can improve event-specific probability prediction from heterogeneous clinical histories.

In this paper, we consider three modalities commonly available in electronic health records: unstructured clinical notes (e.g., physician, nursing, and radiology reports), longitudinal physiological measurements (e.g., blood pressure and heart rate trajectories), and scalar covariates (e.g., age, sex, and ethnicity). We adopt a 24-hour landmark design, using patient information collected during the first

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.

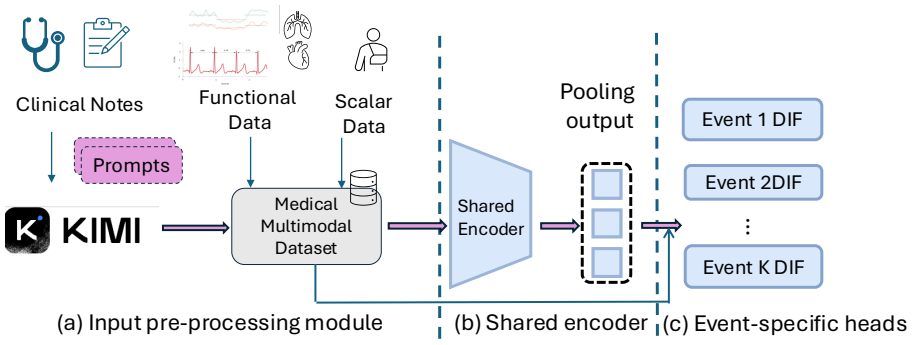


Figure 1: Overview of the proposed M2M-FSurv framework.

24 hours after ICU admission to predict competing outcomes occurring after the landmark. Clinical notes can capture contextual information, including imaging interpretations, treatment responses, and clinician assessments, that may complement physiological measurements and demographic variables. We therefore investigate their joint contribution to post-landmark competing-risks prediction. To capture temporal dynamics in physiological measurements,we treat longitudinal physiological measurements as discretely observed trajectories and encode their values and observation times to capture temporal variation. (Dong et al., 2023; Feng et al., 2025). Given these multimodal observations up to the 24-hour landmark, our goal is to estimate smooth, cause-specific event-time densities for competing outcomes after the landmark.

We propose M2M-FSurv, a multimodal deep neural network for competing-risks prediction. As shown in Fig. 1, M2M-FSurv comprises three components: (i) modality-specific preprocessing and encoding, (ii) an attention-based shared encoder, and (iii) cause-specific prediction heads. Clinical notes are processed using a task-specific prompt and encoded into text representations. These representations, along with longitudinal physiological measurements and scalar covariates, are mapped into a shared embedding space through modality-specific encoders. The shared encoder applies multi-head self-attention to model interactions among the resulting tokens and construct a patient-level representation. Each prediction head outputs nonnegative B-spline coefficients for a cause-specific density incidence function (DIF), yielding smooth continuous-time event-time densities. Unlike DeepHit Lee et al. (2018), which predicts probabilities over a discrete time grid, M2M-FSurv allows event-specific predictions at arbitrary post-landmark times and facilitates visualization of their temporal patterns.

Our contributions are threefold. First, we develop a multimodal landmark prediction framework that integrates clinical text, hourly physiological measurements, and scalar covariates for competing-risks prediction. Second, we parameterize cause-specific event-time subdensities using normalized B-spline mixtures and jointly predicted event probabilities, yielding valid cumulative incidence functions over a finite analysis horizon. Third, we evaluate the complete framework on MIMIC-III, with modality and component ablations, and provide a complementary scalar-only, single-event evaluation on METABRIC. We additionally establish a conditional consistency result for a restricted feedforward coefficient class.

2 Related Work

Survival Analysis and Competing Risks. Classical survival analysis includes nonparametric survival estimation using the Kaplan–Meier estimator (Kaplan & Meier, 1958) and covariate-based regression using the Cox proportional hazards model (Cox, 1972). In competing-risks settings, cumulative incidence functions describe the probability of experiencing a particular event in the presence of competing outcomes. Fine–Gray regression models the subdistribution hazard associated with a cause-specific cumulative incidence function (Fine & Gray, 1999), while transformation models provide alternative regression formulations (Mao & Lin, 2017). Deep learning approaches accommodate nonlinear covariate relationships. DeepHit learns a joint distribution over event type and discretized event time (Lee et al., 2018), whereas SurvTRACE uses a Transformer to model covariate interactions for survival prediction with competing events (Wang & Sun, 2022). These

methods motivate the use of flexible representations and event-specific outputs, while differing in how they parameterize time and incorporate patient information.

Multimodal Learning in Healthcare. Multimodal survival models integrate complementary patient information through modality-specific representations and fusion mechanisms. Transformer-based architectures have also been explored for time-to-event prediction (Wang & Sun, 2022). Existing work has also explored continuous-time competing-risks prediction with multimodal inputs. SAMVAE (Garrido et al., 2025) already addresses continuous-time multimodal competing-risks prediction through a variational autoencoder with parametric time-to-event decoders. M2M-FSurv instead combines attention-based fusion with B-spline density heads, allowing locally varying event-time densities without prescribing a distribution family such as Weibull. This provides a complementary modeling approach, with flexibility controlled by the spline basis and knot placement. Our application additionally integrates clinical narratives, longitudinal physiological measurements, and scalar covariates under a 24-hour ICU landmark design.

Functional Representation and B-splines. Functional data analysis provides tools for representing trajectories through basis expansions (Cardot et al., 2003; Ramsay & Silverman, 2005). A related approach, FCRN, uses a learned basis layer to represent functional covariates for discrete-time competing-risks prediction, together with a missing-data imputation module (Gao et al., 2025). Our use of B-splines concerns the output distribution: M2M-FSurv represents cause-specific event-time DIFs through spline expansions, while encoding physiological trajectories as time-indexed input tokens. We use B-splines because their local support efficiently captures smooth, non-periodic, and locally varying DIFs. In contrast, other basis functions, such as Fourier expansions are naturally suited to periodic or globally oscillatory patterns.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

We adopt a 24-hour landmark design and include only subjects who remain event-free and under observation at the landmark. For subject i, let $W^{(i)} = (x^{(i)}, z^{(i)}(\cdot), u^{(i)})$ denote the multimodal input, comprising scalar covariates, physiological trajectories observed during the first 24 hours after ICU admission, and a clinical-note embedding. All inputs are restricted to information available by the landmark.

Let $T^{(i)}$ and $C^{(i)}$ denote the post-landmark event and censoring times, respectively, and let $J^{(i)} \in \{1, \dots, K\}$ denote the event type. For an analysis horizon τ, we observe $Y^{(i)} = \min(T^{(i)}, C^{(i)}, \tau)$ and $k^{(i)} = J^{(i)}$ if $T^{(i)} \leq \min(C^{(i)}, \tau)$, with $k^{(i)} = 0$ otherwise. The dataset is $\mathcal{D} = \{(\boldsymbol{W}^{(i)}, Y^{(i)}, k^{(i)})\}_{i=1}^{n}$.

Conditional on landmark eligibility, the cause-specific cumulative incidence function (CIF) (Kim, 2007) and its density incidence function (DIF) are

$$F_k(y \mid \boldsymbol{W}) = P(T \leq y, J = k \mid \boldsymbol{W}), \tag{1}$$

$$f_k(y \mid \boldsymbol{W}) = \frac{d}{dy} F_k(y \mid \boldsymbol{W}), \qquad 0 < y < \tau. \tag{2}$$

Here, y is measured in hours after the landmark. Our goal is to estimate these event-specific functions from the multimodal inputs.

3.2 MODEL DESCRIPTION

M2M-FSurv predicts competing ICU outcomes conditional on information available at a 24-hour landmark. The inputs comprise clinical-note summaries, hourly physiological measurements, and scalar covariates. The framework consists of three components: (i) modality-specific input encoding, (ii) a shared Transformer encoder for multimodal fusion, and (iii) cause-specific B-spline prediction heads. Figure 2 summarizes the architecture.

Clinical-text encoder. Clinical notes available within the observation window are processed using Kimi K2.6 with the prompt provided in Appendix A.1. The resulting text is encoded using Clinical Longformer. The token embeddings are projected into the shared embedding space and aggregated

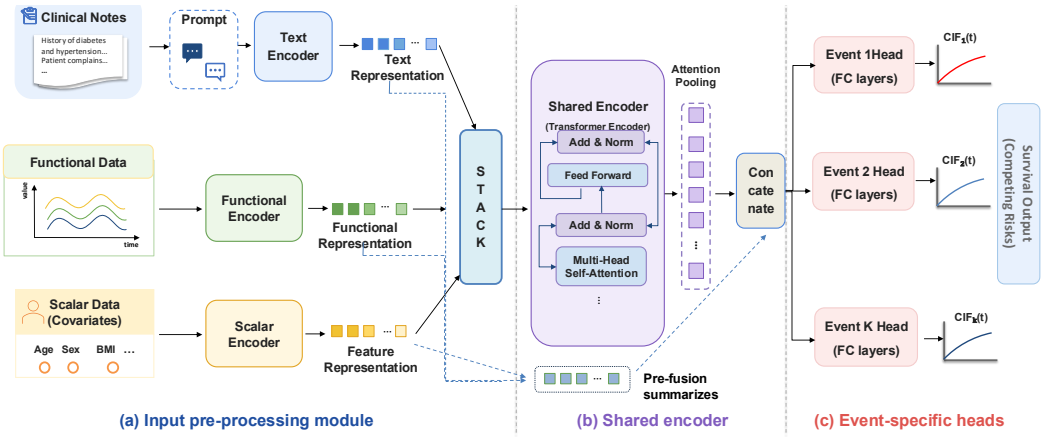


Figure 2: Detailed architecture of the proposed M2M-FSurv framework.

into a single text token using masked mean pooling. Encoding equations and implementation details are provided in Appendices A.2 and A.5, respectively.

Functional encoder. The five physiological signals are represented at 25 hourly positions indexed by $h = 0, \ldots, 24$. Let $z_h^{(i)} \in \mathbb{R}^5$ denote the processed measurements at position h. A linear projection maps these measurements to the shared embedding space, and a Time2Vec representation of the corresponding hour is added: $e_{z,h}^{(i)} = W_z z_h^{(i)} + b_z + \text{Time2Vec}(h) \in \mathbb{R}^d$. This representation preserves distinct hourly positions during multimodal fusion. All measurements used to construct these tokens are available by the landmark.

Scalar-covariate encoder. The processed scalar covariates $x^{(i)}$ are mapped to a single token, $e_x^{(i)} = \text{Enc}_x(x^{(i)}) \in \mathbb{R}^d$, where Enc_x denotes the scalar encoder. Its architecture and covariate preprocessing are described in Appendix A.2.

3.3 Model Framework

Multimodal fusion and skip connections. A shared Transformer encoder models interactions among 27 input tokens: one text token, 25 hourly functional covariate tokens, and one scalar covariate token. Learned attention pooling produces a patient-level representation, which is concatenated with three pre-fusion summaries: the text token, the mean of the functional covariate tokens, and the scalar covariate token. This concatenation provides direct modality-specific paths to each cause-specific head. Details are provided in Appendix A.2.

Cause-specific B-spline heads. Each cause-specific heads outputs spline-mixture logits. These are combined with jointly normalized event probabilities and basis-integral normalization to obtain non-negative coefficients $c_k(W)$ that satisfy the competing-risks probability constraints by construction. The complete parameterization is provided in Appendix A.5. Let $\theta(y)$ denote the B-spline basis on $[0, \tau]$ and $\phi(y) = \int_0^y \theta(v)\, dv$. The cause-specific DIF and CIF are

$$f_k(y \mid \boldsymbol{W}) = \theta(y)^\top c_k(\boldsymbol{W}), \tag{3}$$

$$F_k(y \mid \boldsymbol{W}) = \phi(y)^\top c_k(\boldsymbol{W}). \tag{4}$$

A valid competing-risks distribution requires $c_k(\boldsymbol{W}) \geq 0$ and $\sum_{k=1}^K \phi(\tau)^\top c_k(\boldsymbol{W}) \leq 1$, with event-free survival $S(y \mid \boldsymbol{W}) = 1 - \sum_{k=1}^K F_k(y \mid \boldsymbol{W})$. This representation yields nondecreasing CIFs that can be evaluated at any time within the post-landmark analysis horizon. Spline configurations and prediction-head dimensions are provided in Appendix A.5.

3.4 LOSS FUNCTION

We train the model by minimizing a total loss $\mathcal{L}_{\text{Total}}$, which combines a negative log-likelihood loss $\mathcal{L}_1$ that accounts for both observed and censored outcomes and a cause-specific ranking loss $\mathcal{L}_2$ that encourages correct risk ordering across subjects.

Using the B-spline representations in Eqs. (3)–(4), the cause-specific DIF and CIF for subject i are given by $f_k(y \mid \mathcal{W}^{(i)}) \approx \boldsymbol{\theta}(y)^\top \boldsymbol{c}_k^{(i)}$ and $F_k(y \mid \mathcal{W}^{(i)}) \approx \boldsymbol{\phi}(y)^\top \boldsymbol{c}_k^{(i)}$, respectively. Let $\mathbb{1}(\cdot)$ denote the indicator function. The negative log-likelihood is

$$\mathcal{L}_1 = -\sum_{i=1}^{n} \left[\mathbb{1}\left(k^{(i)} \neq 0\right) \log\left(\boldsymbol{\theta}(y^{(i)})^\top \boldsymbol{c}_{k^{(i)}}^{(i)}\right) + \mathbb{1}\left(k^{(i)} = 0\right) \log\left(1 - \sum_{k=1}^{K} \boldsymbol{\phi}(y^{(i)})^\top \boldsymbol{c}_k^{(i)}\right) \right].$$

The first term corresponds to subjects with an observed event $k^{(i)} \neq 0$ and evaluates the cause-specific density at the observed event time. The second term corresponds to right-censored subjects $k^{(i)} = 0$ and evaluates the probability of remaining event-free up to the censoring time $1 - \sum_{k=1}^{K} F_k(y^{(i)} \mid \mathcal{W}^{(i)})$. To encourage correct risk ordering, we add a concordance-style ranking loss $\mathcal{L}_2$ (Harrell et al., 1982) that penalizes, for each cause k, comparable pairs (i, j) where subject i experiences event k before j is censored or observed (full definition of comparable pairs in Appendix A.3):

$$\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \sum_{i \neq j} A_{k,ij}\, \eta\left(\boldsymbol{\phi}(y^{(i)})^\top \boldsymbol{c}_k^{(i)},\ \boldsymbol{\phi}(y^{(i)})^\top \boldsymbol{c}_k^{(j)}\right),$$

Here, $A_{k,ij} = \mathbf{1}(k^{(i)} = k, Y_i < Y_j)$, $\alpha_k = 1$, and $\eta(x, y) = \exp(-(x - y)/\rho)$. We minimize $\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \lambda \mathcal{L}_2$, selecting λ and $\rho > 0$ on the validation set. Further details are provided in Appendix A.3.

3.5 OPTIMIZATION

The prediction heads and normalization procedure jointly determine the B-spline coefficient vectors $c_k^{(i)}$. We optimize the trainable parameters using AdamW, with gradients computed by backpropagation, a mini-batch size of 16, and an initial learning rate of 5×10^{-5}. A plateau-based scheduler reduces the learning rate when the monitored validation criterion stops improving.

We assess cause-specific discrimination using the time-dependent concordance index C^{td} (Antolini et al., 2005). For event type k, we define

$$C_k^{td} = P\left(\widehat{F}_k(y^{(i)} \mid \mathcal{W}^{(i)}) > \widehat{F}_k(y^{(i)} \mid \mathcal{W}^{(j)}) \middle| D_i = k_i, y^{(i)} < y^{(j)}\right). \tag{5}$$

Here, $y^{(i)}$ denotes the observed post-landmark event or censoring time, and $k^{(i)}$ is the observed event label. Comparable pairs therefore consist of a subject experiencing event k and another subject who remains event-free and under observation beyond the first subject's event time.

We select the loss hyperparameters ρ and λ based on validation performance. Checkpoint selection combines discrimination and probability prediction error. Let $\overline{C}_{\text{val}}^{td}$ and $\overline{\text{IBS}}_{\text{val}}$ denote validation concordance and integrated Brier score aggregated across the competing outcomes. The selected checkpoint maximizes $V_{\text{val}} = \overline{C}_{\text{val}}^{td} - \overline{\text{IBS}}_{\text{val}}$. Early stopping is based on validation loss. Test-fold outcomes are not used for hyperparameter or checkpoint selection.

The number of B-spline basis functions M is determined by the knot configuration and spline degree. For MIMIC-III, non-uniform knot placement provides finer temporal resolution in the early post-landmark period and a coarser representation of the tail. The spline degree is fixed at $d_1 = 3$ for both datasets. Dataset-specific knot configurations and training details are provided in Appendix A.5.

4 THEORETICAL ANALYSIS

Following the approximation and consistency framework of Zhong et al. (2021), we study a feed-forward neural B-spline coefficient model. Let $f_0 = (f_{01}, \ldots, f_{0K})$ denote the true cause-specific

densities, and let

$$f_{g,k}(t \mid w) = \theta_{M_n}(t)^\top g_k(w), \qquad g \in \mathcal{G}_n^{\mathrm{adm}},$$

where θ_{M_n} is a nonnegative B-spline basis satisfying $\sum_m \theta_{M_n,m}(t) = 1$. The class $\mathcal{G}_n^{\mathrm{adm}}$ contains admissible coefficient networks, and $\mathcal{F}_n$ denotes the induced density class. Complete network and function-class definitions are provided in Appendix A.4. Let $\widehat{f}_n$ approximately minimize $\widehat{R}_n(f) + \lambda_n \widehat{Q}_n(f)$, up to optimization error ε_n, where $\widehat{R}_n$ is the average censored negative log-likelihood and $\widehat{Q}_n$ is the ranking loss normalized by $n(n-1)$. Define

$$d^2(f, f_0) = E_W \left[\sum_{k=1}^{K} \int_0^\tau \{f_k(t \mid \boldsymbol{W}) - f_{0k}(t \mid \boldsymbol{W})\}^2 \, dt \right].$$

All distributions refer to the population remaining event-free at the landmark.

4.1 Consistency results

Theorem 4.1 (Consistency of the feedforward B-spline estimator). Let $\widehat{f}_n$ be the approximate minimizer over the admissible feedforward coefficient class defined in Appendix A.4, with normalized ranking-loss weight $\lambda_n \to 0$. Under Assumptions 1–6 in Appendix A.4.2,

$$d(\widehat{f}_n, f_0) \xrightarrow{P} 0.$$

Furthermore,

$$E_W \left[\sum_{k=1}^{K} \sup_{t \in [0,\tau]} |\widehat{F}_{nk}(t \mid \boldsymbol{W}) - F_{0k}(t \mid \boldsymbol{W})|^2 + \sup_{t \in [0,\tau]} |\widehat{S}_n(t \mid \boldsymbol{W}) - S_0(t \mid \boldsymbol{W})|^2 \right] \xrightarrow{P} 0.$$

The proof combines approximation error, uniform likelihood convergence, and control of the ranking term; details are provided in Appendix A.4.3. The result concerns the specified feedforward coefficient class and an asymptotically negligible ranking term. It does not establish consistency for the full attention architecture or a fixed nonzero ranking weight. The theoretical coefficient class also differs from the implemented softmax-based parameterization in Appendix A.5. The theorem establishes conditional consistency for the stated admissible coefficient class; it does not directly establish consistency of the implemented encoder and probability head.

5 Experiments

We evaluate M2M-FSurv on two datasets with distinct prediction settings. MIMIC-III serves as the primary benchmark for multimodal competing-risks prediction, allowing us to assess predictive performance and examine the contributions of individual modalities and architectural components. METABRIC provides a complementary evaluation in a scalar-only, single-event setting.

5.1 Datasets

MIMIC-III. MIMIC-III contains ICU records from Beth Israel Deaconess Medical Center (2001–2012). We adopt a 24-hour landmark design and include patients who remain at risk 24 hours after ICU admission, with all multimodal inputs restricted to this observation window. We use physician, nursing, radiology, and other available clinical notes, with an average length of 19,203 tokens per patient. Repeated measurements of diastolic and systolic blood pressure, heart rate, Glasgow Coma Scale, and peripheral oxygen saturation are treated as functional covariates, while age, sex, ethnicity, and marital status are included as scalar covariates. We consider two mutually exclusive competing events: discharge alive from the ICU and in-ICU mortality. The cohort includes 12,332 patients, with two competing outcomes: discharge alive from the ICU (9,202) and in-ICU death (3,130).

METABRIC. We evaluate single-event survival prediction using 21 clinical covariates, including tumor characteristics and nodal status. Death is the event of interest, and patients alive at their last recorded follow-up are right-censored. The cohort comprises 1,981 patients, including 888 observed deaths and 1,093 censored observations. Missing continuous and categorical covariates are imputed using feature-wise means and modes, respectively.

5.2 Experimental Setup

Evaluation protocol. We employ patient-level 10-fold cross-validation, with all partitions constructed from the original cohort before oversampling. All ICU stays belonging to the same patient are assigned to the same fold. In each iteration, one fold is held out for testing, while 20% of the patients in the remaining nine folds are reserved for validation, model selection, and early stopping. Minority death-event oversampling is applied exclusively to the training subset after splitting. Validation and test subsets retain their original event distributions and contain no duplicates introduced by oversampling. All data-dependent preprocessing is fitted exclusively on the training subset and subsequently applied to the validation and test subsets. Performance is evaluated on each held-out test fold and summarized across the ten folds.

Implementation details. We use cubic B-splines ($d_1 = 3$) and select λ and ρ on the validation set. Architecture, optimization, knot configurations, and analysis horizons are detailed in Appendix A.5.

5.2.1 Baseline

SurvTRACE. To compare with a Transformer-based competing-risks model, we adapted SurvTRACE to the multimodal competing-risks setting using the same fixed-dimensional functional representations, clinical-note embeddings, and scalar covariates as M2M-FSurv. These inputs were concatenated and passed to SurvTRACE for cause-specific prediction. All methods used identical cross-validation folds, event definitions, and prediction horizons. Data-dependent preprocessing was fitted exclusively on each fold's training subset to prevent information leakage.

Concatenation Baseline. We additionally considered a straightforward multimodal concatenation baseline to assess whether the predictive improvement could be explained simply by combining information from multiple modalities. The same fixed-dimensional representation $\tilde{x}(i)$ used for the adapted SurvTRACE was provided to a MLP-based competing-risks model. This comparison separates the benefit of multimodal information itself from the additional benefit of Transformer-based feature interactions and the explicit functional modeling employed by our proposed method.

5.3 Empirical Results

5.3.1 MIMIC-III Results

We evaluate our method on MIMIC-III under a competing-risks setting with two mutually exclusive ICU outcomes: *discharge alive from the ICU* and *in-ICU mortality*. All methods are evaluated using the cause-specific C^{td}-index defined in Eq. (5), together with the *Integrated Brier Score* (IBS).

We additionally evaluate probabilistic prediction accuracy using the cause-specific IBS, computed over $[0, 5000]$ hours after the 24-hour landmark with uniform time weighting. Observed competing events are treated as negative outcomes for the cause of interest, while right censoring is handled using inverse probability of censoring weighting. All methods use the same evaluation interval and time grid. Lower IBS indicates better performance. Details are provided in Appendix A.6.

For **modality ablation study**, we assess each modality's contribution by comparing the full model (*Text* + *Functional* + *Scalar*) with three leave-one-modality-out variants: *Functional* + *Scalar*, *Text* + *Scalar*, and *Text* + *Functional*. For **component ablation study**, With all modalities fixed, we evaluate four variants: (I) removing Time2Vec; (II) replacing six-head attention with a single head; (III) removing the padding mask; and (IV) replacing Time2Vec with an LSTM over the 24-hour functional sequences. All other components, data partitions, and training settings remain unchanged.

Table 1 summarizes the results of both ablation studies alongside the baseline comparisons. On MIMIC-III, M2M-FSurv achieves the strongest discrimination among all methods compared, with a cause-specific C^{td}-index of 0.788 for discharge and 0.866 for death, alongside the lowest IBS (0.020 and 0.109, respectively), improving over both the concatenation-fusion baseline and SurvTRACE. We report 95% confidence intervals estimated using bootstrap resampling.

Comparison with baselines. The comparison with adapted SurvTRACE controls for differences in input feature construction by using the same clinical-note embeddings, functional covariates, and scalar covariates. Under this setup, M2M-FSurv achieves higher C^{td} for both discharge (0.788

Table 1: Comparison of cause-specific C^{td}-index and IBS on the MIMIC-III using inputs observed during the first 24 hours after ICU admission. C^{td}-index values are reported as mean (95% CI).

Model / Experiment	**Input / Modification**	**C^{td}-index ↑**		**IBS ↓**	
		Discharge	**Death**	**Discharge**	**Death**
Baseline Models					
Concatenation Fusion	T + F + S	0.709 (0.696, 0.723)	0.667 (0.615, 0.717)	0.024	0.223
SurvTRACE	T + F + S	0.766 (0.751, 0.779)	0.801 (0.778, 0.823)	0.044	0.165
FCRN	T + F + S	0.742 (0.730 , 0.758)	0.703 (0.689, 0.722)	0.027	0.198
Proposed Model					
Full Model	T + F + S	**0.788 (0.774, 0.802)**	**0.866 (0.821, 0.902)**	**0.020**	**0.109**
Modality Ablation					
Modality I	F + S	0.698 (0.651, 0.742)	0.694 (0.677, 0.711)	0.026	0.201
Modality II	T + S	0.721 (0.710, 0.735)	0.716 (0.681–0.751)	0.023	0.220
Modality III	T + F	0.768 (0.733, 0.801)	0.744 (0.729, 0.758)	0.021	0.254
Component Ablation					
Component I	No Time Encoding	0.778 (0.764, 0.792)	0.656 (0.605, 0.703)	0.024	0.207
Component II	Single-head Attention	0.768 (0.753, 0.782)	0.683 (0.635, 0.727)	0.022	0.238
Component III	No Padding Mask	0.781 (0.767, 0.794)	0.700 (0.641, 0.753)	0.021	0.255
Component IV	LSTM functional encoder	0.781 (0.768, 0.795)	0.685 (0.633, 0.733)	0.025	0.198

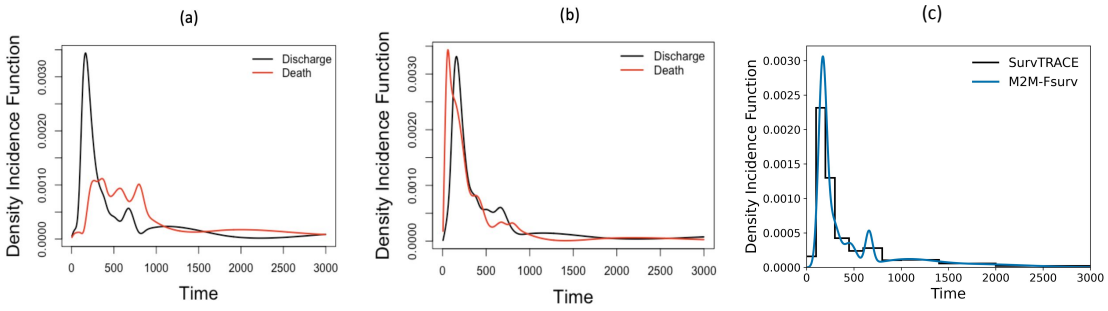


Figure 3: Cause-specific density incidence functions (DIFs). (a, b) Estimated DIFs for discharge and death for two representative patients. (c) Estimated discharge DIFs for a selected patient, comparing SurvTRACE and M2M-Fsurv.

versus 0.766) and death (0.866 versus 0.801), together with lower IBS (0.020 versus 0.044 for discharge; 0.109 versus 0.165 for death). These results support the performance of the proposed framework under matched input representations, without attributing the gains to any individual architectural component.

As shown in Fig. 3(c), M2M-FSurv also differs from SurvTRACE in output form: instead of predicting event probabilities over a fixed discrete time grid, it parameterizes the cause-specific DIF directly in continuous time via the B-spline expansion, yielding smooth, closed-form incidence and cumulative incidence curves at arbitrary post-landmark times rather than only at discretized grid time points.

Component ablations. Mortality discrimination is more sensitive to the evaluated architectural changes than discharge discrimination. Removing Time2Vec yields the largest decrease in death concordance, from 0.866 to 0.656. Using a single attention head, removing the padding mask, and replacing the functional encoder with an LSTM reduce death concordance to 0.683, 0.700, and 0.685, respectively. In contrast, discharge concordance remains between 0.768 and 0.781, compared with 0.788 for the full model. Removing the padding mask produces the highest death IBS among these variants (0.255). These results suggest that the effects of the evaluated components differ across outcomes.

Modality ablations. Removing each input modality reduces discrimination relative to the full model, with effects differing across outcomes. These comparisons assess the contribution of each input representation and its processing pathway. Because clinical notes may also contain physiological measurements and demographic information, the ablations do not isolate mutually exclusive sources of clinical information or establish that the modalities are nonredundant. Each ablated modality is removed from both the fusion input and all modality-specific skip connections.

Qualitative analysis A key feature of our approach is its ability to represent continuous-time DIFs using B-spline expansions. Fig. 3(a) and (b) illustrate the predicted DIFs for two representative patients. In Fig. 3(a), the discharge DIF peaks earlier after the 24-hour landmark, whereas the death DIF is more dispersed over time. In Fig. 3(b), the death DIF is concentrated near the landmark and subsequently declines. These examples illustrate patient-specific differences in the predicted timing of competing events and provide a continuous-time visualization of their temporal patterns. Under the reported evaluation protocol, the full model achieves the lowest IBS among the evaluated methods for both outcomes. These results measure integrated probability prediction error over the specified horizon; they do not establish calibration at individual time points or explain the difference between discharge and mortality IBS.

5.3.2 METABRIC Results

To compare with methods that do not support multimodal inputs, we additionally evaluate M2M-FSurv on METABRIC in a scalar-only, single-event setting. Baselines include Cox (Therneau & Lumley, 2015), ThresReg (Lee & Whitmore, 2006), RSF (Ishwaran et al., 2014), MP-RForest, MP-LogitR, MP-AdaBoost, DeepSurv (Katzman et al., 2018), DeepHit (Lee et al., 2018), and SurvTRACE. We report the performance of M2M-FSurv and the baseline methods on METABRIC in Table 2. M2M-FSurv achieves the highest mean C^{td} among the evaluated methods on METABRIC (0.707 versus 0.701 for SurvTRACE). This experiment provides an additional assessment of the framework in a scalar-only, single-event setting.

Table 2: Comparison of cause-specific C^{td}-index performance (mean and 95% CI) on METABRIC in the single event setting.

Algorithms	**Single Event**	**Algorithms**	**Single Event**
Cox	0.662 (0.643–0.681)	MP-LogitR	0.619 (0.594–0.635)
RSF	0.678 (0.673–0.687)	DeepSurv	0.648 (0.636–0.660)
ThresReg	0.647 (0.641–0.665)	DeepHit	0.690 (0.676–0.701)
MP-RForest	0.679 (0.673–0.684)	SurvTRACE	0.701 (0.686–0.710)
MP-AdaBoost	0.643 (0.627–0.662)	M2M-FSurv	**0.707 (0.700–0.714)**

6 Conclusion

We presented M2M-FSurv, a multimodal framework that combines clinical text, hourly physiological measurements, and scalar covariates for continuous-time competing-risks prediction. The model integrates these inputs through a shared Transformer encoder and represents cause-specific event-time subdensities using B-spline. We established consistency for a constrained feedforward B-spline estimator with an asymptotically vanishing ranking weight. Experiments on MIMIC-III showed improved discrimination and lower IBS relative to the evaluated baselines, while METABRIC provided an additional single-event evaluation. The framework also supports visualization of patient-specific event-time patterns. External validation and assessment of clinical utility remain directions for future work.

Limitations. We do not isolate spline-head effects or establish time-specific calibration and clinical utility. Multimodal competing-risks evaluation is limited to MIMIC-III; METABRIC covers scalar-only, single-event prediction. Consistency holds for a restricted coefficient class, not the full implemented model.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing and code development. Specifically, generative AI was used to improve the clarity, grammar, and readability of the manuscript, as well as to assist with code writing, debugging, and refinement. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was independently inspected, tested, and validated for correctness before being used in the experiments. Generative AI was not used to generate research results, fabricate data, or make scientific conclusions. The authors take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the assistance of generative AI.

REFERENCES

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–591, 2003.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Jianghu Dong, Haolun Shi, Liangliang Wang, Ying Zhang, and Jiguo Cao. Jointly modelling multiple transplant outcomes by a competing risk model via functional principal component analysis. *Journal of Applied Statistics*, 50(1):43–59, 2023.

Jiahui Feng, Haolun Shi, Da Ma, Mirza Faisal Beg, and Jiguo Cao. Supervised functional principal component analysis under the mixture cure rate model: An application to alzheimer's disease. *Statistics in Medicine*, 44(3-4):e10324, 2025.

Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

Penglei Gao, Yan Zou, Abhijit Duggal, Shuaiqi Huang, Faming Liang, and Xiaofeng Wang. Deep survival analysis for competing risk modeling with functional covariates and missing data imputation. *arXiv preprint arXiv:2509.25381*, 2025.

Alba Garrido, Alejandro Almodóvar, Patricia A. Apellániz, Juan Parras, and Santiago Zazo. Deep survival analysis in multimodal medical data: A parametric and probabilistic approach with competing risks, 2025. URL https://arxiv.org/abs/2507.07804.

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.

Hemant Ishwaran, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange, and Bryan M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 04 2014.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

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(1), February 2018.

Haesook T Kim. Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical cancer research*, 13(2):559–565, 2007.

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.

Mei-Ling Ting Lee and G. A. Whitmore. Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statistical Science*, 21(4), nov 2006.

Lu Mao and DY Lin. Efficient estimation of semiparametric transformation models for the cumulative incidence of competing risks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):573–587, 2017.

James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2005.

Terry M. Therneau and Thomas Lumley. *A Package for Survival Analysis in R*, 2015. R package version 2.38.

Zifeng Wang and Jimeng Sun. Survtrace: Transformers for survival analysis with competing events. In *Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics*, pp. 1–9, 2022.

Qixian Zhong, Jonas W Mueller, and Jane-Ling Wang. Deep extended hazard models for survival analysis. *Advances in Neural Information Processing Systems*, 34:15111–15124, 2021.

A Appendix

A.1 Prompt

To convert lengthy, noisy ICU notes into structured clinical narratives, we developed a task-specific prompt that instructs the large language model to preserve all documented clinical information while improving readability. The prompt emphasizes factual fidelity, exact retention of numerical values and timestamps, and preservation of distinct clinical events. The prompt template used in our experiments is provided below.

You are a clinical documentation specialist. Your task is to transform the ICU note below into a comprehensive, factual medical narrative written in continuous prose.

Preserve every documented clinical detail. Do not summarize, interpret, or omit information.

=== MANDATORY INSTRUCTIONS ===

1. CLINICAL CONTENT

Include all of the following, when present:

- The complete admission physical examination, including vital signs and HEENT, cardiovascular, pulmonary, abdominal, and neurological findings.

- All follow-up examination findings.

- Every laboratory result, including its numerical value, unit, and recorded timestamp.

- All vital signs at every recorded time.

- All medications, including documented doses, routes, frequencies, and infusion rates.

- All procedures, including intubation, dialysis, vascular access placement, and surgery.

2. DATA PRESERVATION

- Preserve all numerical values exactly as documented. Never replace specific values with descriptions such as ”normal” or ”stable.”

- Retain all available timestamps for laboratory results, vital signs, and clinical events. Group information by time when possible, while preserving distinct events.

- Expand clinical shorthand into fluent prose without changing its meaning. For example, convert ”bp: 166/87” to ”blood pressure was 166/87.” Include units only when documented.

- Preserve repeated observations and values at each documented occurrence; do not remove them as redundant.

- Do not add information that is absent from the source note.

3. OUTPUT FORMAT

- Output only plain text in continuous prose, without bullet points, headings, Markdown formatting, asterisks, or paragraph breaks.

- Produce a detailed and comprehensive narrative. Aim for 1,800–2,200 tokens when supported by the source material; do not add unsupported content or unnecessary repetition to meet a length target.

- Do not shorten the narrative at the expense of clinical detail.

- Use precise clinical terminology while preserving the meaning of the original documentation.

4. PROHIBITED TRANSFORMATIONS

- Do not infer diagnoses or findings that are not explicitly documented.

- Do not omit seemingly minor findings, such as "no yeast" or "occasional ovalocytes."

- Do not compress timelines or merge distinct clinical events.

- Do not resolve ambiguities or inconsistencies by making assumptions.

=== ICU NOTE ===

[Insert ICU note here]

A.2 INPUT ENCODING AND MULTIMODAL FUSION

Modality-specific input encoding. The shared embedding dimension is $d = 252$. For patient i, let $U_\ell^{(i)}$ denote the Clinical Longformer representation of token ℓ, and let $m_\ell^{(i)}$ indicate a non-padding token. Projection and masked mean pooling yield

$$e_u^{(i)} = \frac{\sum_{\ell=1}^{L_i} m_\ell^{(i)} \left(W_u U_\ell^{(i)} + b_u \right)}{\sum_{\ell=1}^{L_i} m_\ell^{(i)}} \in \mathbb{R}^d.$$

The physiological measurements at hourly position h are encoded as

$$e_{z,h}^{(i)} = W_z z_h^{(i)} + b_z + \mathrm{Time2Vec}(h), \qquad h = 0, \ldots, 24.$$

Scalar covariates are encoded as

$$e_x^{(i)} = \mathrm{Enc}_x(x^{(i)}) \in \mathbb{R}^d.$$

Shared encoder and attention pooling. The input sequence is

$$E^{(i)} = \left[e_u^{(i)}, e_{z,0}^{(i)}, \ldots, e_{z,24}^{(i)}, e_x^{(i)} \right]^\top \in \mathbb{R}^{27 \times 252}.$$

A single Transformer encoder layer with six attention heads and residual connections contextualizes these tokens. Each head has dimension $d_{\mathrm{head}} = 42$.

Suppressing the patient index for clarity, the queries, keys, and values for attention head a are

$$Q_a = EW_{Q,a}, \qquad K_a = EW_{K,a}, \qquad V_a = EW_{V,a}.$$

The attention output is

$$A_a = \mathrm{softmax}\left(\frac{Q_a K_a^\top}{\sqrt{d_{\mathrm{head}}}} + M \right) V_a,$$

where M is the additive attention mask, with zero entries at allowed positions and $-\infty$ at masked positions. When no positions are masked, $M = 0$. The head outputs are concatenated and projected:

$$\mathrm{MHA}(E) = \mathrm{Concat}(A_1, \ldots, A_6) W_O.$$

The Transformer encoder produces contextualized representations $H = \mathrm{TransformerEncoder}(E) \in \mathbb{R}^{27 \times 252}$.

A learned scoring function s_ω determines the attention-pooling weights:

$$a_\ell = \frac{\exp(s_\omega(H_\ell))}{\sum_{j=1}^{27} \exp(s_\omega(H_j))}, \qquad r = \sum_{\ell=1}^{27} a_\ell H_\ell.$$

When fusion-token masking is applied to pooling, normalization and aggregation are restricted to valid positions.

Modality-summary skip connections. The pre-fusion summaries are

$$s_u^{(i)} = e_u^{(i)}, \qquad s_z^{(i)} = \frac{1}{25} \sum_{h=0}^{24} e_{z,h}^{(i)}, \qquad s_x^{(i)} = e_x^{(i)}.$$

Each cause-specific head receives

$$q^{(i)} = \mathrm{Concat}\left(r^{(i)}, s_u^{(i)}, s_z^{(i)}, s_x^{(i)}\right) \in \mathbb{R}^{1008}.$$

Prediction-head configurations are provided in Appendix A.5.

A.3 Ranking Loss Construction

In $\mathcal{L}_2$, for each event type k, we consider comparable pairs (i, j) with $k^{(i)} = k$ and $y^{(i)} < y^{(j)}$, so that subject j has not experienced any event (nor been censored) by time $y^{(i)}$. Define $A_{k,ij} = \mathbb{1}\left(k^{(i)} = k,\ y^{(i)} < y^{(j)}\right)$. For each such pair, we compare the predicted CIF for cause k at the reference time $y^{(i)}$. Ideally, the model should assign a larger $F_k(y^{(i)} \mid \mathcal{W}^{(i)})$ than $F_k(y^{(i)} \mid \mathcal{W}^{(j)})$, since by time $y^{(i)}$ subject i has already experienced event k whereas subject j has not. We therefore apply a convex pairwise loss $\eta(\cdot, \cdot)$ to penalize violations of this ordering, and define

$$\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \sum_{i \neq j} A_{k,ij} \eta\left(\boldsymbol{\phi}(y^{(i)})^\top \boldsymbol{c}_k^{(i)}, \boldsymbol{\phi}(y^{(i)})^\top \boldsymbol{c}_k^{(j)}\right),$$

where α_k are hyperparameters. For simplicity, we set $\alpha_k = 1$ for $k = 1, \ldots, K$. We use $\eta(x, y) = \exp(-(x - y)/\rho)$, where $\rho > 0$ is hyperparameter

A.4 PROOFS OF THEORETICAL RESULTS

A.4.1 Statistical setting and neural function classes

All distributions are defined within the population of subjects who remain event-free at the landmark. Let $W \in \mathcal{W} \subseteq [0, 1]^p$ denote the multimodal input available at the landmark, T the subsequent event time, $J \in \{1, \ldots, K\}$ the event type, and C the censoring time. For a fixed analysis horizon $\tau > 0$, define

$$Y = \min(T, C, \tau), \qquad D = \begin{cases} J, & T \leq C,\ T \leq \tau, \\ 0, & \text{otherwise.} \end{cases}$$

The observed sample is $\{O_i = (W_i, Y_i, D_i)\}_{i=1}^n$. For each cause k, let

$$F_{0k}(t \mid w) = P(T \leq t, J = k \mid W = w) = \int_0^t f_{0k}(u \mid w)\, du.$$

Write $f_0 = (f_{01}, \ldots, f_{0K})$. The corresponding survival function is $S_0(t \mid w) = 1 - \sum_{k=1}^K F_{0k}(t \mid w)$. Consider a feedforward neural network with P hidden layers and width vector $\boldsymbol{d}^* = (d_0^*, \ldots, d_{P+1}^*), d_0^* = p, d_{P+1}^* = KM$. Its hidden representations and output are $h_0(w) = \sigma(W_0 w + b_0)$, $h_\ell(w) = \sigma(W_\ell h_{\ell-1}(w) + b_\ell), \ell = 1, \ldots, P-1$, $g(w) = W_P h_{P-1}(w) + b_P$, where σ is the componentwise ReLU activation, $W_\ell \in \mathbb{R}^{d_{\ell+1}^* \times d_\ell^*}$, $b_\ell \in \mathbb{R}^{d_{\ell+1}^*}, \ell = 0, \ldots, P$. Following the network-class formulation used by Zhong et al. (2021), define

$$\mathcal{G}(P, \boldsymbol{d}^*) = \left\{ g : \max_{0 \leq \ell \leq P} \left(\|W_\ell\|_{\max} \vee \|b_\ell\|_\infty\right) \leq 1 \right\},$$

where $\|W\|_{\max} = \max_{i,j} |W_{ij}|$. We further restrict network sparsity and output magnitude:

$$\mathcal{G}(P, \boldsymbol{d}^*, s, A) = \left\{ g \in \mathcal{G}(P, \boldsymbol{d}^*) : \sum_{\ell=0}^{P} (\|W_\ell\|_0 + \|b_\ell\|_0) \le s, \quad \sup_{w \in \mathcal{W}} \|g(w)\|_\infty \le A \right\}.$$

Here, $\|\cdot\|_0$ counts nonzero entries. The parameters P, $\boldsymbol{d}^*$, and s may increase with the sample size. Let $\theta_M(t)$ contain M nonnegative B-spline basis functions satisfying $\sum_{m=1}^{M} \theta_{Mm}(t) = 1$. Partition $g(w)$ into K coefficient vectors $g_k(w) \in \mathbb{R}^M$, and define $f_{g,k}(t \mid w) = \theta_M(t)^\top g_k(w)$. The admissible network class $\mathcal{G}_n^{\mathrm{adm}}$ is the subset of

$$\mathcal{G}(P_n, \boldsymbol{d}_n^*, s_n, A)$$

whose coefficients are nonnegative and whose induced densities satisfy the probability constraints and bounds specified below. Let $\mathcal{F}_n = \{f_g : g \in \mathcal{G}_n^{\mathrm{adm}}\}$. This restriction defines the theoretical parameter space; it does not imply that unconstrained network training automatically satisfies these conditions.

A.4.2 Assumptions

Assumption 1. The observations are iid. Conditional on W, C is independent of (T, J). There exists $g_* > 0$ such that $P(C \ge t \mid W = w) \ge g_*$ for $t \in [0, \tau]$ and almost every w. Event-time distributions are absolutely continuous on this interval.

Assumption 2. The true density vector and every member of $\mathcal{F}_n$ satisfy $a \le f_k(t \mid w) \le b, \qquad S_f(\tau \mid w) \ge s_*$, for constants $0 < a < b < \infty$ and $s_* > 0$ independent of n. In particular, all candidates define valid competing-risks distributions. These bounds restrict the analysis to a horizon on which the cause-specific densities and survival probabilities remain bounded away from zero.

Assumption 3. There exist coefficient maps $c_M^\star : \mathcal{W} \to \mathbb{R}^{KM}$, constants $C_{\mathrm{sp}} > 0$ and $\beta > 0$, such that

$$\max_k \sup_{t,w} \left| f_{0k}(t \mid w) - \theta_M(t)^\top c_{M,k}^\star(w) \right| \le C_{\mathrm{sp}} M^{-\beta}.$$

Each coordinate of $c_M^\star$ belongs to a composite Hölder class $\mathcal{H}(L, \boldsymbol{p}, \boldsymbol{p}^*, \boldsymbol{\zeta}, B)$ with bounds uniform in M. Specifically, each coordinate admits a composition $h_L \circ \cdots \circ h_0$, where the component functions at level ℓ have Hölder smoothness ζ_ℓ and depend on at most p_ℓ^* input coordinates. The coefficient-map regularity is assumed separately from smoothness in time.

Assumption 4. There exist $g_n^\star \in \mathcal{G}_n^{\mathrm{adm}}$ such that

$$\max_{k,m} \sup_w |g_{n,km}^\star(w) - c_{M_n,km}^\star(w)| \le \eta_n,$$

where $M_n \to \infty, \eta_n \to 0$.

Thus, the approximation must remain within the admissible probability model. An unconstrained approximation result alone does not establish this assumption.

Assumption 5. The network classes and B-spline bases are deterministic. With

$$\|g - \widetilde{g}\|_{\infty,\mathrm{coef}} = \max_{k,m} \sup_w |g_{km}(w) - \widetilde{g}_{km}(w)|,$$

the admissible classes satisfy $\log N\left(n^{-1}, \mathcal{G}_n^{\mathrm{adm}}, \|\cdot\|_{\infty,\mathrm{coef}}\right) = o(n)$.

Assumption 6. The number of causes K is fixed, $0 \le \alpha_k \le \bar{\alpha}, \rho_n \ge \rho_* > 0, , \lambda_n \ge 0, , \lambda_n \xrightarrow{P} 0$, and $\varepsilon_n \xrightarrow{P} 0$.

Assumptions 3–5 separate spline approximation, constrained neural approximation, and statistical complexity. Their verification must account for the increasing output dimension KM_n. The composite Hölder structure motivates the approximation analysis but does not, by itself, establish admissible approximation or the required covering-number bound.

A.4.3 PROOF OF THEOREM 4.1

Let $f_n^\star = f_{g_n^\star}$. The partition-of-unity property gives

$$\|f_n^\star - f_0\|_\infty \le \eta_n + C_{\mathrm{sp}} M_n^{-\beta} \to 0.$$

Furthermore,

$$\|S_f - S_{\widetilde{f}}\|_\infty \le K\tau \|f - \widetilde{f}\|_\infty.$$

Under Assumption 2, the logarithm is uniformly Lipschitz on the relevant ranges. Hence

$$R(f_n^\star) - R(f_0) \to 0.$$

Define

$$U_n = \sup_{f \in \mathcal{F}_n} |\widehat{R}_n(f) - R(f)|.$$

The losses are uniformly bounded, and

$$\|f_g - f_{\widetilde{g}}\|_\infty \le \|g - \widetilde{g}\|_{\infty,\mathrm{coef}}.$$

A finite-cover concentration argument, using Assumption 5 and the Lipschitz property of the losses, therefore yields

$$U_n = o_P(1).$$

Because every CIF lies in $[0, 1]$,

$$0 \le \widehat{Q}_n(f) \le K\bar{\alpha} e^{1/\rho_*} =: B_Q.$$

Approximate minimization implies

$$\widehat{R}_n(\widehat{f}_n) \le \widehat{R}_n(f_n^\star) + \lambda_n B_Q + \varepsilon_n.$$

Consequently,

$$R(\widehat{f}_n) - R(f_0) \le 2U_n + \{R(f_n^\star) - R(f_0)\} + \lambda_n B_Q + \varepsilon_n = o_P(1).$$

It remains to relate excess risk to density error. Let P_f^{obs} denote the observed-data distribution induced by f, with the true covariate and censoring distributions held fixed. Conditional independent censoring gives

$$R(f) - R(f_0) = \mathrm{KL}(P_0^{\mathrm{obs}} \| P_f^{\mathrm{obs}}).$$

Using squared Hellinger distance

$$H^2(P, Q) = \int (\sqrt{dP} - \sqrt{dQ})^2,$$

we have $H^2(P, Q) \le \mathrm{KL}(P\|Q)$. The observed-event density for cause k, conditional on $W = w$, is

$$G(t \mid w) f_k(t \mid w), \qquad G(t \mid w) = P(C \ge t \mid W = w).$$

Therefore,

$$H^2(P_0^{\mathrm{obs}}, P_f^{\mathrm{obs}}) \ge g_* E_W \left[\sum_k \int_0^\tau (\sqrt{f_k} - \sqrt{f_{0k}})^2 \, dt \right].$$

Since $f_k, f_{0k} \le b$,

$$(f_k - f_{0k})^2 \le 4b(\sqrt{f_k} - \sqrt{f_{0k}})^2.$$

Thus,

$$d^2(f, f_0) \le \frac{4b}{g_*} \{R(f) - R(f_0)\}.$$

Applying this inequality to $\widehat{f}_n$ proves the theorem. By Cauchy–Schwarz,

$$\sup_{t \in [0,\tau]} |\widehat{F}_{nk}(t \mid w) - F_{0k}(t \mid w)|^2 \le \tau \int_0^\tau |\widehat{f}_{nk}(u \mid w) - f_{0k}(u \mid w)|^2 \, du.$$

Summing over causes and integrating over W,

$$E_W\left[\sum_k \sup_{t\in[0,\tau]} |\widehat{F}_{nk} - F_{0k}|^2\right] \le \tau d^2(\widehat{f}_n, f_0) \xrightarrow{P} 0.$$

Finally,

$$\widehat{S}_n - S_0 = -\sum_k (\widehat{F}_{nk} - F_{0k}),$$

so

$$E_W\left[\sup_{t\in[0,\tau]} |\widehat{S}_n - S_0|^2\right] \le K\tau d^2(\widehat{f}_n, f_0) \xrightarrow{P} 0.$$

This proves both conclusion.

A.5 Hyperparameters and Training Details

Input representations and shared encoder. M2M-FSurv uses a shared embedding dimension of 252. Clinical-note summaries are encoded using Clinical Longformer, projected to the shared embedding space, and aggregated through masked mean pooling into a single text token. The five physiological signals are represented at 25 hourly positions within the 24-hour observation window. Each position is encoded using a linear projection of the measurements plus a Time2Vec embedding of the corresponding hour. Scalar covariates are encoded into one additional token. The resulting sequence contains 27 tokens: one text token, 25 physiological tokens, and one scalar token.

The shared encoder consists of a single Transformer encoder layer with six attention heads and residual connections. Learned attention pooling aggregates the contextualized tokens into a fused patient representation. Each cause-specific prediction head receives the fused representation concatenated with the three modality-specific summaries. The head architecture and probability normalization are described below.

Optimization and model selection. The model is trained using AdamW with a batch size of 16 and an initial learning rate of 5×10^{-5}. A plateau-based learning-rate scheduler is used during training. The training objective combines the censored negative log-likelihood and a cause-specific ranking loss, without inverse-event-frequency class weights. Checkpoints are selected using validation time-dependent concordance minus integrated Brier score. All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 4060 GPU.

Cross-validation and oversampling Cross-validation partitions were constructed from the original cohort before oversampling, with all records from each patient assigned exclusively to the training, validation, or test subset within each iteration. Minority-event oversampling was performed only within the training subset. Validation and test data were neither oversampled nor otherwise rebalanced. Each original test observation was evaluated once in its held-out fold. The oversampling procedure and sampling ratio are reported below.

B-spline configuration. The B-spline degree is fixed at $d_1 = 3$ for both datasets. For METABRIC, we use 41 equally spaced internal knots over the observed time domain. For MIMIC-III, survival times are expressed relative to the 24-hour landmark, with $Y = 0$ corresponding to 24 hours after ICU admission. To allocate greater temporal resolution to the early post-landmark period, we use a non-uniform knot configuration: 20 internal knots within $[0, 1000]$ hours, 6 within $[1000, 4000]$ hours, and 2 within $[4000, 5000]$ hours. The knot locations and analysis horizon were prespecified and held fixed across all cross-validation folds. This configuration provides a finer representation of early event-time variation and a coarser representation of the sparsely observed tail.

Cause-specific heads and probability normalization. Let $q(W)$ denote the patient representation obtained by concatenating the fused representation and the three modality-specific summaries. Each cause-specific MLP maps $q(W)$ to M mixture logits $a_k(W) \in \mathbb{R}^M$, where M is the number of B-spline basis functions. Each MLP contains a hidden layer with 128 units and a ReLU activation, followed by a linear output layer. For the MIMIC-III configuration, $M = 32$. An additional linear

layer maps $q(W)$ to $K+1$ logits $b(W) \in \mathbb{R}^{K+1}$, representing the K competing events within the analysis horizon and the residual event-free probability.

Let $\theta_m(t)$ be the mth nonnegative B-spline basis function on $[0, \tau]$, and define

$$A_m = \int_0^\tau \theta_m(u)\, du = \phi_m(\tau), \qquad \bar{\theta}_m(t) = \frac{\theta_m(t)}{A_m}, \qquad \bar{\phi}_m(t) = \frac{\phi_m(t)}{A_m},$$

where every retained basis function has $A_m > 0$. Consequently, $\bar{\theta}_m$ integrates to one over $[0, \tau]$, and $\bar{\phi}_m(\tau) = 1$.

We normalize the event logits jointly across the $K+1$ categories:

$$\pi_j(W) = \frac{\exp(b_j(W))}{\sum_{r=0}^{K} \exp(b_r(W))}, \qquad j = 0, \ldots, K.$$

For each cause, the mixture logits are normalized across the M basis functions:

$$w_{km}(W) = \frac{\exp(a_{km}(W))}{\sum_{\ell=1}^{M} \exp(a_{k\ell}(W))}, \qquad m = 1, \ldots, M.$$

Here, $\pi_k(W)$ is the probability of experiencing event k by τ, while $\pi_0(W)$ is the probability of remaining event-free beyond τ. The weights $w_{km}(W)$ specify the conditional event-time density given that event k occurs within the analysis horizon.

The cause-specific density and cumulative incidence functions are

$$f_k(t \mid W) = \pi_k(W) \sum_{m=1}^{M} w_{km}(W) \bar{\theta}_m(t),$$

$$F_k(t \mid W) = \pi_k(W) \sum_{m=1}^{M} w_{km}(W) \bar{\phi}_m(t).$$

This parameterization is equivalent to Eqs. (3)–(4), with

$$c_{km}(W) = \frac{\pi_k(W) w_{km}(W)}{A_m}.$$

Validity of the competing-risks distribution. Because the basis functions and mixture weights are nonnegative, $f_k(t \mid W) \geq 0$, and every $F_k(t \mid W)$ is nondecreasing. Moreover, for all $t \in [0, \tau]$,

$$0 \leq F_k(t \mid W) \leq \pi_k(W),$$

and therefore

$$\sum_{k=1}^{K} F_k(t \mid W) \leq \sum_{k=1}^{K} \pi_k(W) = 1 - \pi_0(W) \leq 1.$$

Thus,

$$S(t \mid W) = 1 - \sum_{k=1}^{K} F_k(t \mid W)$$

is nonnegative and nonincreasing, with $S(0 \mid W) = 1$ and $S(\tau \mid W) = \pi_0(W)$. The distributional constraints in Section 3.3 therefore hold by construction, without a separate constraint penalty.

The residual mass $\pi_0(W)$ represents survival beyond the finite analysis horizon; it is not a cure probability. The model specifies event-time subdensities on $[0, \tau]$ and does not specify the distribution of event times beyond τ.

A.6 Integrated Brier Score

Let $Y_i = \min(T_i, C_i)$ denote the observed follow-up time and $D_i \in \{0, 1, \ldots, K\}$ the observed event label, where $D_i = 0$ indicates censoring. For cause k, the time-specific Brier score is

$$\widehat{\mathrm{BS}}_k(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbf{1}(Y_i \le t, D_i \neq 0)}{\widehat{G}(Y_i^-)} \left\{ \mathbf{1}(D_i = k) - \widehat{F}_k(t \mid W_i) \right\}^2 + \frac{\mathbf{1}(Y_i > t)}{\widehat{G}(t)} \widehat{F}_k(t \mid W_i)^2 \right],$$

where $\widehat{G}(t)$ estimates the censoring survival function $P(C > t)$. An observed competing event contributes a target value of zero for cause k; it is not treated as censoring. When censoring is absent over the evaluation interval, the weights reduce to one. We compute

$$\widehat{\mathrm{IBS}}_k = \frac{1}{\tau - t_{\min}} \int_{t_{\min}}^{\tau} \widehat{\mathrm{BS}}_k(t)\, dt$$

For MIMIC-III, evaluation times are measured from the 24-hour landmark.