Replicating Patient Follow-Up with Hierarchical Directed Graphs for Head and Neck Cancer Survival Analysis

Theo Di Piazza[1,*] THEO.DIPIAZZA@CREATIS.INSA-LYON.FR
Hugo Miccinilli[*]
[1] *INSA Lyon, University of Lyon, CNRS, INSERM, CREATIS UMR 5220, U1294, Lyon, France*
[*] *Equal contribution*

Editors: Under Review for MIDL 2026

Abstract

Head and neck cancer is a common malignancy with persistently limited survival outcomes, making accurate clinical prognosis particularly challenging. To establish a diagnosis, patients typically undergo a series of examinations producing heterogeneous data. This includes clinical data review, blood tests, tissue sampling, and lymph node analysis, encompassing multiple imaging and non-imaging modalities prior to treatment, which often involves surgery for disease treatment. Despite advances in diagnostic imaging and clinical assessment, treatment decisions remain largely dependent on the disease stage. This highlights the critical need for automated and reliable tools capable of accurately estimating patient survival to further assist clinicians in personalized treatment planning. Existing survival analysis methods, typically rely on shallow architectures or early-fusion schemes that struggle to exploit the complexity and structure of multimodal clinical data. To address these limitations, we introduce H2DGSurv, a deep learning framework for survival prediction in head and neck cancer that models multimodal patient data as a directed hierarchical heterogeneous graph tracing the clinical workflow from initial diagnosis to surgery. The proposed architecture organizes modality-specific leaf nodes under clinical step-level parent nodes, and integrates a global patient node to capture consolidated representations prior to survival prediction. Experimental results demonstrate that H2DGSurv substantially improves survival prediction performance compared with established baselines, while ablation studies confirm the importance of each model component.
Code: https://github.com/dpmc-lab/h2dg-surv.
Models: https://huggingface.co/dpmc/h2dg-surv.
Keywords: Survival analysis, Head and neck cancer, Multimodal, Graph neural networks.

1. Introduction

The integration of machine learning into precision oncology has significantly improved survival prediction and treatment planning across various malignancies (Wang et al., 2022). Among these, head and neck cancer (HNC) remains one of the most prevalent cancers worldwide (Bray et al., 2024). Despite substantial progress in diagnostic imaging and therapeutic strategies, patient outcomes remain poor, with 5-year survival rates still limited (Budach and Tinhofer, 2019). This persistent gap underscores the urgent need for robust and automated tools capable of assisting oncologists in clinical decision-making and personalized prognosis estimation. However, developing accurate survival models for HNC remains particularly challenging due to the scarcity of large-scale, publicly available datasets with paired, multimodal clinical and imaging data (Dörrich et al., 2024).

In clinical practice, the diagnostic process for HNC involves a sequence of complementary examinations, including clinical data review, blood tests, tissue microarrays and whole slides images analysis to support metastasis detection and provide information on tumor type. These examinations generate heterogeneous and high-dimensional data spanning multiple modalities, which are essential to determine the disease stage and estimate patient survival.

Although prior works have successfully applied survival analysis approaches across various clinical modalities (Barnwal et al., 2021), applying survival framework to medical multimodal data present significant challenges. While deep learning survival analysis frameworks have improved survival prediction accuracy (Katzman et al., 2018), existing works rely on early fusion strategies that insufficiently capture inter-modality relationship, thus limiting their ability to fully exploit the complementary nature of multimodal clinical data.

Therefore, we introduce H2DGSurv, a Hierarchical Heterogeneous Directed Graph framework for survival prediction in head and neck cancer. Our approach explicitly models the clinical workflow, from initial diagnosis through imaging, reporting, and surgery, using a directed graph in which each clinical step is represented as a node aligned with the patient's follow-up timeline. To capture modality-specific information while enabling a unified patient representation, we design a three-level hierarchical structure: (1) a patient-level root node, (2) step-level nodes corresponding to key clinical events, and (3) leaf-level nodes encoding each imaging or report modality. This hierarchical heterogeneous formulation facilitates richer feature aggregation by jointly modeling distinct data types and their asymmetric relationships. Our contributions can be summarized as:

- We introduce a directed multimodal graph representation that captures the patient follow-up trajectory from cancer diagnosis to surgery.
- We develop a hierarchical heterogeneous graph architecture that integrates modality-specific representations while preserving clinically meaningful structure, ultimately improving survival prediction.
- We conduct ablations to quantify the impact of each model component, clinical step, and hierarchical levels. We further provide additional qualitative analyses.
- We make our trained models and source codes publicly available to support reproducibility in precision oncology for head and neck cancer.

2. Related Work

2.1. Survival Analysis

Survival analysis is a fundamental field with critical applications ranging from engineering reliability to clinical prognosis, where the goal is to predict the time until an event of interest occurs. It has evolved from classical statistical methods to sophisticated machine learning approaches, driven by the need to model complex relationships in real world data.

Statistical and Machine Learning Approaches. A cornerstone of the field is the **Kaplan-Meier** estimator (Kaplan and Meier, 1958), a non-parametric method that estimates the survival function from censored data but cannot account for patient-specific covariates. The **Cox Proportional Hazards (CoxPH)** model (Cox, 1972) addresses this by formulating the hazard rate as a product of a baseline hazard and a covariate-dependent term. It relies on the semi-parametric Proportional Hazards (PH) assumption,

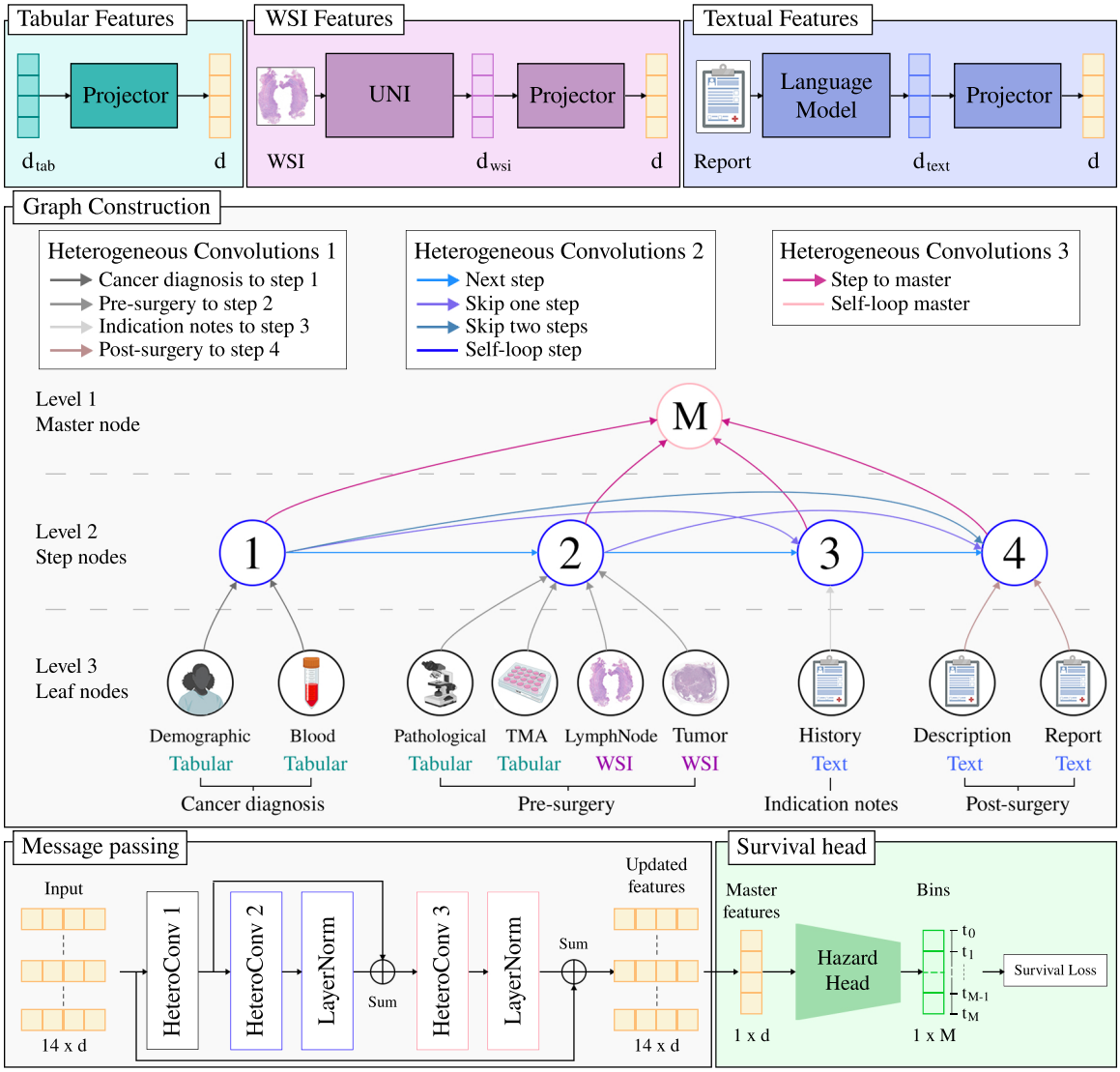


Figure 1: H2DGSurv features a new hierarchical directed heterogeneous framework for survival analysis, following patient follow-up from initial diagnosis to local surgery.

where the effect of covariates is constant over time. Parametric models like **Weibull** regression (Weibull, 1951) may offer greater efficiency but require strong assumptions about the underlying event time distribution. To capture non-linear feature interactions without such strict constraints, machine learning ensemble methods were adapted for survival tasks, notably **Random Survival Forests (RSF)** (Ishwaran et al., 2008) and gradient boosting variants like **Survival Gradient Boosting** (Chen et al., 2013) and **Survival-XGBoost** (Barnwal et al., 2021).

Deep Learning for Survival. Deep neural networks have further advanced the field by enabling end-to-end learning of complex representations. **DeepSurv** (Katzman et al.,

2018) retains the structure of the Cox model but estimates the hazard function with a deep neural network, effectively modeling non-linear risk functions while maintaining the PH assumption. Other approaches seek to relax this constraint entirely: discrete-time models like **DeepHit** (Lee et al., 2018) estimate the probability mass function directly, allowing the hazard to vary freely over time. More recently, the widespread success of attention mechanisms has inspired architectures such as **TransDSA** (Hu et al., 2021), which adapts Transformer blocks (Vaswani et al., 2017) to survival analysis, demonstrating the versatility of modern deep learning architectures for time-to-event prediction.

2.2. Graph Neural Networks

Graph Neural Networks (GNNs) (Gori et al., 2005) have emerged as a powerful paradigm for learning representations from non-Euclidean data structures. Standard GNN architectures typically operate via message-passing mechanisms. A key development was the Graph Convolutional Network (GCN)(Kipf and Welling, 2017), which approximated spectral graph convolutions via a localized first-order approximation. This was generalized by attention mechanisms in Graph Attention Networks (GAT) (Veličković et al., 2018), enabling nodes to weigh the importance of their neighbors adaptively. Limitations in the static attention mechanism of GAT prompted the development of GATv2 (Brody et al., 2022), which introduced dynamic attention to improve expressivity. To better capture the complexity of real-world data, research has expanded beyond homogeneous simple graphs. Heterogeneous GNNs were developed to model graphs with multiple types of nodes and edges, employing relation-specific message passing to preserve semantic distinctions (Zhang et al., 2019). On the other hand, Hierarchical GNNs have been proposed to process data at multiple levels of abstraction, typically through pooling operations or multi-level architectures, allowing models to capture both local interactions and global structural patterns (Sobolevsky, 2021). In medical imaging, prior work leveraged hierarchical representation to model different resolutions of Whole Slide Images (Guo et al., 2023), while heterogeneous graph learning demonstrates promising results for multi-modal medical data fusion (Kim et al., 2023).

3. Method

3.1. Discrete-Time Survival Analysis

Let T^* and C be positive random variables denoting the event time and the censoring time, respectively and let X be a random feature vector. We consider the observed time $T := \min(T^*, C)$ and the event indicator $\delta := \mathbb{1}(T^* \leq C)$. We assume a finite prediction horizon T_{max} and we adopt the standard non-informal censoring assumption $T \perp\!\!\!\perp C \mid X$. We work on a fixed discrete timeline $0 = t_0 < t_1 < \cdots < t_M = T_{\max}$ and take $T^* \in \{t_1, \ldots, t_M\}$. The survival function S and the discrete-time hazard λ are defined as:

$$S(t_j \mid X) := \mathbb{P}(T^* > t_j \mid X), \qquad \lambda(t_j \mid X) := \mathbb{P}(T^* = t_j \mid T^* > t_{j-1}, X) \tag{1}$$

These quantities satisfy $S(t_j \mid X) = \prod_{k=1}^{j}\big(1 - \lambda(t_k \mid X)\big)$. For an observed time t, let $\kappa(t)$ be the unique index such that $t = t_{\kappa(t)}$. Using hazards over the at-risk indices $k = 1, \ldots, \kappa(t)$ and defining $y_k := \mathbb{1}(k = \kappa(t), \delta = 1)$, the event-time likelihood can be written as $L(x, t, \delta) = \prod_{k=1}^{\kappa(t)} \lambda(t_k \mid x)^{y_k} \left(1 - \lambda(t_k \mid x)\right)^{1-y_k}$ (Brown, 1975). Given N samples

$\{(x_i, t_i, \delta_i)\}_{i=1}^N$ observed from (X, T, δ) and $y_{ik} := \mathbb{1}(k = \kappa(t_i), \delta_i = 1)$, the mean negative log-likelihood for discrete-time survival is:

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{\kappa(t_i)} \Big(y_{ik} \log \lambda(t_k \mid x_i) + (1 - y_{ik}) \log\big(1 - \lambda(t_k \mid x_i)\big) \Big) \tag{2}$$

3.2. H2DGSurv

We propose H2DGSurv (Figure 1), a GNN architecture designed to model the sequential nature of the patient's clinical pathway as well as the heterogeneous nature of the data used for their diagnostic. This architecture maps the raw feature vector X into a patient-specific heterogeneous hierarchical graph structure, processes it through a message-passing mechanism, and outputs a latent survival representation used as input in the survival head.

Feature Initialization. For each patient i, we consider a fine-grained decomposition of the patient data x_i into specific sub-modalities indexed by $k \in \mathcal{K}$. The index set is partitioned into WSI, tabular, and text sub-modalities, with $\mathcal{K}_{\text{wsi}} = \{\text{lymph}, \text{tumor}\}$, $\mathcal{K}_{\text{tab}} = \{\text{clinical}, \text{blood}, \text{pathological}, \text{TMA}\}$, and $\mathcal{K}_{\text{text}} = \{\text{history}, \text{surgery report}, \text{surgery desc}\}$, so that $\mathcal{K} = \mathcal{K}_{\text{tab}} \cup \mathcal{K}_{\text{wsi}} \cup \mathcal{K}_{\text{text}}$. For each sub-modality k, we employ a learnable projector f_{θ_k} to map features $x_{i,k}$ to a shared latent space $\mathbb{R}^d$:

$$\mathbf{h}_{i,k} = \begin{cases} f_{\theta_k}(x_{i,k}) & \text{if } k \in \mathcal{K}_{\text{tab}}, \\ f_{\theta_k}\big(\text{UNI}(x_{i,k})\big) & \text{if } k \in \mathcal{K}_{\text{wsi}}, \\ f_{\theta_k}\big(\text{LM}(x_{i,k})\big) & \text{if } k \in \mathcal{K}_{\text{text}}. \end{cases} \tag{3}$$

Here, UNI (Chen et al., 2023) denotes a frozen WSI foundation model and LM a frozen language model.

Message Passing. We define the patient representation as a heterogeneous directed graph $\mathcal{G}_i = (\mathcal{V}, \mathcal{E})$. The node set $\mathcal{V}$ is partitioned into leaf nodes (typed by $k \in \mathcal{K}$), step nodes $\mathcal{V}_S$, and a master node v_m. The leaf nodes features are initialized from $\mathbf{h}_{i,k}$, the steps nodes features are initialized to zero and the master node feature $\mathbf{h}_{i,v_m}$ by mean pooling over $\{\mathbf{h}_{i,k}\}_{k \in \mathcal{K}}$. The edge set $\mathcal{E}$ is defined as $\mathcal{V} \times \mathcal{V} \times \mathcal{R}$, where $\mathcal{R} := \mathcal{R}_{\text{agg}} \cup \mathcal{R}_{\text{prog}} \cup \mathcal{R}_{\text{fusion}}$ specify the relation type and $\mathcal{N}_r(v) = \{u \in \mathcal{V} \mid (u, v, r) \in \mathcal{E}\}$ be the set of neighbors of v connected via relation r:

- $\mathcal{R}_{\text{agg}} = \{(k \to \text{step}) \mid k \in \mathcal{K}\}$: aggregation from leaf nodes to clinical steps.
- $\mathcal{R}_{\text{prog}} = \{(\text{step} \xrightarrow{\text{next}} \text{step}), (\text{step} \xrightarrow{\text{skip}} \text{step})\}$: defines the clinical pathway flow.
- $\mathcal{R}_{\text{fusion}} = \{(\text{step} \to \text{master})\}$: final aggregation from steps to the patient master node.

We employ a three-layer Heterogeneous GNN based on GATv2 (Brody et al., 2022) to produce the patient-level embedding z_i. The layer-wise updates are:

$$\begin{aligned}
\mathbf{h}_{i,v}^{(1)} &= \mathrm{ReLU}\left(\sum_{r\in\mathcal{R}_{\text{agg}}}\sum_{u\in\mathcal{N}_r(v)} \mathrm{GATv2}_r(\mathbf{h}_{i,u},\mathbf{h}_{i,v})\right), && v\in\mathcal{V}_S,\\
\mathbf{h}_{i,v}^{(2)} &= \mathrm{ReLU}\left(\mathbf{h}_{i,v}^{(1)} + \mathrm{LayerNorm}\left(\sum_{r\in\mathcal{R}_{\text{prog}}}\sum_{u\in\mathcal{N}_r(v)} \mathrm{GATv2}_r(\mathbf{h}_{i,u}^{(1)},\mathbf{h}_{i,v}^{(1)})\right)\right), && v\in\mathcal{V}_S,\\
z_i &= \mathrm{ReLU}\left(\mathbf{h}_{i,v_m} + \mathrm{LayerNorm}\left(\sum_{r\in\mathcal{R}_{\text{fusion}}}\sum_{u\in\mathcal{N}_r(v_m)} \mathrm{GATv2}_r(\mathbf{h}_{i,u}^{(2)},\mathbf{h}_{i,v_m})\right)\right)
\end{aligned} \tag{4}$$

Hazard Head. While raw event times lie on a fine-grained discrete timeline, we assume the hazard function is piecewise constant. We partition the time axis $[\![0, T_{\max}]\!]$ into M disjoint bins $[t_{j-1}, t_j)$ and estimate the discrete hazard vector $\hat{\lambda}_i \in [0,1]^M$ using a Multi-Layer Perceptron (MLP) with temperature-scaled sigmoid activation:

$$\hat{\lambda}_i = \sigma\left(\frac{\mathrm{MLP}(z_i)}{\tau}\right) \tag{5}$$

where $\tau > 0$ is a temperature parameter. Each component $\hat{\lambda}_{i,j}$ represents the conditional probability of the event occurring in the j-th bin, given survival up to t_{j-1}. We optimize the model by minimizing $\mathcal{L} := \mathcal{L}_{\text{NLL}} + \beta\mathcal{L}_{\text{X-CAL}}$, where $\mathcal{L}_{\text{X-CAL}}$ is explicit calibration penalty (Goldstein et al., 2020). $\mathcal{L}_{\text{X-CAL}}$ encourages the predicted CDF values to follow a uniform distribution and so penalize not-well calibrated probabilities by turning the D-Calibration (Haider et al., 2020) into a differentiable objective. To recover the estimation of the fine-grained survival function $\hat{S}_i(t)$, we assign the cumulative probability of the corresponding bin to all time steps within it. The predicted survival time in days, denoted as $\hat{T}_i \in \mathbb{R}^+$, is then derived via the Restricted Mean Survival Time (Han and Jung, 2022) as follows:

$$\hat{T}_i = \sum_{t=0}^{T_{max}} \hat{S}_i(t) \tag{6}$$

4. Dataset

Database. We leverage `HANCOCK` (Dörrich et al., 2024) to train and evaluate our method. The multi-modal dataset includes 793 unique patients with three types of modalities: tabular, textual and imaging. Tabular features cover *clinical*, *blood*, *tissue microarray (TMA) measurements* and *pathological* data. Textual data consist of *history*, *surgery descriptions* and *reports*. Imaging data include WSIs of *primary tumors* and *lymph nodes*. For each patient, the time-to-event and censoring indicator are provided. The *deceased tumor specific* label is considered at the positive uncensored event. **Preprocessing.** Textual report sequences are truncated to 200, 512 and 60 tokens for history, surgery, and description reports, respectively. For WSIs, we use the precomputed UNI embeddings provided with the

dataset (Chen et al., 2024), and replace missing WSIs with zero vectors. Tabular features with missing values are imputed with zeros and augmented with binary missing value indicators. All preprocessing steps are applied consistently across all compared methods and are released with our source code. **Protocol.** We perform a 5-fold cross-validation. Each fold is split into a train, validation and test sets following a 70/15/15 rule.
Implementation details. Training runs for 20,000 iterations using AdamW with a cosine annealing scheduler, a 0.0003 learning rate, a 0.0005 weight decay, and a batch size of 8.

5. Experimental results

5.1. Evaluation results

Method	C-index ↑	td-AUC ↑	IBS ↓	MAE_H ↓	KM_C ↓	D-CAL ↑
Random predictions	0.500±0.001	0.500±0.001	0.265±0.026	550.045±31.780	7.070±1.406	0/5
■ Weibull (Weibull, 1951)	0.630±0.046	0.664±0.065	0.187±0.026	537.848±66.633	<u>0.453</u>±0.106	**5**/5
■ CoxPH (Cox, 1972)	0.673±0.028	0.703±0.028	<u>0.175</u>±0.022	505.517±55.501	0.835±0.451	**5**/5
■ RSF (Ishwaran et al., 2008)	0.647±0.023	0.678±0.034	0.178±0.020	529.016±56.860	0.808±0.453	**5**/5
■ S-GB (Chen et al., 2013)	0.690±0.046	<u>0.736</u>±0.043	0.199±0.029	**446.606**±49.178	1.263±0.410	1/5
■ S-XGB (Barnwal et al., 2021)	0.665±0.045	0.694±0.052	0.223±0.044	547.653±34.185	1.560±0.812	1/5
■ DeepSurv (Katzman et al., 2018)	<u>0.694</u>±0.019	0.728±0.031	0.181±0.025	490.405±63.531	0.860±0.438	**5**/5
■ DeepHit (Lee et al., 2018)	0.663±0.021	0.693±0.022	0.225±0.041	579.465±98.572	3.076±2.589	**5**/5
■ TransDSA (Hu et al., 2021)	0.676±0.030	0.724±0.015	0.180±0.016	505.351±36.563	**0.444**±0.086	0/5
■ H2DGSurv (ours)	**0.726**±0.013	**0.763**±0.024	**0.165**±0.033	<u>465.082</u>±81.309	0.486±0.102	<u>4</u>/5

Table 1: Prediction performance in survival analysis. Mean and standard deviation are reported using a 5-fold cross-validation. Random predictions are from a uniform distribution. **Best results** are in bold, <u>second best</u> are underlined.
■ Classical Stastical. ■ Traditional Machine Learning. ■ Deep Learning Survival.

Evaluation on Survival Analysis. Table 1 details discrimination, accuracy and calibration metrics. H2DGSurv yields the best performance on the C-index, td-AUC and IBS, demonstrating an overall improvement of $+\Delta 9\%$ in C-index, $+\Delta 9\%$ in td-AUC and $+\Delta 14\%$ in IBS. Additionally, our proposed method achieves competitive results when compared to Deep Learning Survival state-of-the-art baselines (■). Specifically, H2DGSurv shows an overall improvement of $+\Delta 11\%$ in MAE_H, and achieves the second best results in KM_C.

Evaluation on Event Prediction. We complement standard survival evaluation with 3-year and 5-year mortality prediction, two clinically meaningful endpoints for oncologists (Dörrich et al., 2024). For a given horizon t, we interpret the model output as a survival probability $\hat{p}_i = \hat{S}_i(t) \in [0, 1]$. Patients censored before t are excluded from the horizon-based evaluation. Among the remaining cases, individuals who experienced the event prior to t are assigned positive labels ($y = 1$), and all others are considered negative ($y = 0$). Table 2 reports classification metrics to quantify discrimination performance at each horizon. For a 3-year survival prediction, H2DGSurv achieves the best results, with an overall improvement of $+\Delta 25\%$ in F1-Score and $+\Delta 14\%$ in AUROC. At horizon 5-year, our proposed model yields the best AUROC, demonstrating an overall improvement of $+\Delta 9\%$, highlighting H2DGSurv ability to deliver competitively accurate survival predictions.

Table 2: Evaluation to 3 and 5-year death prediction task, filtered on uncensored patients at each time horizon. **Best results** are in bold, second best are underlined. ■ Classical Stastical. ■ Traditional Machine Learning. ■ Deep Learning Survival.

	3-year prediction		5-year prediction	
Method	**F1-Score**	**AUROC**	**F1-Score**	**AUROC**
Random predictions	0.411±0.038	0.500±0.001	0.591±0.049	0.500±0.001
■ Weibull (Weibull, 1951)	0.402±0.055	0.622±0.064	0.588±0.055	0.648±0.031
■ CoxPH (Cox, 1972)	0.415±0.150	0.724±0.035	0.584±0.033	0.687±0.034
■ RSF (Ishwaran et al., 2008)	0.452±0.084	0.679±0.056	0.613±0.036	0.677±0.020
■ S-GB (Chen et al., 2013)	0.469±0.050	0.698±0.050	0.619±0.043	0.717±0.038
■ S-XGB (Barnwal et al., 2021)	0.385±0.030	0.648±0.050	0.597±0.043	0.657±0.039
■ DeepSurv (Katzman et al., 2018)	0.492±0.085	0.745±0.037	0.626±0.023	0.729±0.032
■ DeepHit (Lee et al., 2018)	0.434±0.089	0.715±0.043	0.634±0.055	0.713±0.043
■ TransDSA (Hu et al., 2021)	0.512±0.047	0.733±0.048	**0.656**±0.069	0.725±0.032
■ H2DGSurv (ours)	**0.554**±0.064	**0.789**±0.047	0.632±0.022	**0.758**±0.025

5.2. Ablation study

Effect of the aggregation module. To evaluate the effectiveness of our proposed hierarchical heterogeneous graph representation, we conduct an ablation study in which H2DGSurv is replaced by several widely used feature aggregation baselines: (i) mean pooling over all modality features, (ii) a Transformer encoder allowing full pairwise feature interactions via self-attention, (iii) a Graph Convolutional Network applied to a fully connected feature graph, and (iv) a multilayer perceptron operating on the concatenation of all features. All models share the same survival prediction head and hyperparameters to ensure a fair comparison. As reported in Table 3, H2DGSurv improves C-index by $+\Delta 8\%$ over the graph convolution baseline, and $+\Delta 13\%$ over the transformer encoder baseline. This suggests that explicitly modeling hierarchical structure and heterogeneity provides substantial benefits over fully connected or structure-agnostic aggregation strategies.

	Survival Analysis			3-year prediction	
Module	**C-index** ↑	**IBS** ↓	**KM$_C$** ↓	**F1-Score** ↑	**AUROC** ↑
Mean pooling	0.568±0.084	0.210±0.031	0.521±0.139	0.362±0.042	0.594±0.117
Transformer Encoder	0.633±0.053	0.211±0.039	0.732±0.269	0.428±0.080	0.711±0.080
GraphConv	0.667±0.020	0.238±0.037	1.136±0.265	0.449±0.095	0.724±0.038
Multilayer Perceptron	0.672±0.063	0.212±0.022	0.643±0.147	0.522±0.023	0.755±0.040
H2DGSurv (ours)	**0.726**±0.013	**0.165**±0.033	**0.486**±0.102	**0.554**±0.064	**0.789**±0.047

Table 3: Ablation study on feature aggregation strategies. We compare our hierarchical directed heterogeneous graph (H2DG) with commonly used aggregation modules.

Table 4: Leave-one-out ablation on the clinical steps, hierarchical levels, and heterogeneity. For each section, **worst results** are in bold, and second-worst are underlined to mark which component have the largest performance contribution.

	Survival Analysis			3-year prediction	
Module	**C-index** ↑	**IBS** ↓	**KM$_C$** ↓	**F1-Score** ↑	**AUROC** ↑
H2DGSurv (ours)	0.726±0.013	0.165±0.033	0.486±0.102	0.554±0.064	0.789±0.047
(1.a) without step 1	**0.553**±0.033	**0.232**±0.049	**1.023**±0.875	**0.408**±0.069	**0.588**±0.049
(1.b) without step 2	0.699±0.032	0.194±0.027	0.618±0.195	0.497±0.067	0.747±0.059
(1.c) without step 3	0.661±0.060	0.198±0.036	0.586±0.296	0.419±0.124	0.725±0.023
(1.d) without step 4	0.700±0.029	0.199±0.022	0.609±0.149	0.460±0.057	0.756±0.049
(2.a) without level 2	0.711±0.013	**0.204**±0.061	**0.733**±0.485	0.544±0.043	0.783±0.038
(2.b) without level 3	**0.696**±0.044	0.190±0.031	0.595±0.197	**0.491**±0.013	**0.749**±0.061
(3) without heterogeneity	0.512±0.027	0.199±0.029	0.531±0.222	0.411±0.038	0.500±0.001

Effect of clinical steps (1.a) → (1.d). We validate the effectiveness of each clinical step by excluding the corresponding nodes from H2DGSurv. Table 4 presents a comparison of key survival analysis and 3-year prediction metrics in our H2DGSurv framework with and without each clinical step. The largest drop in performance comes from removing step 1, which includes blood and clinical data at initial cancer diagnosis, yielding a -Δ24% drop in C-index. The second largest drop is observed by removing step 3, associated with indication reports preceding surgery, which leads to a -Δ9% decline. Removing Steps 2 and 4 similarly reduces performance, though to a lesser extent. These results indicate that each stage provides complementary prognostic information, with early clinical and laboratory data forming the most critical component for accurate survival prediction.

Effect of hierarchical levels (2.a) & (2.b). Table 4 further evaluates the contribution of each hierarchical level by removing the nodes associated with that level. A large performance decrease occurs when removing level 3, corresponding to modality-specific leaf nodes, which results in a -Δ4% drop in C-index. Incorporating all hierarchical levels jointly yields the best performance, indicating that the model benefits from integrating fine-grained modality information, step-level clinical context, and patient-level global representations.

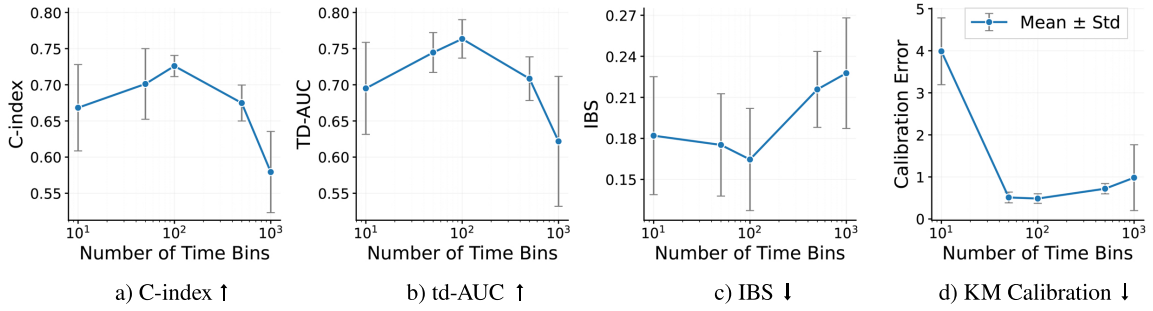
Effect of heterogeneity (3). Disabling heterogeneous node and edge types leads to a significant drop across all metrics. This suggests that modeling modality-specific semantics, via heterogeneous message passing, is essential for effective cross-modal feature aggregation.

Effect of the number of bins. Figure 2 shows that $M = 100$ bins yields the best overall combination of discrimination and calibration. Both coarser and finer discretizations lead to performance drops on most metrics illustrating a bias–variance trade-off.

5.3. Qualitative results

Figure 3 provides examples of two predictions with the predicted survival curves alongside the observed time-to-event. Example 1 represents a case where the event occurred within three years post-diagnosis, while Example 2 corresponds to an event time exceeding three years, highlighting H2DGSurv ability to make accurate predictions from multimodal inputs.

Figure 2: Effect of the number of bins $M \in \{10, 50, 100, 500, 1000\}$ from the hazard head.



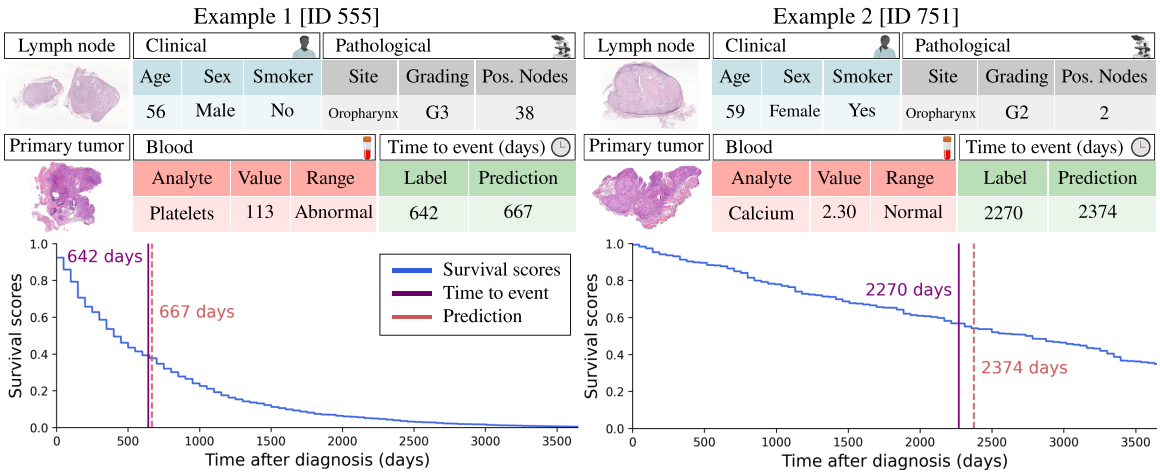


Figure 3: Examples of patient-level survival predictions and survival curves. Example 1 (left) corresponds to a patient who experienced an event within 3 years of diagnosis, whereas Example 2 (right) illustrates a patient whose event occurred after 3 years. Given the large number of input modalities, only a randomly selected subset of covariates is visualized for clarity.

6. Conclusion and Discussion

We present H2DGSurv, a hierarchical directed heterogeneous graph framework for survival analysis in head and neck cancer. Experiments show that our approach demonstrate state-of-the-art performance in survival prediction from multimodal data. Ablation studies highlight the impact of hierarchical levels, clinical steps and heterogeneity graph modeling.

Limitations and Future work. Due to the limited number of monocentric retrospective publicly available datasets for head and neck cancer (Dörrich et al., 2024), our study is limited to a single institution dataset. Future work includes extending our adaptable framework to other modalities for broader medical applicability.

Acknowledgments. We acknowledge contributors from `HANCOCK` (Dörrich et al., 2024).

References

Avinash Barnwal, Hyunsu Cho, and Toby Dylan Hocking. Survival regression with accelerated failure time model in XGBoost, August 2021. arXiv:2006.04920 [cs].

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024. ISSN 1542-4863.

Shaked Brody, Uri Alon, and Eran Yahav. How Attentive are Graph Attention Networks?, January 2022.

C. C. Brown. On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics*, 31(4):863–872, December 1975. ISSN 0006-341X.

Volker Budach and Ingeborg Tinhofer. Novel prognostic clinical factors and biomarkers for outcome prediction in head and neck cancer: a systematic review. *The Lancet. Oncology*, 20(6):e313–e326, June 2019. ISSN 1474-5488.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H. Song, Muhammad Shaban, Mane Williams, Anurag Vaidya, Sharifa Sahai, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Walt Williams, Long Phi Le, Georg Gerber, and Faisal Mahmood. A General-Purpose Self-Supervised Model for Computational Pathology, August 2023.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. ISSN 1546-170X. Publisher: Nature Publishing Group.

Yifei Chen, Zhenyu Jia, Dan Mercola, and Xiaohui Xie. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and Mathematical Methods in Medicine*, 2013:873595, 2013. ISSN 1748-6718.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 0035-9246. Publisher: [Royal Statistical Society, Oxford University Press].

Marion Dörrich, Matthias Balk, Tatjana Heusinger, Sandra Beyer, Hassan Kanso, Christian Matek, Arndt Hartmann, Heinrich Iro, Markus Eckstein, Antoniu-Oreste Gostian, and Andreas M. Kist. A multimodal dataset for precision oncology in head and neck cancer, May 2024. Pages: 2024.05.29.24308141.

Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, and Rajesh Ranganath. X-CAL: Explicit Calibration for Survival Analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 18296–18307. Curran Associates, Inc., 2020.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. *A New Model for Earning in Raph Domains*, volume 2. January 2005. ISBN 978-0-7803-9048-5. doi: 10.1109/IJCNN.2005.1555942.

Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis, September 2023. arXiv:2309.07400 [cs].

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *J. Mach. Learn. Res.*, 21(1):85:3289–85:3351, January 2020. ISSN 1532-4435.

Kyunghwa Han and Inkyung Jung. Restricted Mean Survival Time for Survival Analysis: A Quick Guide for Clinical Researchers. *Korean Journal of Radiology*, 23(5):495, 2022. ISSN 1229-6929, 2005-8330.

Shi Hu, Egill Fridgeirsson, Guido van Wingen, and Max Welling. Transformer-Based Deep Survival Analysis. In *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, pages 132–148. PMLR, May 2021. ISSN: 2640-3498.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), September 2008. ISSN 1932-6157. arXiv:0811.1645 [stat].

E.L. Kaplan and P Meier. Nonparametric Estimation from Incomplete Observations., 1958.

Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, 18(1): 24, December 2018. ISSN 1471-2288. arXiv:1606.00931 [stat].

Sein Kim, Namkyeong Lee, Junseok Lee, Dongmin Hyun, and Chanyoung Park. Heterogeneous Graph Learning for Multi-modal Medical Data Analysis, March 2023. arXiv:2211.15158 [cs].

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, February 2017.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468, 2159-5399.

Stanislav Sobolevsky. Hierarchical Graph Neural Networks, May 2021. arXiv:2105.03388 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks, February 2018.

Ching-Wei Wang, Muhammad-Adil Khalil, and Nabila Puspita Firdi. A Survey on Deep Learning for Precision Oncology, June 2022.

Waloddi Weibull. A Statistical Distribution Function of Wide Applicability. Journal of Applied Mechanics:293–297, 1951.

Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V. Chawla. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 793–803, Anchorage AK USA, July 2019. ACM. ISBN 978-1-4503-6201-6.