



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

ChronoSurv: A Clinical Pathway-Guided Graph Framework for Multimodal Survival Analysis

Hugo Miccinilli^{1,*} and Theo Di Piazza^{2,*}

¹Université Paris-Saclay, CentraleSupélec, MICS, France

²University of Lyon, INSA Lyon, CREATIS, France

*Equal contribution

{hugo.miccinilli@centralesupelec.fr, theo.dipiazza@creatis.insa-lyon.fr}

Abstract. Accurate survival prediction is essential for personalized treatment planning in head and neck cancer, yet remains challenging due to the heterogeneous and high-dimensional nature of multimodal clinical data. While deep survival models have improved predictive performance over classical statistical approaches, existing methods typically rely on static fusion strategies or temporally agnostic modeling, limiting their ability to capture structured clinical workflows. In this work, we propose ChronoSurv, a heterogeneous hierarchical directed graph framework for multimodal survival analysis. ChronoSurv represents patient care as a progression-aware clinical trajectory using directed graphs aligned with key diagnostic steps. A hierarchical topology incorporates fine-grained, coarse, and global representations, further supporting flexible adaptation to missing modalities, while heterogeneous message passing models complex and asymmetric relationships across modalities and clinical steps. Experimental results on two public datasets demonstrate that ChronoSurv achieves state-of-the-art discriminative performance while maintaining statistically reliable calibration. Comprehensive ablation studies further confirm the contribution of each architectural component, highlighting the potential of trajectory-aware graph modeling for multimodal survival prediction. Access our code at <https://github.com/MICS-Lab/ChronoSurv>.

Keywords: Survival analysis · Multimodality · Head and neck cancer.

1 Introduction

The integration of machine learning into precision oncology has significantly advanced survival prediction and treatment planning across a wide range of malignancies [36]. Among these, head and neck cancer (HNC) remains one of the most prevalent cancers worldwide [4]. Despite progress in diagnostic and therapeutic strategies, patient outcomes remain suboptimal, with five-year survival rates still limited [7]. A major challenge lies in the heterogeneous and high-dimensional nature of HNC patient data, spanning clinical variables, medical imaging, pathological findings, and reports [15]. Effectively integrating these complementary yet structurally diverse data sources remains an open problem.

Survival analysis aims to predict the time until an event of interest while accounting for censored observations. In HNC, multimodal survival prediction is complicated by substantial discrepancies in scale and granularity across data sources. Bridging these heterogeneous representations, ranging from compact clinical variables to high-dimensional Whole Slide Images (WSIs) and textual reports, requires models that can effectively capture cross-modal interactions while accounting for their distinct structural characteristics.

Early survival analysis methods rely primarily on statistical models [37,13], which are limited in their ability to capture non-linear relationships between patient characteristics and survival risk. Deep learning approaches [24,27,21] have demonstrated improved performance over classical methods [22,12], yet most are designed for unimodal inputs.

A prominent line of research in multimodal survival analysis combines WSIs with genomic profiles [10,11,38,31,23], typically extracting modality-specific features followed by feature fusion and survival prediction. While these approaches demonstrate improved performance over unimodal baselines, they often focus on limited modality combinations. In head and neck cancer specifically, prior work has primarily targeted PET/CT-based imaging pipelines [30,32,39], notably around the HECKTOR benchmark [2]. These studies predominantly address segmentation and recurrence-free survival prediction from PET/CT imaging rather than structured fusion across diverse clinical modalities and care stages. More recently, broader multimodal strategies, incorporating tabular data, textual reports, and histopathology, have been explored through late fusion [35], hierarchical pooling [28], graph-based [33] or attention-based [14,20] fusion. However, most existing methods treat patient data as an unordered collection of features, overlooking the hierarchical organization of clinical information and the temporal structure of care pathways. In particular, the sequential progression of clinical decision-making, from patient history review through diagnosis to surgical intervention, remains underexplored in current survival frameworks.

Graph Neural Networks (GNNs) provide a principled framework for modeling relational structure in data. In computational pathology, GNNs have been extensively used to capture spatial dependencies between tissue patches within whole slide images [9,40,29], organs at risk [3] or to model patient similarity [25]. Their ability to encode heterogeneous, multi-scale relationships makes them a natural candidate for representing structured clinical workflows. To our knowledge, however, no prior work leverages GNNs to explicitly model the chronological progression of clinical care pathways for multimodal survival prediction.

To address these challenges, we propose ChronoSurv, a novel hierarchical directed graph framework for multimodal survival analysis. Inspired by the clinical care workflow, our method explicitly models patient trajectories from background review to diagnosis and surgical intervention, enabling progression-aware prediction. Notably, ChronoSurv integrates a heterogeneous multi-level message passing scheme that jointly captures sub-modality features, clinical steps and patient-level representations while supporting adaptable topology for missing-modality robustness. In a multi-cohort setting combining two public datasets,

ChronoSurv demonstrates state-of-the-art discriminative performance alongside reliable calibration, supported by comprehensive ablation studies.

2 Method

As shown in Fig. 1, we propose ChronoSurv, a multimodal framework for survival analysis in head and neck cancer that mirrors the clinical care pathway. Our approach extracts per-modality features (Sec. 2.1), models each patient as a hierarchical directed graph (Sec. 2.2), enables feature interaction through heterogeneous message passing (Sec. 2.3), and predicts discrete-time hazards (Sec. 2.4).

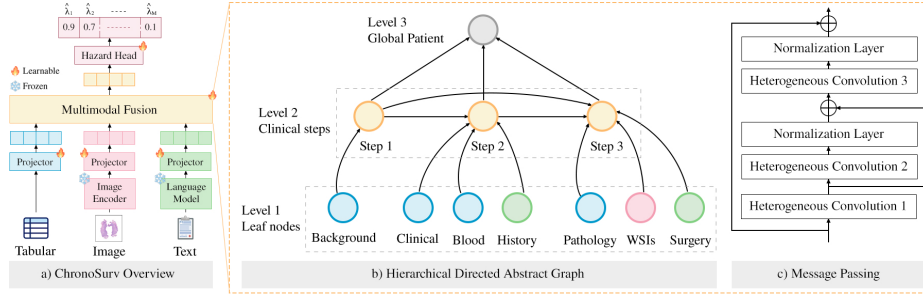


Fig. 1: ChronoSurv overview. (a) Multimodal features are integrated through a dedicated fusion module prior to survival prediction. (b) Each patient is modeled as a hierarchical directed graph encoding modality-, step-, and patient-level structure. (c) Heterogeneous convolutions enable cross-level message passing.

Survival Modeling Setup. We consider right-censored survival data with patient covariates X , event time T^* and censoring time C . The observed time is $T = \min(T^*, C)$ with event indicator $\delta = \mathbb{1}(T^* \leq C)$. We assume conditional non-informative censoring, i.e., $T^* \perp\!\!\!\perp C \mid X$. To enable flexible neural modeling of time-to-event outcomes, we adopt a discrete-time survival formulation [16]. Given a prediction horizon $T_{\max}$ and a time grid $0 = t_0 < t_1 < \dots < t_M = T_{\max}$, for each interval $j \in \{1, \dots, M\}$, we model the hazard, denoted as $\lambda(t_j \mid X)$ and defined as $\lambda(t_j \mid X) = \mathbb{P}(T^* = t_j \mid T^* \geq t_j, X)$, from which the conditional survival function follows as $S(t_j \mid X) = \mathbb{P}(T^* > t_j \mid X) = \prod_{k=1}^j (1 - \lambda(t_k \mid X))$, enabling computation of survival probabilities and likelihood-based training.

2.1 Feature Initialization

The patient data span a wide range of modalities including WSI ($\mathcal{K}_{\text{wsi}}$), tabular ($\mathcal{K}_{\text{tab}}$), and text ($\mathcal{K}_{\text{text}}$) sub-modalities, with $\mathcal{K} = \mathcal{K}_{\text{tab}} \cup \mathcal{K}_{\text{wsi}} \cup \mathcal{K}_{\text{text}}$. Each sub-modality is projected into a shared d -dimensional latent space:

$$h_k = f_k \left(\mathbb{1}_{k \in \mathcal{K}_{\text{tab}}} x_k + \mathbb{1}_{k \in \mathcal{K}_{\text{wsi}}} \Phi_{\text{enc}}^{\text{visual}}(x_k) + \mathbb{1}_{k \in \mathcal{K}_{\text{text}}} \Phi_{\text{model}}^{\text{language}}(x_k) \right), \quad (1)$$

where $h_k \in \mathbb{R}^d$ is the encoded representation of sub-modality k , each f_k is a learnable projector, $\Phi_{\text{enc}}^{\text{visual}}$ a pretrained visual encoder and $\Phi_{\text{model}}^{\text{language}}$ a pretrained language model, both kept frozen.

2.2 Graph Construction

Each patient is represented as a hierarchical directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose node set is partitioned into three levels (Fig. 1b). First-level leaf nodes $\mathcal{V}_1$ carry the sub-modality representations $\{h_k\}_{k \in \mathcal{K}}$ from Sec. 2.1. Second-level nodes $\mathcal{V}_2$ encode the temporal structure of the clinical care pathway through three key steps: (1) patient background, (2) initial cancer diagnosis, and (3) local surgery, their features are zero-initialized. Each leaf node is connected to its corresponding clinical step via a directed edge (e.g. blood $\rightarrow$ step 2, WSIs $\rightarrow$ step 3), forming edge set $\mathcal{E}_{1 \rightarrow 2}$. Clinical-step nodes are further linked by directed edges $\mathcal{E}_{\text{temporal}}$ aligned with the sequential ordering of clinical events. The third level is a single patient node $\mathcal{V}_3$, initialized by mean-pooling over available leaf embeddings and connected to all step nodes through $\mathcal{E}_{2 \rightarrow 3}$. Node and edge sets are defined as:

$$\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \mathcal{V}_3, \quad \mathcal{E} = \mathcal{E}_{1 \rightarrow 2} \cup \mathcal{E}_{\text{temporal}} \cup \mathcal{E}_{2 \rightarrow 3}. \quad (2)$$

All modalities are static covariates ($t = 0$ at pathologic diagnosis). Directed edges encode their acquisition order, not dynamic changes during follow-up. Importantly, the graph topology is patient-specific: when a sub-modality k is unavailable, due to differences in clinical protocols, institutional practices, or incomplete records, the corresponding leaf node and its incident edges are omitted from $\mathcal{G}$, providing a principled mechanism for handling missing modalities and making ChronoSurv applicable to heterogeneous real-world clinical settings.

2.3 Message Passing

ChronoSurv employs three heterogeneous graph attention layers [5], each operating on a distinct relation type to progressively build a patient-level representation. The first layer Ψ_1 pools fine-grained sub-modality features into coarser clinical-step representations through $\mathcal{R}_{1 \rightarrow 2} = \{k \rightarrow s \mid k \in \mathcal{K}\}$, as follows:

$$h_s^{(1)} = \sum_{r \in \mathcal{R}_{1 \rightarrow 2}} \sum_{u \in \mathcal{N}_r(s)} \Psi_1(h_u, h_s), \quad s \in \mathcal{V}_2, \quad (3)$$

where $\mathcal{N}_r(s)$ denotes the set of neighbors of s under relation r . The second layer Ψ_2 captures progression-aware dependencies by propagating information along the clinical trajectory through next ($s_i \xrightarrow{N} s_j, j-i=1$) and skip ($s_i \xrightarrow{S} s_j, j-i>1$) directed temporal edges between step nodes, such that:

$$h_s^{(2)} = h_s^{(1)} + \text{LayerNorm} \left(\sum_{r \in \mathcal{R}_t} \sum_{u \in \mathcal{N}_r(s)} \Psi_2(h_u^{(1)}, h_s^{(1)}) \right), \quad s \in \mathcal{V}_2. \quad (4)$$

At this stage, step nodes jointly capture aggregated sub-modality information and temporal clinical progression, serving as intermediate representations that bridge fine-grained features and the patient-level embedding. Finally, the third layer Ψ_3 fuses all clinical-step representations into a global patient embedding through $\mathcal{R}_{2 \rightarrow 3} = \{s \rightarrow g \mid s \in \mathcal{V}_2\}$, following:

$$h_g = h_g^{(0)} + \text{LayerNorm} \left(\sum_{r \in \mathcal{R}_{2 \rightarrow 3}} \sum_{u \in \mathcal{N}_r(g)} \Psi_3(h_u^{(2)}, h_g^{(0)}) \right), \quad (5)$$

where $h_g^{(0)}$ denotes the mean-pooled initialization of the patient node.

2.4 Survival Head

The survival head receives the patient embedding h_g and predicts bin-wise hazards over the discrete grid defined as follows:

$$\hat{\lambda} = \sigma(\Psi_{\text{head}}^{\text{hazard}}(h_g)) \in [0, 1]^M, \quad (6)$$

where $\hat{\lambda}_j$ estimates the conditional event probability in bin j given survival up to bin $j-1$. Let $\kappa(t_i)$ denote the bin index of observed time t_i and $y_{ik} := \mathbb{1}(k = \kappa(t_i), \delta_i = 1)$. Training minimizes $\mathcal{L} = \mathcal{L}_{\text{NLL}} + \beta \mathcal{L}_{\text{X-CAL}}$, where:

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{\kappa(t_i)} \left[y_{ik} \log \hat{\lambda}_{ik} + (1 - y_{ik}) \log(1 - \hat{\lambda}_{ik}) \right], \quad (7)$$

is the discrete-time negative log-likelihood [6], $\mathcal{L}_{\text{X-CAL}}$ a differentiable calibration penalty [17], and β a tunable weighting coefficient. Day-level survival predictions are obtained by assigning each day to its corresponding time bin.

3 Experiments

Datasets. We leverage two public head and neck cancer datasets for multi-modal survival analysis: **HANCOCK** [15] and **TCGA-HNSC** [26], summarized in Table 1. **HANCOCK** provides clinical, blood, and pathological data, textual surgical reports, and WSIs from both primary tumors and lymph nodes. **TCGA-HNSC** covers clinical and pathological data with tumor WSIs only, lacking blood data, textual reports, and lymph-node slides. We further construct **UnihNC** by combining both sources, creating a multi-cohort setting that reflects real-world clinical heterogeneity where modality availability varies across institutions (Sec. 2.2).

Setup. Experiments use 5-fold cross-validation with patient-level 70/15/15 splits. Models are trained with early stopping on the validation set loss and evaluated on the held-out test fold, reporting mean and standard deviation across folds. All models leverage the same frozen feature extraction module: UNI [8] as visual encoder and BioClinicalBERT [1] as language model. ChronoSurv is trained for 20,000 iterations with a batch size of 8 using AdamW, a 3×10^{-4} learning rate, a

Table 1: Dataset overview. **UniHNC** is the union of **HANCOCK** and **TCGA-HNSC**. ✓/✗ indicate modality availability.

Dataset	Origin	Samples	Censoring	Tabular	WSI	Text
HANCOCK [15]	Germany	763	72.1 %	✓	✓	✓
TCGA-HNSC [26]	USA	526	57.6 %	✓	✓	✗
UniHNC	Germany & USA	1,289	66.1 %	✓	✓	✓

5×10^{-4} weight decay and a cosine annealing scheduler. The number of time bins is set to $M = 100$ (tuned over $\{10, 50, 100, 500, 1000\}$), the calibration weight to $\beta = 7$ (tuned over $\{0.1, 1, 3, 5, 7, 15\}$) and the embedding dimension to $d = 512$. Inference takes approximately 30 ms per sample on a single RTX 2080 Ti.

3.1 Evaluation results

Table 2 compares ChronoSurv against *Non-Deep Learning* (■), *Deep Learning* (■), and *Multimodal Deep Learning* (◆) baselines. Non-Deep Learning and Deep Learning baselines receive the concatenation of all features and missing sub-modalities are zero-imputed for all baselines. Discrimination performance is evaluated using C_{index} [34], while the IBS [18] assesses overall survival probability accuracy. ChronoSurv achieves the highest C_{index} on both **UniHNC** and **HANCOCK** while obtaining the best or tied-best IBS across all settings. Among multimodal baselines, SurvPCG reaches the closest C_{index} on **UniHNC** but with a higher IBS, suggesting that although cross-modal attention captures relevant interactions for risk ranking, it yields less accurate survival probability estimates. On **TCGA-HNSC**, where text and blood modalities are unavailable, ChronoSurv matches the best IBS while maintaining competitive discrimination performance, including against MMD, which is designed to handle incomplete multimodal data. Importantly, when applying the D-Calibration test [19] with a Type I error rate of $\alpha = 0.05$, ChronoSurv passes the test in all 5 folds ($p \geq 0.05$), whereas SurvPCG passes only 1/5. Baselines with consistent calibration (Cox, RSF, MultiSurv: 5/5) also achieve competitive IBS but substantially lower discrimination.

3.2 Ablation study

Effect of multimodal fusion. Table 3 compares our multimodal fusion module with unimodal baselines (■) and variants replacing the fusion module by widely used aggregation mechanisms (■), while keeping identical feature initialization and survival heads. Unimodal baselines confirm that each modality carries prognostic signal. In fact, tabular features alone match or surpass several multimodal strategies, highlighting that naïve fusion can be detrimental. In contrast, ChronoSurv consistently outperforms all alternative aggregation schemes, supporting the effectiveness of structured multimodal interaction modeling.

Table 2: Performance on the UniHNC dataset, and ^(†) dataset-specific results on HANCOCK and TCGA-HNSC. **Best** in bold, second best underlined.

■ Non-Deep Learning ■ Deep Learning ♦ Multimodal Deep Learning.

Method	UniHNC		HANCOCK ^(†)		TCGA-HNSC ^(†)	
	C _{index} (↑)	IBS (↓)	C _{index} (↑)	IBS (↓)	C _{index} (↑)	IBS (↓)
■ Weibull [37]	.636 ±.036	.167 ±.018	.589 ±.059	.150 ±.016	.593 ±.065	.195 ±.024
■ CoxPH [13]	.652 ±.032	.155 ±.008	.623 ±.035	.138 ±.012	.619 ±.059	.183 ±.008
■ RSF [22]	.654 ±.032	.154 ±.012	.622 ±.019	.136 ±.009	.584 ±.080	.183 ±.018
■ DeepSurv [24]	.674 ±.028	.165 ±.013	.644 ±.048	.149 ±.022	.639 ±.068	.191 ±.015
■ DeepHit [27]	.619 ±.030	.178 ±.006	.516 ±.013	.143 ±.014	.619 ±.058	.231 ±.020
■ TransDSA [21]	.629 ±.024	.263 ±.045	.602 ±.027	.255 ±.077	.514 ±.056	.280 ±.018
♦ MMD [14]	.683 ±.024	.184 ±.028	.659 ±.031	.187 ±.025	.659 ±.067	.182 ±.035
♦ GraphMMP [33]	.667 ±.038	.161 ±.026	.657 ±.019	.145 ±.014	.633 ±.073	.187 ±.045
♦ HFBSurv [28]	.679 ±.019	.302 ±.030	.656 ±.042	.349 ±.063	.660 ±.067	.238 ±.036
♦ MultiSurv [35]	.651 ±.045	.156 ±.009	.637 ±.041	.141 ±.014	.591 ±.058	.182 ±.006
♦ SurvPCG [20]	.695 ±.023	.163 ±.012	.672 ±.038	.140 ±.011	.659 ±.060	.198 ±.029
♦ ChronoSurv (Ours)	.702 ±.039	.153 ±.019	.704 ±.037	.134 ±.020	.652 ±.048	.182 ±.022

Table 3: Comparison with unimodal baselines and alternative multimodal aggregation schemes. ■ Unimodal. ■ Multimodal.

Method	UniHNC		HANCOCK ^(†)		TCGA-HNSC ^(†)	
	C _{index} (↑)	IBS (↓)	C _{index} (↑)	IBS (↓)	C _{index} (↑)	IBS (↓)
■ Text-only	.606 ±.037	.160 ±.011	.543 ±.036	.143 ±.017	-	-
■ Image-only	.638 ±.012	.172 ±.024	.652 ±.044	.147 ±.020	.549 ±.055	.212 ±.033
■ Tabular-only	.668 ±.065	.170 ±.039	.665 ±.090	.138 ±.021	.596 ±.044	.219 ±.076
■ Mean pooling	.641 ±.056	.228 ±.043	.643 ±.061	.222 ±.041	.522 ±.102	.239 ±.066
■ MLP	.655 ±.063	.214 ±.043	.638 ±.059	.213 ±.035	.579 ±.114	.222 ±.070
■ Gated Fusion	.624 ±.066	.226 ±.043	.592 ±.063	.224 ±.044	.543 ±.127	.238 ±.065
■ FC GNN	.674 ±.040	.216 ±.051	.672 ±.047	.206 ±.050	.598 ±.052	.233 ±.068
■ Cross-attention	.689 ±.031	.235 ±.053	.677 ±.043	.227 ±.044	.656 ±.052	.245 ±.092
■ Self-attention	.695 ±.038	.213 ±.055	.697 ±.039	.200 ±.050	.644 ±.025	.236 ±.076
■ ChronoSurv (Ours)	.702 ±.039	.153 ±.019	.704 ±.037	.134 ±.020	.652 ±.048	.182 ±.022

Component-wise ablation. Table 4 reports a leave-one-out ablation study on ChronoSurv’s components. Removing any individual clinical step and its corresponding leaf nodes (1-3) degrades performance across all datasets, highlighting that each clinical stage provides complementary prognostic information. Notably, excluding step 2 (initial cancer diagnosis) yields the largest reduction in C_{index}. Removing hierarchical levels (4 & 5) results in a performance decline, indicating that incorporating coarser information captured by intermediate clinical steps and global patient representations is critical for effective multimodal integration. Disabling heterogeneous message passing (6) consistently degrades C_{index} and IBS across datasets, demonstrating the importance of relation-specific modeling for structured, multi-level message passing. Finally, replacing directed edges with undirected ones (7) also degrades performance, suggesting that preserving the temporally ordered topology leads to more accurate prediction.

Table 4: Leave-one-out ablation on clinical-steps modeling (1-3), hierarchical levels (4 and 5), graph heterogeneity (6) and edge directionality (7).

Method	UniHNC		HANCOCK ^(†)		TCGA-HNSC ^(†)	
	C _{index} (↑)	IBS (↓)	C _{index} (↑)	IBS (↓)	C _{index} (↑)	IBS (↓)
ChronoSurv	.702 $\pm$.039	.153 $\pm$.019	.704 $\pm$.037	.134 $\pm$.020	.652 $\pm$.048	.182 $\pm$.022
(1) w/o step 1	.644 $\pm$.088	.162 $\pm$.016	.638 $\pm$.045	.143 $\pm$.018	.596 $\pm$.112	.193 $\pm$.029
(2) w/o step 2	.586 $\pm$.056	.182 $\pm$.016	.552 $\pm$.058	.145 $\pm$.011	.586 $\pm$.026	.239 $\pm$.036
(3) w/o step 3	.607 $\pm$.050	.171 $\pm$.021	.607 $\pm$.070	.141 $\pm$.021	.564 $\pm$.047	.217 $\pm$.031
(4) w/o level 2	.546 $\pm$.061	.186 $\pm$.018	.546 $\pm$.066	.147 $\pm$.019	.529 $\pm$.088	.246 $\pm$.024
(5) w/o level 3	.598 $\pm$.067	.181 $\pm$.022	.579 $\pm$.052	.149 $\pm$.018	.552 $\pm$.056	.231 $\pm$.043
(6) w/o heterogeneity	.644 $\pm$.043	.157 $\pm$.009	.598 $\pm$.062	.144 $\pm$.012	.631 $\pm$.047	.179 $\pm$.018
(7) w/o directionality	.681 $\pm$.058	.199 $\pm$.039	.679 $\pm$.031	.196 $\pm$.035	.646 $\pm$.054	.211 $\pm$.049

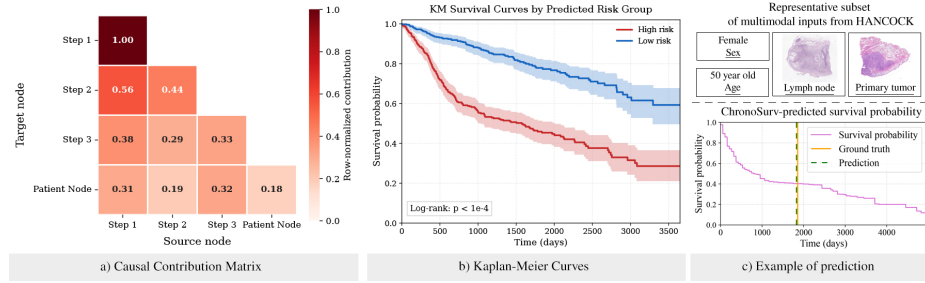


Fig. 2: Qualitative results. (a) Contribution matrix between clinical steps and the global patient node, illustrating the row-normalized magnitude of message-passing contributions from each source node to its corresponding target. (b) Kaplan-Meier survival curves stratified by predicted risk groups. (c) Patient-level prediction displaying the estimated survival curve and survival time.

3.3 Qualitative results

Fig. 2 provides interpretability insights into ChronoSurv’s learned representations. The contribution matrix (a) indicates how information flows through the graph hierarchy, extracted from normalized contribution between two nodes. These results indicate that step 1 (background) acts as a self-contained source, step 2 (initial diagnosis) predominantly aggregates background patient information from step 1, while step 3 (local surgery) draws from all preceding steps, further highlighting the complementary nature of clinical steps. At the patient level, step 1 and step 3 contribute most to the global representation, reflecting the clinical importance of both patient history and surgical findings for prognosis. Kaplan-Meier analysis (b) demonstrates clear risk stratification, showing a statistically significant difference between groups (log-rank test $p < 10^{-4}$). The example of prediction for a test sample from HANCOCK (c) further illustrates that ChronoSurv can produce clinically consistent survival predictions.

4 Discussion and Conclusion

We introduce ChronoSurv, a heterogeneous hierarchical directed graph framework for multimodal survival analysis in head and neck cancer. By modeling the clinical care pathway as a temporally directed graph with heterogeneous message passing, ChronoSurv captures progression-aware dependencies while naturally handling missing modalities through adaptive topology. Experiments on two public datasets show that structured, progression-aware graph modeling yields both accurate risk ranking and well-calibrated survival estimates, highlighting the potential of clinically-guided modeling for multimodal survival analysis.

Limitations and Future work. Although our study primarily focuses on multimodal fusion, we anticipate that opportunities for improvement could come from fine-tuning initial feature extractors or exploring alternative visual encoders and language models, which is left for future work. Additionally, future work will explore scaling the framework to larger cohorts and more diverse malignancies as additional annotations become available.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.B.A.: Publicly Available Clinical BERT Embeddings (Jun 2019)
2. Andrearczyk, V., Oreiller, V., Abobakr, M., Hatt, M., Depeursinge, A., et al.: Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT (2023)
3. Bae, J., et al.: HoG-Net: Hierarchical Multi-Organ Graph Network for Head and Neck Cancer Recurrence Prediction from CT Images . In: MICCAI (2024)
4. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **74**, 229–263 (2024)
5. Brody, S., et al.: How Attentive are Graph Attention Networks? (2022)
6. Brown, C.C.: On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics* **31**(4), 863–872 (Dec 1975)
7. Budach, V., Tinhofer, I.: Novel prognostic clinical factors and biomarkers for outcome prediction in head and neck cancer. *The Lancet. Oncology* (2019)
8. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Williams, W., Mahmood, F., et al.: Towards a general-purpose foundation model for computational pathology. *Nature Medicine* (2024)
9. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F.K., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: MICCAI (2021)
10. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: ICCV (2021)

11. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Shaban, M., Shady, M., Williams, M., Joo, B., Noor, Z., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* (2022)
12. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 785–794 (Aug 2016), arXiv:1603.02754 [cs]
13. Cox, D.R.: Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2), 187–220 (1972)
14. Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Zhao, Y.W., Yang, H., Landman, B.A., Huo, Y.: Survival Prediction of Brain Cancer with Incomplete Radiology, Pathology, Genomics, and Demographic Data (Jul 2022), arXiv:2203.04419 [cs]
15. Dörrich, M., Balk, M., Heusinger, T., Beyer, S., Kanso, H., Matek, C., Hartmann, A., Iro, H., Eckstein, M., Gostian, A.O., Kist, A.M.: A multimodal dataset for precision oncology in head and neck cancer (May 2024), pages: 2024.05.29.24308141
16. Gensheimer, M.F., Narasimhan, B.: A scalable discrete-time survival model for neural networks **7**, e6257. <https://doi.org/10.7717/peerj.6257>, <https://pmc.ncbi.nlm.nih.gov/articles/PMC6348952/>
17. Goldstein, M., Han, X., Puli, A., Perotte, A., Ranganath, R.: X-CAL: Explicit Calibration for Survival Analysis. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 18296–18307. Curran Associates, Inc. (2020)
18. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* (1999)
19. Haider, H., Hoehn, B., Davis, S., Greiner, R.: Effective ways to build and evaluate individual survival distributions (2018), <https://arxiv.org/abs/1811.11347>
20. Hou, J., Zhang, R., Xie, Y., Li, C., Qin, W.: Multimodal deep learning for cancer prognosis prediction with clinical information prompts integration. *npj Digital Medicine* **9**(1), 76 (Dec 2025)
21. Hu, S., Fridgeirsson, E., Wingen, G.v., Welling, M.: Transformer-Based Deep Survival Analysis. In: *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*. pp. 132–148. PMLR (May 2021)
22. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. *The Annals of Applied Statistics* **2**(3) (Sep 2008), arXiv:0811.1645 [stat]
23. Jaume, G., Vaidya, A., Chen, R., Williamson, D., Liang, P., Mahmood, F.: Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction (Apr 2024), arXiv:2304.06819 [cs]
24. Katzman, J., Shaham, U., Bates, J., Cloninger, A., Jiang, T., Kluger, Y.: DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology* **18** (2018)
25. Kim, S.Y.: GNN-surv: Discrete-Time Survival Prediction Using Graph Neural Networks. *Bioengineering* **10**(9), 1046 (Sep 2023)
26. Lawrence, M.S., Sougnez, C., Lichtenstein, L., Cibulskis, K., Lander, E., Burnett, K., et al.: Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**(7536), 576–582 (Jan 2015)
27. Lee, C., Zame, W., Yoon, J., Van Der Schaar, M.: DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence* **32**(1) (Apr 2018)
28. Li, R., Wu, X., Li, A., Wang, M.: HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics* (2022)
29. Lu, W., Toss, M., Rakha, E., Rajpoot, N., Minhas, F.: Slidegraph+: Whole slide image level graphs to predict her2status in breast cancer (2021)

30. Meng, M., Bi, L., Kim, J., et al.: Merging-Diverging Hybrid Transformer Networks for Survival Prediction in Head and Neck Cancer. In: MICCAI (2023)
31. Qu, M., Yang, G., Di, D., Su, T., Fan, L., et al.: Multimodal cancer survival analysis via hypergraph learning with cross-modality rebalance. In: IJCAI (2025)
32. Saeed, N., Ridzuan, M., Maani, F.A., Yaqub, M., et al.: SurvRNC: Learning Ordered Representations for Survival Prediction using Rank-N-Contrast (2024)
33. Shan, X., Ge, R., Liu, J., Wu, L., Zhang, C., Liu, S., Qin, W., Min, W., Elazab, A., Wang, C.: GraphMMP: A Graph Neural Network Model with Mutual Information and Global Fusion for Multimodal Medical Prognosis (Aug 2025)
34. Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., Wei, L.J.: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**(10), 1105–1117 (May 2011)
35. Vale-Silva, L.A., Rohr, K.: Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports* **11**(1), 13505 (Jun 2021)
36. Wang, C.W., Khalil, M.A., Puspita Firdi, N.: A Survey on Deep Learning for Precision Oncology (Jun 2022)
37. Weibull, W.: A Statistical Distribution Function of Wide Applicability. *Journal of Applied Mechanics* pp. 293–297 (1951)
38. Xiong, C., Chen, H., Zheng, H., Wei, D., Zheng, Y., Sung, J.J., King, I.: Mome: Mixture of multimodal experts for cancer survival prediction (2024)
39. Zhang, H., Liu, Y., Sun, Y., Shen, L., Wei, L.: Improving Survival Prediction of Head-and-Neck Cancer with Medical Image, Foundation Models and Multi-modal Fusion. In: ML4H (2025)
40. Zhou, Y., Graham, S., Koohbanani, N.A., Shaban, M., Heng, P.A., Rajpoot, N.: Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In: ICCV (2019)