

The Cost of Binarizing Survival Outcomes in Clinical Prognostic Modeling

Shashank Yadav

*Department of Radiation Oncology
Mayo Clinic, Rochester, MN, USA 55905*

YADAV.SHASHANK@MAYO.EDU

David M. Routman

*Department of Radiation Oncology
Mayo Clinic, Rochester, MN, USA 55905*

ROUTMAN.DAVID@MAYO.EDU

Andrew Y.K. Foong

*Department of Radiation Oncology
Mayo Clinic, Rochester, MN, USA 55905*

FOONG.ANDREW@MAYO.EDU

Abstract

Survival analysis is an established framework for analyzing time-to-event data, yet many clinical machine learning studies still binarize the outcome before model training. This practice excludes censored patients, collapses temporal information into a single threshold, and can affect which features are selected as prognostically relevant. We examine the cost of this binarization in the context of Bayesian network (BN) feature selection, using two recent publications as case studies: one that applies BN-based feature selection to a head-and-neck cancer cohort and a second surgical cohort study that, while not BN-based, likewise binarizes its survival endpoint. We replace the binary scoring function with the Cox partial log-likelihood for feature-to-outcome edges, a modification we call the Survival-Aware Bayesian network, and recover prognostic features that binarization misses. Our ablation experiment confirms that the improvement is driven by the time-to-event scoring formulation rather than by retaining more patients. The results generalize across five endpoint-cohort combinations in head-and-neck cancer and extend to three further cancer types (breast, colorectal, and kidney). We propose that clinical studies with survival outcomes should use time-to-event methods by default, as binarization discards the prognostic signal retained by survival analysis.

1. Introduction

Survival analysis has been the standard framework for time-to-event outcomes, with well-understood methods for handling censoring, estimating hazard functions, and comparing treatment effects (Kleinbaum and Klein, 2012). However, a large fraction of clinical machine learning studies still convert time-to-event outcomes into binary targets before predictive modeling. The practice is especially prevalent in Bayesian network (BN) modeling of cancer outcomes (Gevaert et al., 2006; Rabinowicz et al., 2017; Shah et al., 2026). For instance, in head-and-neck cancer (HNC), Kotevski et al. (2023) predicted 2-year HNC survival as a binary endpoint across multiple institutions. Kazmierski et al. (2023) chose binary 2-year survival for multi-institutional HNC modeling, noting that this formulation “is commonly used in the literature.”

Binarization, however, is not a neutral modeling choice. A time-to-event outcome, when collapsed into a fixed-horizon indicator, discards information about when events occur, inconsistently censors patients depending on follow-up length, and limits the use of hazard ratios and time-dependent risk estimates. These costs are well understood in the survival analysis literature. Still, they are rarely discussed in BN-based prognostic studies, where the graph structure and Markov blanket are otherwise leveraged for interpretable feature selection (Friedman et al., 1997; Pearl, 2009; Lucas et al., 2004; Kyrimi et al., 2021; McLachlan et al., 2020; Tsamardinos et al., 2003; Aliferis et al., 2010).

In this work, we quantify the cost of endpoint binarization using two head-and-neck cancer cohorts as case studies: one based on a well-established public dataset and the other drawn from a newly published cohort. The first case study is a Bayesian network prognostic model on a radiation cohort (Shah et al., 2026) that binarizes survival for BN structure learning; the second is a prognostic analysis on a surgical cohort (Dörrich et al., 2025) that does not use BNs but adopts the same endpoint binarization. For both studies, we show that binarizing the survival endpoint results in a measurable loss of prognostic information. We demonstrate how survival analysis can be performed without sacrificing the interpretability advantages that motivated the use of BNs in the first place. We refer to our approach as Survival-Aware Bayesian network (SA-BN) and evaluate it on these two HNC cohorts, which differ in treatment modality (radiation, $n=2,994$; surgical, $n=763$), and ask three questions: (1) Which features does binary scoring miss and why? (2) Is the improvement driven by the Cox scoring formulation or by retaining more patients? (3) Do the survival-selected features degrade binary classification performance?

Generalizable Insights about Machine Learning in the Context of Healthcare

This work provides three insights relevant to studies that binarize a survival outcome:

- i) **Endpoint binarization hides features with graded survival effects.** Fixed-horizon thresholds (e.g., 2-year survival) may not detect prognostic variables whose influence is continuous rather than concentrated at a cutpoint. In our case studies, switching from a binary to a survival-aware scoring function recovers established HNC prognostic factors that the binarization step misses, including smoking pack-years, overall stage, hemoglobin, neutrophil-to-lymphocyte ratio (NLR), positive lymph node count, and age. Any study that binarizes a survival outcome should verify that its selected features are not a consequence of this formulation.
- ii) **The scoring function determines which features are considered prognostic.** A common response to censoring-induced exclusion is to obtain more data. In the radiation cohort, running the survival-aware pipeline on the same 2,635 patients with sufficient (≥ 2 years) survival information recovers 8 of 9 Markov blanket features; the additional 359 patients with insufficient survival information contribute only one. The endpoint-scoring formulation primarily determines which features are considered prognostic.
- iii) **Survival-aware feature selection matches or improves binary performance while enabling effect estimation in units of time.** When features selected under survival-aware scoring are used for binary 2-year classification, performance

improves over features selected under binary scoring. The same features also support interventional RMST, which expresses each factor in months of expected life after confounder adjustment. A binary-proxy model cannot produce these quantities for any feature, including the ones it selects.

2. Related Work

Binarization of survival endpoints. The practice of converting time-to-event outcomes into binary classification targets is widespread in clinical ML. In head-and-neck cancer alone, [Shah et al. \(2026\)](#) used 2-year survival for BN-based feature selection on the RADCURE cohort ([Welch et al., 2024](#)). [Diamant et al. \(2019\)](#) trained deep learning models on binary 2-year overall survival, and [Starke et al. \(2023\)](#) adopted the same endpoint for multi-institutional radiomics modeling. We hypothesize that in each case, the binary endpoint method was chosen for methodological convenience rather than clinical necessity. The clinical question concerns the time to death, which a fixed threshold only partially summarizes. No prior work has systematically quantified the cost of this binarization in terms of feature-selection quality or tested whether the lost features can be recovered by switching to a time-to-event scoring function.

Bayesian network Markov blankets and the target specification problem. The Markov blanket (MB) of a BN is the minimal set of variables that makes the target conditionally independent of all other variables, with theoretical optimality guarantees under faithfulness ([Tsamardinos et al., 2003](#); [Aliferis et al., 2010](#)). These guarantees are conditional on the target being correctly specified. We argue that when the target is a binary proxy for a continuous survival outcome, the MB is optimal for the proxy, not for survival itself. A separate line of work ([Scutari et al., 2019](#)) has shown that no single BN structure-learning algorithm dominates across datasets and recommends ensemble approaches. Our method addresses both issues: we replace the binary target with a survival-aware scoring function and require consensus across three structurally different BN learning algorithms (Appendix A).

Survival-native feature selection. Feature selection methods such as LASSO-penalized Cox regression ([Tibshirani, 1997](#)), random-survival-forest variable importance ([Ishwaran et al., 2008](#)), and Cox stepwise selection use time-to-event information without binarizing. Each returns a ranked or penalized feature list, leaving the conditional independence structure among features implicit. SA-BN returns a graph and its Markov blanket, a structure that the ranked selectors do not produce. This blanket serves as the adjustment set for G-computation ([Keil et al., 2014](#)) to estimate the interventional RMST for each prognostic feature.

3. Methods

Our pipeline has four phases: (1) SA-BN structure learning on all patients (including censored) to discover the survival-aware Markov blanket (SA-MB). As a baseline, we compare against the binary-proxy approach of [Shah et al. \(2026\)](#), which learns BN structure on a discretized “SVy2” target (survived ≥ 2 years). (2) Survival modeling with Cox PH, RSF and Gradient Boosting Survival (GBS) models on the selected features ([Cox, 1972](#);

Ishwaran et al., 2008; Hothorn et al., 2006). (3) Causal inference using interventional RMST and hazard ratios to validate that the selected features have interpretable graded effects. (4) Ablation experiments to disentangle the contributions of the Cox scoring formulation from the inclusion of additional censored patients and to verify that survival-selected features do not degrade binary classification.

3.1. Survival-Aware BN Structure Learning (SA-BN)

Standard BN structure learning scores node families using a classification-based scoring function (e.g., BDeu (Heckerman et al., 1995) and K2 (Cooper and Herskovits, 1992)) that requires a discrete target.¹ SA-BN decouples feature–feature edges from feature–target edges, enabling the target node to represent time-to-event information. Feature–target edges are scored by the Cox partial log-likelihood and added by a likelihood-ratio test (LRT, $p < 0.05$), so censored patients contribute to structure learning through their time and event indicators rather than being excluded by a fixed-horizon cutpoint. All scoring equations, algorithm hyperparameters, and cycle-breaking rules are provided in Appendix A.

3.2. Markov Blanket Discovery and Downstream Modeling

We identify the survival Markov blanket by consensus ($\geq 2/3$) across three methods: DAG-derived parents, children, and co-parents of the survival target; Cox stepwise selection; and univariate C-index top- k . For comparison, we also extract the binary-proxy Markov blanket, following Shah et al. (2026), for the SVy2 target. Three survival models (Cox PH, Random Survival Forest, Gradient-Boosted Survival) are then trained on the selected features and evaluated with concordance index (Harrell et al., 1996) (1,000-bootstrap 95% CIs). Causal effects are estimated via interventional RMST by time-to-event G-computation (Keil et al., 2014; Hernán and Robins, 2020), which fits a Cox model with the remaining Markov blanket parents as confounders, applies the do-operation to the feature of interest, and integrates the resulting individual survival curves to 5 years. Algorithm details, scoring thresholds, model hyperparameters, evaluation formulas, and G-computation derivations are in Appendices A, B, and E. Per-feature hazard-ratio p -values from the DAG-informed Cox models are corrected for multiple comparisons within each cohort using the Benjamini–Hochberg false-discovery-rate (FDR) procedure (Benjamini and Hochberg, 1995), and the proportional-hazards assumption is assessed by Schoenfeld-residual tests (Grambsch and Therneau, 1994; Schoenfeld, 1982) with the same FDR control (Appendix G).

4. Cohort

4.1. Cohorts and features

The two HNC cohorts used in this study are summarized in Table 1.

Cohort 1: Radiation: 2,994 patients with HNC treated with definitive radiation at a single institution (2005–2017) (Welch et al., 2024). For binary BN learning, 2,635 patients

1. These are sometimes called multinomial scores because they model the target’s conditional distribution as a multinomial over discrete categories.

Table 1: Cohort characteristics. Survival modeling retains all patients; binary BN learning uses evaluable patients only.

	Radiation	Surgical
Treatment	Definitive RT	Surgery primary
Total patients	2,994	763
Binary-evaluable	2,635	620
Features (BN)	13	15
Temporal split	2015	2016
Train / test (surv.)	2,174 / 820	565 / 198
Event rate	32.7%	27.9%
Has HPV	Yes	Yes
Has invasion markers	No	Yes

are evaluable for 2-year survival (88%), for survival modeling, all 2,994 are retained. Temporal split at year 2015, training $n=2,174$, test $n=820$. Event rate: 32.7% (978 deaths). Thirteen features, including demographics (age, sex), tumor characteristics (T stage, N stage, primary tumor site, Gross Tumor Volume, Histological Subtype), patient factors (Eastern Cooperative Oncology Group Performance Status (ECOG PS), smoking pack-years, Human Papillomavirus status), treatment (modality), overall stage, and a T×N interaction.

Cohort 2: Surgical: 763 patients with HNC treated with primary surgery at a single institution (2005–2019) (Dörrich et al., 2025). For binary BN learning, 620 are evaluable for 2-year survival (81%) and for survival modeling, all 763 are retained. Temporal split at year 2016, training $n=565$, test $n=198$. Event rate: 27.9% (213 deaths). Fifteen features consisting of primary tumor site, pT, pN, grade, HPV, sex, smoking, adjuvant radiotherapy, age, infiltration depth, positive lymph node count, blood hemoglobin, NLR, invasion burden composite, and T×N interaction. Six additional variables available in the raw dataset (resection status, resected lymph node count, resection margin, leukocytes, platelets, platelet-to-lymphocyte ratio (PLR)) were excluded before structure learning due to near-constant values (resection status: 90% R0) or high collinearity with retained features (leukocytes, platelets, and PLR with NLR; resected count with positive count).

5. Results

5.1. SA-BN Recovers Features That Binary Scoring Misses

The binary-proxy and survival-aware Markov blankets are compared and summarized in Table 2. In the radiation cohort, binary scoring excludes 359 patients (12%) and discovers an 8-feature Markov blanket. SA-BN retains all patients and discovers a 9-feature survival Markov blanket (Figure 1). The features unique to SA-BN that no binary implementation recovers are smoking pack-years and overall HNC stage, both with continuous, graded-effect relationships that binarization collapses (Ang et al., 2010; Ma et al., 2022b). SA-BN also drops the T×N interaction, whose signal is subsumed by the T stage and the overall stage

Table 2: All candidate features and their Markov blanket membership. $\bullet$ = selected, $-$ = not selected. SA-BN test-set C-index: radiation 0.801 [.76–.84], surgical 0.726 [.64–.81]; binary: 0.770 [.73–.81], 0.648 [.58–.75].

Radiation Cohort (13 features)			Surgical Cohort (15 features)		
Feature	Binary BN	SA-BN	Feature	Binary BN	SA-BN
Age	$\bullet$	$\bullet$	Tumor Site	$-$	$-$
ECOG PS	$\bullet$	$\bullet$	pT (grouped)	$-$	$-$
GTV	$\bullet$	$\bullet$	pN (grouped)	$\bullet$	$-$
HPV status	$\bullet$	$\bullet$	Grade	$-$	$-$
Tumor Site	$\bullet$	$\bullet$	HPV (p16)	$-$	$-$
T stage	$\bullet$	$\bullet$	Sex	$-$	$-$
Treatment modality	$\bullet$	$\bullet$	Smoking	$-$	$-$
T $\times$ N	$\bullet$	$-$	Adjuvant radiotherapy	$-$	$-$
Sex	$-$	$-$	Age	$-$	$\bullet$
N stage	$-$	$-$	Infiltration depth	$\bullet$	$-$
Histological Subtype	$-$	$-$	Positive Lymph Node count	$-$	$\bullet$
Smoking (pack-yr)	$-$	$\bullet$	Hemoglobin	$-$	$\bullet$
Overall stage	$-$	$\bullet$	NLR	$-$	$\bullet$
			Invasion burden	$\bullet$	$\bullet$
			T $\times$ N	$-$	$-$

separately. In the surgical cohort, SA-BN replaces the binary Markov blanket’s coarse categorical variables (grouped pN and infiltration depth) with continuous alternatives, such as the actual count of positive lymph nodes, blood hemoglobin, and NLR (Figure 2). The SA-BN Markov blanket improves test-set concordance index by +0.031 (radiation) and +0.078 (surgical) while retaining >95% of all-feature performance in both cohorts. Seven of nine radiation features and all five surgical features exceed the 70% robustness threshold across 50 bootstrap resamples (Figure 3), confirming that the feature selection is stable. Per-feature selection frequencies, hazard ratios, BH-FDR q -values, and directed-edge stabilities are reported in Table 10 (Appendix H); all features survive multiple-comparison correction at $q < 0.05$ except radiation T stage.

5.2. Cox scoring recovers the Markov blanket on the binary-evaluable patients

Two factors could explain the difference in the Markov blanket: Cox scoring (formulation) and additional censored patients (data). In the radiation cohort, running SA-BN on the 2,635 patients (excluding 359 censored before 2 years) recovers 8 of 9 survival MB features (Table 3), and only the overall stage is dropped. In the surgical cohort, running SA-BN on the 620 patients (excluding 143 censored before 2 years) recovers all 5 of 5 features and adds T $\times$ N. In both cohorts, the Cox scoring formulation is the primary driver, and the extra censored patients contribute at most one feature.

5.3. Switching to survival MB features is also beneficial for binary classification

We then ask whether features selected for survival prediction remain useful for binary 2-year classification. Table 4 compares the two MBs on binary prediction for both cohorts. In the

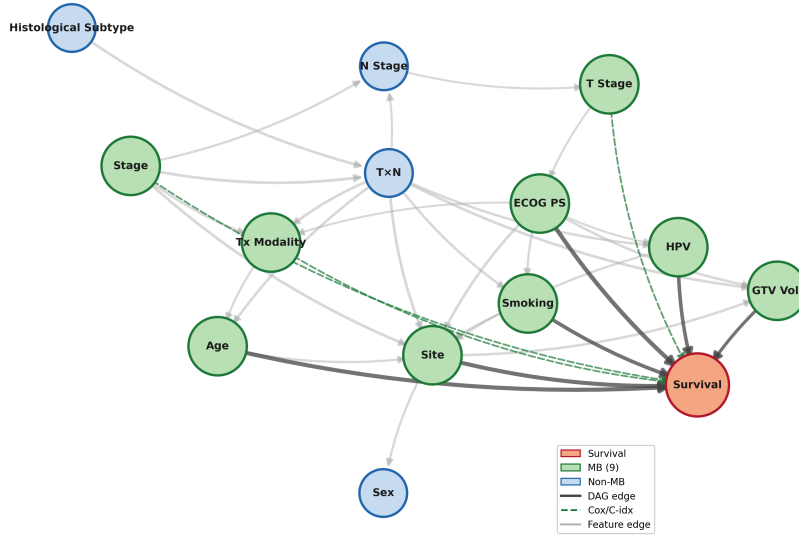


Figure 1: SA-BN consensus structure. Radiation cohort (14 nodes, 28 edges, 9-feature Markov blanket). Orange = survival target, green = consensus survival Markov blanket, blue = non-Markov blanket. Dark arrows = DAG-derived edges to survival; dashed green = Cox/C-index selected; gray = feature–feature edges.

radiation cohort, the SA-BN MB improves binary classification on all metrics: AUC-ROC from 0.797 to 0.826 (+0.029), AUC-PR from 0.943 to 0.950, and MCC from 0.369 to 0.407. The SA-BN features (which include smoking and overall stage) provide information that helps discriminate the 2-year thresholds and predict survival time. In the surgical cohort, the SA-BN MB also improves binary classification: AUC-ROC from 0.740 to 0.807 and MCC from 0.266 to 0.487. The test set here is small (35 events), so the bootstrap confidence intervals overlap, and this improvement is directional. Across both cohorts, the survival-aware feature set improves performance on the binary 2-year task while also supporting survival curves and interventional RMST, which a binarized endpoint cannot provide. The survival-modeling advantage is model-dependent: in the surgical cohort, linear Cox and LASSO-Cox slightly favor the binary feature set on discrimination (Appendix I), and the survival-aware gain comes from RSF and the integrated Brier score.

5.4. Feature-Level Interventional RMST

Binary BNs can report $P(SV_{y2} = 1 \mid \text{do}(X = x))$ via do-calculus (Pearl, 2009; Shah et al., 2026), but this collapses survival beyond 2 years into a single probability. The survival-endpoint analog, interventional RMST via G-computation, preserves the full temporal structure and expresses effects in months of life (Figures 6 and 7). In the radiation cohort, moving from ECOG PS 0 to PS 3+ is associated with a loss of 8 months of interventional RMST after adjustment for confounders, consistent with the established prognostic role of performance status in HNC (Chalker et al., 2022). HPV-positive patients gain 8 months of interventional RMST relative to HPV-negative patients, in line with the RTOG 0129 finding

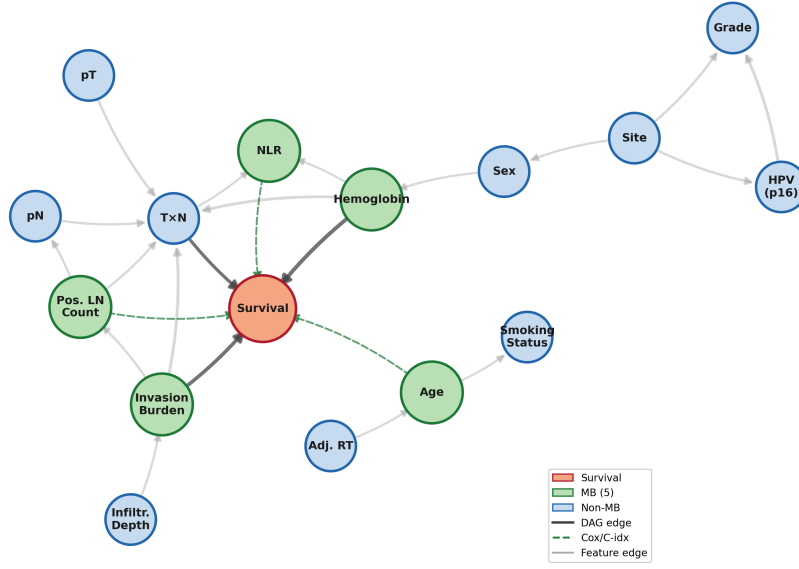
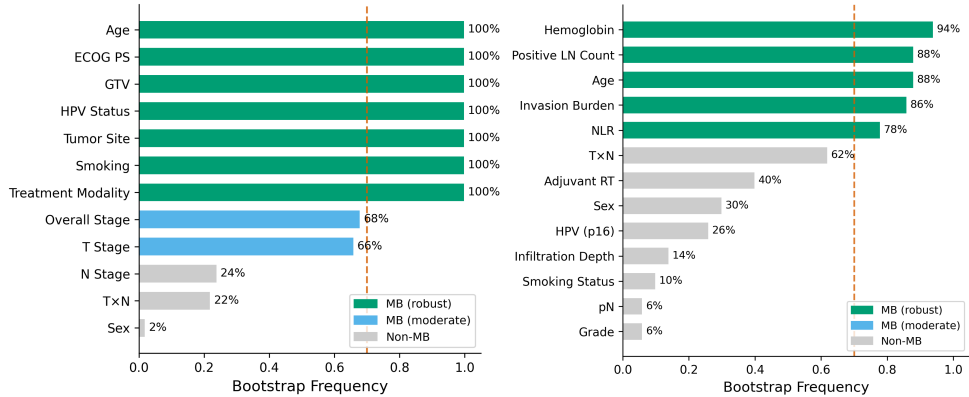


Figure 2: SA-BN consensus structure. Surgical cohort (16 nodes, 25 edges, 5-feature Markov blanket). Orange = survival target, green = consensus survival Markov blanket, blue = non-Markov blanket. Dark arrows = DAG-derived edges to survival; dashed green = Cox/C-index selected; gray = feature–feature edges.



(a) Radiation Cohort: 7/9 Markov blanket at 100%; stage (68%), T (66%) moderate. (b) Surgical Cohort: all 5 Markov blanket members above 78%.

Figure 3: SA-BN Markov blanket stability across 50 bootstrap resamples. Dashed line = 70% threshold.

that HPV is the strongest independent prognostic factor in oropharyngeal cancer (Ang et al., 2010). In the surgical cohort, hemoglobin ≤ 12.1 g/dL costs 13.9 months of interventional RMST relative to 15.1–16.2 g/dL, reflecting the well-documented dose-dependent association

Table 3: Ablation: all candidate features and their MB membership across three conditions.
 • = selected, – = not selected. Binary uses evaluable patients only (2,635/620);
 SA-BN (eval.) uses the same patients with Cox scoring; SA-BN (all) adds censored
 patients.

Radiation Cohort (13 features)				Surgical Cohort (15 features)			
Feature	Binary	SA-BN (ev.)	SA-BN (all)	Feature	Binary	SA-BN (ev.)	SA-BN (all)
Age	•	•	•	Tumor Site	–	–	–
ECOG PS	•	•	•	pT (grouped)	–	–	–
GTV	•	•	•	pN (grouped)	•	–	–
HPV status	•	•	•	Grade	–	–	–
Tumor Site	•	•	•	HPV (p16)	–	–	–
T stage	•	•	•	Sex	–	–	–
Treatment modality	•	•	•	Smoking	–	–	–
T×N	•	–	–	Adjuvant RT	–	–	–
Sex	–	–	–	Age	–	•	•
N stage	–	–	–	Infiltration depth	•	–	–
Histological Subtype	–	–	–	Positive LN count	–	•	•
Smoking (pk-yr)	–	•	•	Hemoglobin	–	•	•
Overall stage	–	–	•	NLR	–	•	•
				Invasion burden	•	•	•
				T×N	–	•	–
MB	8	8	9	MB	3	6	5

Table 4: Binary 2-year classification on evaluable patients. Mean $\pm$ std, 200 bootstrap resamples. Bold: best per cohort per model. LR: Logistic Regression; RF: Random Forest

Cohort	Features	Model	AUC-ROC	AUC-PR	F1	MCC
Radiation	Binary MB (8)	LR	.797 $\pm$.024	.943 $\pm$.011	.815 $\pm$.016	.369 $\pm$.041
	Binary MB (8)	RF	.785 $\pm$.026	.938 $\pm$.012	.828 $\pm$.016	.329 $\pm$.043
	SA-BN MB (9)	LR	.826 $\pm$.023	.950 $\pm$.011	.836 $\pm$.015	.407 $\pm$.043
	SA-BN MB (9)	RF	.820 $\pm$.024	.948 $\pm$.011	.849 $\pm$.014	.404 $\pm$.045
Surgical	Binary MB (3)	LR	.697 $\pm$.061	.903 $\pm$.029	.843 $\pm$.024	.231 $\pm$.094
	Binary MB (3)	RF	.740 $\pm$.053	.925 $\pm$.023	.860 $\pm$.023	.266 $\pm$.097
	SA-BN MB (5)	LR	.737 $\pm$.058	.933 $\pm$.021	.872 $\pm$.022	.239 $\pm$.097
	SA-BN MB (5)	RF	.807 $\pm$.052	.950 $\pm$.019	.935 $\pm$.015	.487 $\pm$.102

between pretreatment anemia and survival in HNC (Prosnitz et al., 2005; Ma et al., 2022a). Positive lymph node count ≥ 8 costs 12 months, mirroring the continuous mortality gradient reported for metastatic nodal burden in oral cavity cancer (Roberts et al., 2016; Lee et al., 2019). NLR ≥ 4.4 costs 7 months, consistent with the meta-analytic HR of 1.69–1.84 for elevated NLR across HNC cohorts (Yang et al., 2018; Takenaka et al., 2018). All SA-BN

Table 5: SA-BN Markov blanket membership across survival endpoints. • = selected, – = not selected. Features appearing across multiple endpoints are more robust to endpoint choice.

Radiation cohort			Surgical cohort			
Feature	OS	Recurrence	Feature	OS	PFS	Recurrence
Age	•	–	Age	•	•	–
ECOG PS	•	•	Hemoglobin	•	•	–
GTV	•	•	NLR	•	•	–
HPV status	•	–	Positive LN count	•	•	•
Tumor Site	•	•	Invasion burden	•	–	•
T stage	•	–	T×N	–	•	–
Treatment modality	•	•	Smoking	–	•	–
Smoking (pack-yr)	•	–	Adjuvant RT	–	–	•
Overall stage	•	–	pT (grouped)	–	–	–
T×N	–	•	pN (grouped)	–	–	–
N stage	–	–	Grade	–	–	–
Sex	–	–	HPV (p16)	–	–	–
Histological Subtype	–	–	Sex	–	–	–
			Infiltration depth	–	–	–
			Tumor Site	–	–	–
MB	9	5	MB	5	6	3
Events	978	303	Events	213	297	177

unique features (except T stage) retain monotone graded effect gradients after confounding adjustment. DAG-informed hazard ratios confirm these patterns in standard clinical units of Hazard Ratios (HR) (ECOG PS HR 1.49–4.22; invasion burden HR 1.22–2.45; hemoglobin HR 0.85; Table 8, Appendix F).

5.5. Generalization Across Survival Endpoints

We repeat the SA-BN analysis on loco-regional recurrence (radiation cohort: 303 events, 10.1%; 22.3% excluded by binarization), PFS (surgical cohort: 297 events, 39%) and recurrence (surgical cohort: 177 events, 23%). Table 5 shows that the binarization cost applies across all endpoint types. In the radiation cohort, the recurrence Markov blanket shares 4 of 9 OS features and drops OS-specific factors (HPV, smoking, staging). In the surgical cohort, 4 of 5 OS members reappear in PFS; positive LN count is the only feature shared across all three surgical endpoints. Continuous biomarkers (hemoglobin, NLR, positive LN count) consistently appear in the SA-BN Markov blanket and are consistently missed by other methods, regardless of endpoint.

5.6. Generalization Across Cancer Types

Both primary cohorts are single-institution and head-and-neck. To test whether the binarization cost is specific to that setting, we repeat the comparison on three multi-institutional cohorts from The Cancer Genome Atlas (TCGA) spanning different cancer types: breast (BRCA), colorectal (COAD) and kidney (KIRC), using the same 2-year binarization and the

Table 6: Generalization to three multi-institutional TCGA cohorts at the 2-year endpoint. Each cohort reports the binary-proxy and survival-aware Markov blankets, along with their test-set metrics (C-index and time-dependent AUC, higher is better; integrated Brier score, lower is better). “% drop” is the fraction of patients excluded by binarization.

Cohort (n, % drop)	Feature set	C-index	tAUC	IBS↓
BRCA breast (1083, 41%)	Binary (5): adj. pharm/RT, N, PR, stage	0.652	0.647	0.081
	SA-BN (5): age, menopause, N, PR, stage	0.711	0.724	0.077
COAD colorectal (605, 40%)	Binary (2): M stage, pos. LN	0.649	0.668	0.110
	SA-BN (2): age, M stage	0.702	0.738	0.105
KIRC kidney (535, 17%)	Binary (1): overall stage	0.741	0.785	0.143
	SA-BN (5): age, grade, N, M, stage	0.767	0.809	0.142

same SA-BN pipeline (Table 6). On all three, the survival-aware Markov blanket recovers prognostic variables the binary endpoint drops and leads on every metric. In kidney cancer, binarization collapses the blanket to overall stage alone, while SA-BN additionally recovers age, grade and nodal/metastatic stage; in breast cancer, the binary blanket latches onto treatment variables (adjuvant pharmacotherapy and radiotherapy) while SA-BN selects biology (age, menopausal status). The magnitude of the gain tracks the fraction of patients that binarization excludes (BRCA and COAD drop $\sim 40\%$ and gain $+0.06/+0.05$ C-index; KIRC drops 17% and gains $+0.03$), consistent with the mechanism described in Section 5.2 and the boundary conditions in Section 6.4.

6. Discussion

6.1. Binarization Distorts the Markov Blanket

Binarization of time-to-event information fails to identify features that oncologists already consider established prognostic factors for HNC. The ablation in Section 5.2 shows this failure is caused by the scoring function itself, not by sample size. The mechanism is twofold: binarization excludes censored patients (thereby biasing the training set) and collapses continuous, graded effect relationships into a single threshold (making graded effects invisible). We illustrate three examples of how this plays out in practice. i) *Smoking (pack-years)* in the radiation cohort: Smoking is an established prognostic factor in HNC, with a continuous graded effect relationship: heavier smokers have progressively worse survival. Neither Shah et al. (2026) nor our binary baseline selects smoking, because binarizing at 2 years collapses the graded relationship between cumulative tobacco exposure and survival time into a single threshold. SA-BN recovers smoking in 100% of bootstrap resamples because the Cox partial log-likelihood preserves the continuous gradient. ii) *Hemoglobin* in the surgical cohort: Anemia is a known poor prognostic factor in HNC (Hoff, 2012) and its effect is graded: each 1 g/dL increase in hemoglobin multiplies the hazard by 0.85 (Table 8). Binary discretization collapses this continuous graded effect into a single threshold, rendering the gradient invisible. iii) *NLR* in the surgical cohort: Elevated neutrophils

relative to lymphocytes is a blood-based marker of systemic inflammation, indicating a pro-tumor immune environment and impaired anti-tumor surveillance. NLR has a continuous, monotone effect on survival (interventional RMST: 4.30 yr at $\text{NLR} \leq 2.6$ vs. 3.76 yr at $\text{NLR} \geq 4.4$, a 7-month gap after confounder adjustment). Binary scoring misses NLR entirely because this gradient does not concentrate at any single time threshold. SA-BN recovers it at 78% bootstrap stability. All three examples share the same mechanism; the feature has a continuous, graded effect on survival time that binary scoring cannot detect because the signal is spread across the entire survival curve rather than concentrated at the 2-year threshold.

6.2. Clinical Implications

The immediate downstream benefit of survival-aware scoring is that causal effects can be reported in units of time rather than probability. Applying G-computation to the SA-BN Markov blanket yields an interventional RMST for each feature level, which translates directly into months of expected life gained or lost after adjustment for confounders. Hemoglobin, missed by the binary-proxy BN in the surgical cohort, shows a monotonic dose-response in interventional RMST. NLR, also missed by binary scoring, shows the same pattern across its range. These gradients are precisely the signal that binarization erases; neither feature has a sharp effect at the 2-year cutpoint, but both substantially shift survival time across their full ranges. A binary-proxy BN cannot produce these quantities, even for the features it selects. When do-calculus is applied to the SVy2 target, it yields $P(\text{SVy2}=1 \mid \text{do}(X=x))$, which collapses the entire post-2-year survival experience into a single probability and offers no way to distinguish a feature that shortens life by 2 months from one that shortens it by 12.

6.3. Integrating SA-BN into Clinical Workflows

SA-BN is a drop-in change to an existing feature-selection step, and it fits clinical pipelines in three ways. First, the survival-aware Markov blanket is a compact, intake-level risk panel built from variables already recorded at first presentation (age, ECOG performance status, smoking history, blood hemoglobin, nodal count), and it can be scored without additional data collection. Second, the G-computation interventional RMST turns each blanket feature into a patient-facing quantity, e.g., months of expected life under a modifiable exposure, which is more directly actionable at the bedside than a shift in 2-year survival probability; pre-operative anemia correction (hemoglobin) is the clearest such example in the surgical cohort. Third, the blanket can serve purely as a feature-selection front-end ahead of an institution’s existing Cox or RSF prognostic model, supplying a smaller, survival-justified set of covariates without changing the downstream model or its calibration. Each use replaces a binarized feature set with a survival-aware one, and the added computational cost is modest, dominated by a one-time offline stability analysis (Appendix K).

6.4. Scope and Limitations of SA-BN

The advantage of SA-BN is greatest when the dataset contains features with graded survival effects (e.g., hemoglobin, smoking pack-years) and substantial censoring. In the radiation cohort (33% event rate, 12% censored-excluded), SA-BN gains +0.031 C-index. In the surgical

cohort (28% event rate, 19% censored-excluded), it increases by 0.078. The generalization across five endpoint-cohort combinations (Section 5.5) suggests this is not endpoint-specific. If censoring is minimal or all features are binary (e.g., treatment received yes/no), SA-BN and binary BN will likely produce similar BN structures. We also compare SA-BN against established survival-native selectors that avoid binarization: LASSO-Cox and RSF. At matched model size, a paired bootstrap C-index difference test with BH-FDR (Benjamini and Hochberg, 1995) does not distinguish SA-BN from a size-matched LASSO-Cox or RSF variable-importance set on either cohort (all $q > 0.5$; Appendix J). The only significant difference favors a larger, cross-validated 12-feature LASSO-Cox over the 6-feature radiation blanket ($q = 0.04$), where the estimator is fixed, and the feature count varies. SA-BN matches these selectors on prediction and adds what they do not produce: a Markov blanket that serves as the adjustment set for the interventional RMST analysis in Section 5.4.

Limitations. (1) The surgical test set contains only 35 events, so bootstrap CIs on the SA-BN vs. binary C-index comparison (0.726 vs. 0.648) overlap. The direction of the effect is consistent with the radiation cohort, but a larger surgical cohort is needed to confirm the magnitude. (2) SA-BN is a hybrid method: feature–feature edges use BDeu/K2 scoring while target–parent edges use Cox scoring. A fully survival-native scoring function for all edges would be more coherent but requires new theoretical development.

7. Conclusion

The binarization of survival outcomes for BN structure learning is a widespread practice in clinical machine learning; our experiments show that this practice incurs a measurable cost. In the two HNC cohorts, binary scoring misses established prognostic factors (smoking pack-years, overall stage, hemoglobin, NLR) because their effects are graded rather than threshold-based, producing unstable Markov blankets that vary across implementations. The fix requires updating how the graph scores the link between each feature and the survival outcome, while leaving the rest of the pipeline unchanged. A BN trained to predict how long patients survive, rather than whether they cross a 2-year threshold, recovers the established prognostic features and improves test-set performance. Causal analysis confirms that the recovered features have graded effects after adjusting for confounders. A controlled ablation shows that 8 of 9 features are recovered even on the same patients used by the binarized analysis, so the improvement comes from how the BN scores time-to-event survival information rather than from additional data. The survival-selected features also improve binary 2-year classification. The broader lesson is that the choice of endpoint formulation during BN structure learning, not just during model fitting, determines which features are considered prognostic. Clinical studies that binarize a survival outcome should verify that their Markov blanket is not an artifact of that binarization.

Data and Code availability. The radiation primary cohort is derived from the RAD-CURE dataset (Welch et al., 2024) and is available for download from <https://www.cancerimagingarchive.net/collection/radcure/>. The surgical cohort is derived from the HANCOCK dataset (Dörrieh et al., 2025) and is available for download from <https://hancock.research.fau.eu/>. Code is available at <https://github.com/xinformatics/COBS>.

References

- Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification. *Journal of Machine Learning Research*, 11:171–234, 2010.
- K Kian Ang, Jonathan Harris, Richard Wheeler, Randal Weber, David I Rosenthal, Phuc Felix Nguyen-Tân, William H Westra, Christine H Chung, Richard C Jordan, Charles Lu, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *New England Journal of Medicine*, 363(1):24–35, 2010.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.
- Cameron Chalker, Jenna M Voutsinas, Qian Vicky Wu, et al. Performance status (PS) as a predictor of poor response to immune checkpoint inhibitors (ICI) in recurrent/metastatic head and neck cancer (RMHNSCC) patients. *Cancer Medicine*, 11(24):4815–4825, 2022.
- Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.
- David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–220, 1972.
- Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.
- André Diamant, Avishek Chatterjee, Martin Vallières, George Shenouda, and Jan Seuntjens. Deep learning in head & neck cancer outcome prediction. *Scientific reports*, 9(1):2764, 2019.
- Marion Dörrieh, Matthias Balk, Tatjana Heusinger, Sandra Beyer, Hamed Mirbagheri, David J Fischer, Hassan Kanso, Christian Matek, Arndt Hartmann, Heinrich Iro, et al. A multimodal dataset for precision oncology in head and neck cancer. *Nature Communications*, 16(1):7163, 2025.
- Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2):131–163, 1997.
- Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.
- Patricia M Grambsch and Terry M Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

- Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996.
- David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.
- Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.
- Camilla Molich Hoff. Importance of hemoglobin concentration and its modification for the outcome of head and neck cancer patients treated with radiotherapy. *Acta Oncologica*, 51(4):419–432, 2012.
- Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.
- Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.
- Michal Kazmierski, Mattea Welch, Sejin Kim, Chris McIntosh, Katrina Rey-McIntyre, Shao Hui Huang, Tirth Patel, Tony Tadic, Michael Milosevic, Fei-Fei Liu, et al. Multi-institutional prognostic modeling in head and neck cancer: evaluating impact and generalizability of deep learning and radiomics. *Cancer Research Communications*, 3(6):1140–1151, 2023.
- Alexander P Keil, Jessie K Edwards, David B Richardson, Ashley I Naimi, and Stephen R Cole. The parametric g-formula for time-to-event data: intuition and a worked example. *Epidemiology*, 25(6):889–897, 2014.
- David G. Kleinbaum and Mitchel Klein. Survival analysis: a self-learning text, 2012.
- Damian P Kotevski, Robert I Smee, Claire M Vajdic, and Matthew Field. Machine learning and nomogram prognostic modeling for 2-year head and neck cancer-specific survival using electronic health record data: a multisite study. *JCO Clinical Cancer Informatics*, 7:e2200128, 2023.
- Evangelia Kyrimi, Scott McLachlan, Kudakwashe Dube, Mariana R. Neves, Ali Fahmi, and Norman Fenton. A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. *Artificial Intelligence in Medicine*, 117:102108, 2021.
- Hojun Lee et al. Number of positive lymph nodes better predicts survival for oral cavity cancer. *Journal of Surgical Oncology*, 119(6):675–682, 2019.
- Peter J.F. Lucas, Linda C. van der Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30(3):201–214, 2004.
- Sung Jun Ma, Han Yu, Michael Khan, et al. Defining the optimal threshold and prognostic utility of pre-treatment hemoglobin level as a biomarker for survival outcomes in head and neck cancer patients receiving chemoradiation. *Oral Oncology*, 133:106054, 2022a.

- Sung Jun Ma, Han Yu, Brian Yu, Olivia Waldman, Michael Khan, Udit Chatterjee, Sharon Santhosh, Jasmin Gill, Austin J Iovoli, Mark Farrugia, et al. Association of pack-years of cigarette smoking with survival and tumor progression among patients treated with chemoradiation for head and neck cancer. *JAMA network open*, 5(12):e2245818, 2022b.
- Nathan Mantel et al. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–170, 1966.
- Scott McLachlan, Kudakwashe Dube, Graham A. Hitman, Norman E. Fenton, and Evangelia Kyrimi. Bayesian networks in healthcare: Distribution by medical condition. *Artificial Intelligence in Medicine*, 107:101912, 2020.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Robert G Prosnitz, Bin Yao, Christopher L Farrell, Robert Clough, and David M Brizel. Pretreatment anemia is correlated with the reduced effectiveness of radiation and concurrent chemotherapy in advanced head and neck cancer. *International Journal of Radiation Oncology, Biology, Physics*, 61(4):1087–1095, 2005.
- Simon Rabinowicz, Arjen Hommersom, Rebecca Butz, and Marc Williams. A prognostic model of glioblastoma multiforme using survival Bayesian networks. In *Artificial Intelligence in Medicine (AIME 2017)*, volume 10259 of *Lecture Notes in Computer Science*, pages 81–85. Springer, 2017. doi: 10.1007/978-3-319-59758-4_9.
- Timothy J Roberts, A Dimitrios Colevas, Wendy Hara, et al. Number of positive nodes is superior to the lymph node ratio and American Joint Committee on Cancer N staging for the prognosis of surgically treated head and neck squamous cell carcinomas. *Cancer*, 122(9):1388–1397, 2016.
- James Robins. A new approach to causal inference in mortality studies with a sustained exposure period: Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.
- David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.
- Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253, 2019.
- Chirag Shah et al. Clinically interpretable survival risk stratification in head and neck cancer using Bayesian networks and Markov blankets. *International Journal of Radiation Oncology, Biology, Physics*, 124(3):856–867, 2026.
- Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. Springer Science & Business Media, 2012.
- Sebastian Starke, Alexander Zwanenburg, Karoline Leger, Klaus Zöphel, Jörg Kotzerke, Mechthild Krause, Michael Baumann, Esther GC Troost, and Steffen Löck. Longitudinal

- and multimodal radiomics models for head and neck cancer outcome prediction. *Cancers*, 15(3):673, 2023.
- Yukinori Takenaka, Ryohei Oya, Takahiro Kitamiura, et al. Prognostic role of neutrophil-to-lymphocyte ratio in head and neck cancer: A meta-analysis. *Head & Neck*, 40(3):647–655, 2018.
- Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.
- Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Algorithms for large scale Markov blanket discovery. In *Proceedings of the Florida AI Research Society (FLAIRS)*, 2003.
- Mattea L Welch, Sejin Kim, Andrew J Hope, Shao Hui Huang, Zhibin Lu, Joseph Marsilla, Michal Kazmierski, Katrina Rey-McIntyre, Tirth Patel, Brian O’Sullivan, et al. Radcure: An open-source head and neck cancer ct dataset for clinical radiation therapy insights. *Medical Physics*, 51(4):3101–3109, 2024.
- Liang Yang, Yi Huang, Li Zhou, et al. Pretreatment neutrophil to lymphocyte ratio in determining the prognosis of head and neck cancer: a meta-analysis. *BMC Cancer*, 18(1): 383, 2018.
- In-Kwon Yeo and Richard A Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

Appendix A. SA-BN Algorithm Details

Survival scoring for target–parent edges. Let (t_i, δ_i) denote the observed time and event indicator for patient i , and let $\mathbf{x}_i$ be the feature vector. For a candidate parent set Pa of the survival node, the Cox partial log-likelihood is:

$$\ell(\beta; \text{Pa}) = \sum_{i: \delta_i=1} \left[\beta^\top \mathbf{x}_{i, \text{Pa}} - \log \sum_{j \in \mathcal{R}(t_i)} \exp(\beta^\top \mathbf{x}_{j, \text{Pa}}) \right] \quad (1)$$

where $\mathcal{R}(t_i)$ is the risk set and $\mathbf{x}_{i, \text{Pa}}$ is the subvector of features in Pa. Starting from an empty parent set with the null-model log-likelihood $\ell_0 = -\sum_{i: \delta_i=1} \log |\mathcal{R}(t_i)|$, at each iteration we add the candidate X_k whose inclusion yields the smallest likelihood-ratio p -value, provided the test $\Lambda = 2[\ell(\hat{\beta}; \text{Pa} \cup \{X_k\}) - \ell(\hat{\beta}; \text{Pa})] \sim \chi_{\text{df}}^2$ satisfies $p < 0.05$; selection stops when no remaining feature meets this threshold or when the maximum in-degree (5) is reached. Cox models are fit using the Breslow approximation with a small L_2 penalty ($\lambda = 0.01$) for numerical stability (Cox, 1972). All LRT comparisons within a single selection run use the same complete-case subset so that nested models are validly comparable.

We instantiate this survival-aware scoring within three structurally different BN algorithms and take a consensus ($\geq 2/3$) of their outputs, reducing sensitivity to any single search strategy.

- i) SA-HC/Cox: Hill-Climbing with BDeu scoring (Heckerman et al., 1995) (equivalent sample size 10, max in-degree 5) for feature–feature edges; target–parent edges selected by forward Cox LRT as described above ($p < 0.05$).
- ii) SA-HC/LR: Hill-Climbing with K2 scoring (Cooper and Herskovits, 1992) for feature–feature edges; target–parent edges selected by forward conditional multi-group log-rank test (Mantel et al., 1966) ($p < 0.01$), where at each step the candidate X_k is tested for conditional independence from survival given the previously selected parents. Continuous candidates are stratified using the same GMM discretization as feature–feature scoring.
- iii) SA-PC/LR: The PC (Peter–Clark) constraint-based algorithm (Spirtes et al., 2012) with conditioning-set size up to 3 and significance level $\alpha = 0.05$. Feature–feature edges use χ^2 conditional independence tests on discretized strata; feature–target edges use a stratified log-rank test where the conditioning set defines the strata, and strata with fewer than 10 observations are skipped. The stratified p -value combines per-stratum log-rank statistics by summing χ^2 values and degrees of freedom.

Continuous features are discretized via Yeo–Johnson normalization (Yeo and Johnson, 2000) followed by Gaussian mixture clustering, with the number of components selected by AIC from $\{2, \dots, \min(6, \lfloor N/15 \rfloor + 1)\}$ per feature (Dempster et al., 1977). All transformations are fit only to the training split; test-set values are transformed using the fitted parameters. Discretization is applied only to BDeu/K2 feature–feature scoring and to the log-rank tests; the Cox partial log-likelihood for feature–target edges operates on the original continuous values. The survival Markov blanket is identified by three methods applied to the training split and combined by consensus ($\geq 2/3$): i) DAG-derived: parents, children, and co-parents of the survival node in the SA-BN consensus graph. ii) Cox stepwise: forward selection with LRT $p < 0.05$, followed by backward pruning at $p > 0.10$, both using the Cox partial log-likelihood on the complete-case training subset with L_2 penalizer 0.01. iii) Univariate

C-index top- k : features ranked by Harrell’s C-index (Harrell et al., 1996) on the training split; the top k are selected, where $k = \max(|\text{MB}_{\text{DAG}}|, |\text{MB}_{\text{Cox}}|, 5)$. A feature enters the consensus Markov blanket if selected by at least 2 of these 3 methods. The resulting blanket is stable across LRT thresholds $p \in [0.01, 0.10]$ and BDeu equivalent sample sizes $\in [1, 50]$. Markov blanket stability (Figure 3) is assessed by 50 bootstrap resamples of the training split, drawn with replacement at the original sample size. Each resample runs through the full structure-learning, consensus, and MB-discovery pipeline; a feature’s stability is the fraction of resamples in which it enters the consensus Markov blanket. The 70% threshold in Figure 3 corresponds to selection in at least 35 of 50 resamples. All structure learning, MB discovery, and discretization fitting use only the training split; the temporally held-out test set is reserved for final evaluation of the survival model.

Appendix B. BN Structure, Bootstrap Stability and Survival Modeling

Figures 1 and 2 show the SA-BN consensus structures for the radiation and surgical cohorts, respectively. Nodes are colored by their role relative to the survival target, and edges reflect consensus structure across bootstrap resamples. Table 7 compares Cox PH, RSF, and GBS across feature sets. The SA-BN MB retains >95% of all-feature performance in both cohorts (radiation: RSF 0.801 vs. 0.803; surgical: RSF 0.726 vs. 0.758), while non-MB features alone drop to $C = 0.685$ in the radiation cohort. The small gap between MB and all-feature models is expected; the MB is designed to be the minimal *sufficient* set that captures the outcome’s conditional independence structure, not the set that maximizes predictive performance. Its value is interpretive as it answers “which features matter and why” rather than “what is the highest achievable C-index”.

Table 7: Survival model C-index [95% CI]. Bold: best per cohort per feature set.

Features	Model	Radiation	Surgical
All (13/15)	Cox PH	0.792 [.75-.83]	0.686 [.57-.80]
	RSF	0.803 [.77-.84]	0.758 [.67-.83]
	GBS	0.806 [.77-.84]	0.681 [.59-.78]
SA-MB (9/5)	Cox PH	0.791 [.75-.83]	0.625 [.51-.74]
	RSF	0.801 [.76-.84]	0.726 [.64-.81]
	GBS	0.798 [.76-.83]	0.672 [.57-.77]
Non-MB	Cox PH	0.681 [.63-.73]	0.681 [.59-.77]
	RSF	0.685 [.64-.73]	0.736 [.65-.81]

Appendix C. Risk Stratification

We stratify patients into tertiles based on RSF-predicted risk scores using SA-MB features. In both cohorts, the resulting Kaplan–Meier curves are well separated (Figure 4). In the radiation cohort, the high-risk tertile has approximately 40% survival at 4 years, compared with over 90% in the low-risk tertile. In the surgical cohort, separation is smaller but still

significant ($p < 0.001$), reflecting the wider confidence intervals from the smaller test set ($n=198$, 35 events).

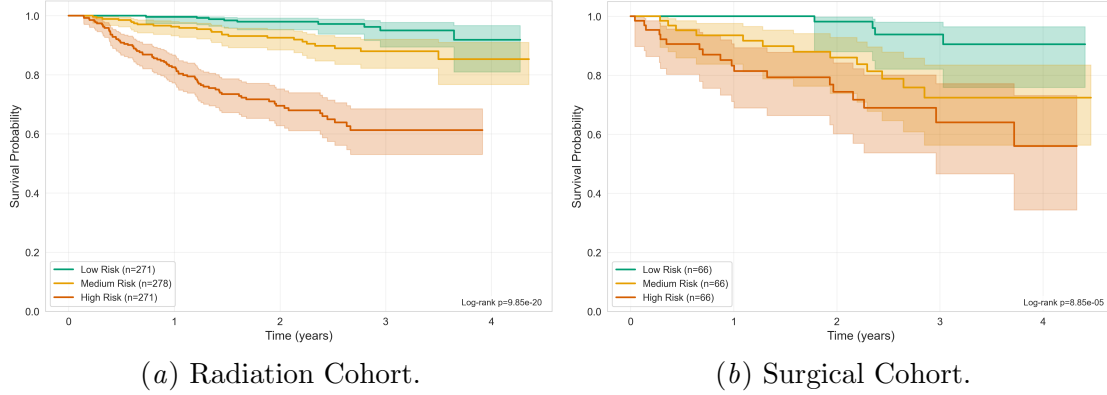


Figure 4: KM survival curves stratified by RSF risk score tertiles using SA-BN MB features.

Appendix D. Ablation: Feature Selection by Method

Table 3 compares feature selection across three conditions: binary-proxy BN (evaluable patients only), SA-BN on evaluable patients only (same patients as binary, but with Cox scoring), and SA-BN on all patients (Cox scoring + censored patients included). In the radiation cohort, SA-BN on evaluable patients recovers 8 of 9 features, confirming that the Cox scoring formulation is the primary driver; the extra 359 censored patients add only overall stage. In the surgical cohort, SA-BN on evaluable patients recovers all 5 features (plus $T \times N$). Features not selected by any method (sex, N stage, Histological Subtype on radiation; site, pT, grade, HPV, sex, smoking on surgical) are consistently excluded, confirming that the MB differences are not random.

Appendix E. G-Computation Details

For a feature X_k at level x , the interventional RMST is:

$$\text{RMST}(\text{do}(X_k=x)) = \int_0^\tau \frac{1}{n} \sum_{i=1}^n \hat{S}(t | X_k=x, \mathbf{C}_i) dt \quad (2)$$

where $\tau=5$ years, $\hat{S}$ is the Cox survival function, and $\mathbf{C}_i$ are the other Markov blanket parents at patient i 's actual values. Setting $X_k=x$ for all patients while keeping $\mathbf{C}_i$ fixed implements the do-operation; averaging marginalizes out confounders (Robins, 1986; Hernán and Robins, 2020). Shah et al. (2026) use do-calculus on the BN's conditional probability tables for the same purpose, which requires manually reversing edges from the outcome to its predictors before the intervention can be computed. G-computation avoids this by using the BN to identify the adjustment set and delegating causal estimation to a Cox model that does not depend on edge directions. The key assumption is that the MB parents constitute

a sufficient adjustment set for the backdoor criterion, which is plausible but not guaranteed from observational data (Hernán and Robins, 2020). Bootstrap 95% CIs from 50 resamples are reported.

Appendix F. Causal Analysis: Hazard Ratios and Interventional RMST

The hazard ratios for the primary SA-BN exposures are shown in Table 8 and Figure 5, complementing the interventional RMST figures (Figures 6 and 7) with per-unit effect sizes. In the radiation cohort, ECOG performance status exhibits a monotonic hazard gradient from PS 1 (HR = 1.49) to PS 4 (HR = 4.22), supporting its selection as a direct survival parent. In the surgical cohort, invasion burden shows a threshold effect at $IB \geq 2$, with hemoglobin as a continuous risk factor.

Table 8: Causal hazard ratios from DAG-informed Cox PH. $^*p < 0.05$.

Cohort	Variable	HR	p
Radiation	ECOG PS 1 vs. 0	1.485	$< 0.001^*$
Radiation	ECOG PS 2 vs. 0	2.120	$< 0.001^*$
Radiation	ECOG PS 3 vs. 0	2.646	$< 0.001^*$
Radiation	ECOG PS 4 vs. 0	4.216	$< 0.001^*$
Surgical	Age (per year)	1.017	0.005^*
Surgical	Hemoglobin (per g/dL)	0.853	$< 0.001^*$
Surgical	NLR (per unit)	1.093	$< 0.001^*$
Surgical	Positive LN (per node)	1.052	$< 0.001^*$
Surgical	IB 1 vs. 0	1.224	0.153
Surgical	IB 2 vs. 0	2.148	$< 0.001^*$
Surgical	IB 3 vs. 0	2.452	0.017^*

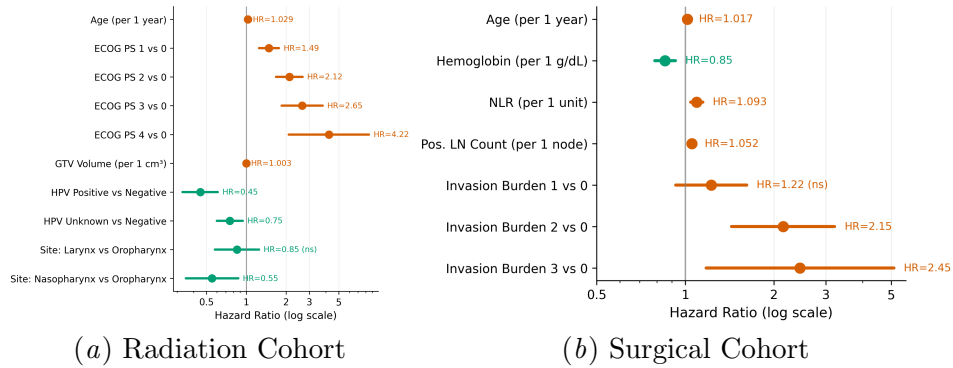


Figure 5: Causal hazard ratio forest plots with 95% CIs on log scale. DAG-informed covariate adjustment.

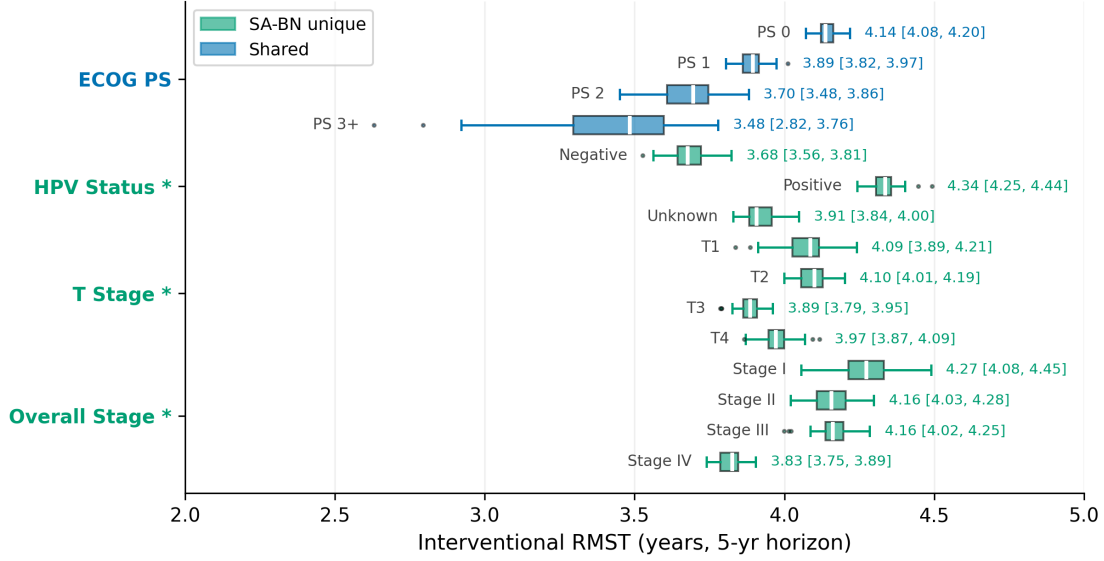


Figure 6: Interventional RMST (5-year horizon) on the radiation cohort by feature level estimated via G-computation. Green = SA-BN unique; blue = shared with binary Markov blanket.

Appendix G. Proportional-Hazards Testing

We assess the proportional hazards (PH) assumption of the DAG-informed Cox models using Schoenfeld residual tests, applied per feature within each cohort’s survival Markov blanket, and corrected for multiple comparisons with the Benjamini–Hochberg (BH) FDR procedure (Table 9). After correction, the only violation is overall stage in the radiation cohort ($q = 7.2 \times 10^{-5}$); no surgical Markov-blanket feature violates PH ($q \geq 0.17$). As a sensitivity analysis, we refit the radiation Cox model stratified on overall stage: all hazard ratios retain direction and magnitude (ECOG PS $1.95 \rightarrow 1.82$, HPV $0.47 \rightarrow 0.43$, site $0.34 \rightarrow 0.42$, T stage $0.69 \rightarrow 0.77$, treatment modality $1.61 \rightarrow 1.76$), so the hazard-ratio and interventional-RMST interpretations are unaffected.

Appendix H. Quantitative Markov-Blanket Membership

For each survival Markov-blanket feature, its 50-bootstrap selection frequency, DAG-informed hazard ratio, BH-FDR-corrected Cox p -value (q), and the bootstrap stability of its directed edge into the survival node are reported in Table 10. Every feature survives multiple-comparison correction at $q < 0.05$ except radiation T stage ($q = 0.42$); in the surgical cohort, all five features survive. Directed-edge-into-survival stabilities from the same 50 bootstraps confirm that the highest-frequency features also anchor the graph: in the radiation cohort, ECOG PS enters survival in 100% of resamples, age in 98%, and site in 96%; in the surgical cohort, hemoglobin in 82% and invasion burden in 74%.

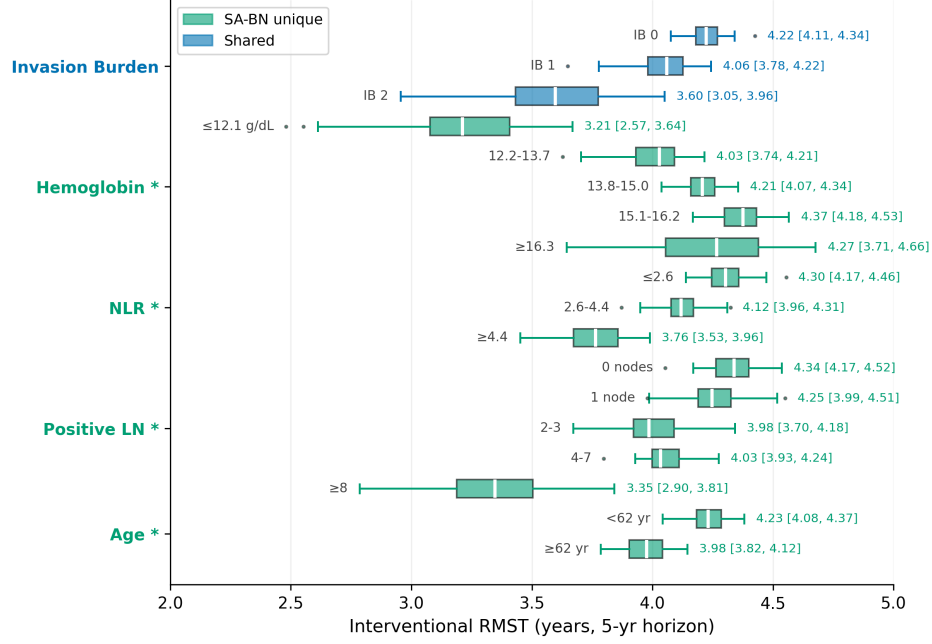


Figure 7: Interventional RMST (5-year horizon) on the surgical cohort by feature level estimated via G-computation.

Table 9: Schoenfeld PH residual tests with BH-FDR within each cohort’s survival Markov blanket. Bold $q < 0.05$ indicates a PH violation.

Cohort	Feature	stat	p_{raw}	q_{BH}
Radiation	overall stage	19.9	8×10^{-6}	7.2×10^{-5}
Radiation	smoking (pack-yr)	5.64	0.018	0.079
Radiation	age	3.67	0.055	0.107
Radiation	site	3.57	0.059	0.107
Radiation	GTV	3.37	0.066	0.107
Radiation	T stage	3.26	0.071	0.107
Radiation	HPV status	1.61	0.205	0.264
Radiation	treatment modality	1.38	0.241	0.271
Radiation	ECOG PS	1.09	0.296	0.296
Surgical	blood hemoglobin	4.50	0.034	0.169
Surgical	invasion burden	3.19	0.074	0.185
Surgical	positive LN count	2.35	0.126	0.209
Surgical	NLR	0.54	0.464	0.580
Surgical	age	0.25	0.616	0.616

Appendix I. Extended Downstream-Model Comparison

We compare four survival models: Cox PH, RSF, gradient-boosted survival (GBS), and LASSO-penalized Cox, holding the model fixed and varying only the feature set (survival-

Table 10: Survival Markov-blanket membership: 50-bootstrap selection frequency, hazard ratio (DAG-informed Cox) and BH-FDR q -value.

Cohort	Feature	Sel. freq.	HR	q_{BH}
Radiation	HPV status	100%	0.43	2.5×10^{-13}
	age	100%	1.03	2.0×10^{-12}
	treatment modality	100%	1.84	6.0×10^{-10}
	GTV	100%	1.01	4.3×10^{-8}
	ECOG PS	100%	1.83	1.3×10^{-7}
	smoking (pack-yr)	100%	1.01	2.3×10^{-7}
	site	100%	0.41	6.9×10^{-6}
	overall stage	68%	2.09	3.3×10^{-4}
Surgical	T stage	66%	0.81	0.42
	blood hemoglobin	94%	0.82	4.8×10^{-4}
	positive LN count	88%	1.06	7.5×10^{-5}
	age	88%	1.03	9.0×10^{-5}
	invasion burden	86%	2.87	4.2×10^{-5}
	NLR	78%	1.08	6.9×10^{-3}

aware vs. binary Markov blanket), across three metrics (Table 11). In the radiation cohort, the survival-aware set wins every model and metric. In the surgical cohort, it wins for the non-linear RSF and on the integrated Brier score across all models, while a linear Cox or LASSO-Cox narrowly favors the binary set on discrimination, an observation consistent with non-linear models transferring better on the smaller surgical cohort.

Table 11: Downstream survival models on the survival-aware vs. binary Markov blanket. Each cell is **SA-MB** / binary-MB; Harrell C and tAUC higher is better, IBS lower is better; bold marks the winning feature set. Values are from the extended multi-model benchmark, whose feature encoding/preprocessing differs slightly from the primary run behind the Table 2 headline. Radiation values and the survival-aware surgical values agree with the primary run to within bootstrap noise; the surgical binary-MB C-index is higher here (RSF 0.678 vs. the primary-run 0.648), so the survival-aware surgical gain under this benchmark (+0.043 for RSF) is smaller than the headline +0.078 while remaining in the same direction.

Cohort	Model	Harrell C	tAUC	IBS↓
Radiation	Cox PH	.792 /.766	.840 /.812	.067 /.070
	RSF	.804 /.770	.848 /.815	.069 /.072
	GBS	.799 /.766	.844 /.811	.067 /.071
	LASSO-Cox	.790 /.757	.837 /.802	.069 /.073
Surgical	Cox PH	.625/ .639	.671/ .695	.093 /.096
	RSF	.721 /.678	.767 /.722	.089 /.096
	GBS	.672 /.663	.700/ .714	.094 /.099
	LASSO-Cox	.636/ .686	.680/ .730	.093 /.095

Appendix J. Parity with Survival-Native Selectors

We compare the survival Markov blanket against size-matched LASSO-Cox and RSF variable-importance feature sets, with a downstream Cox PH model fit on each selected set for a like-for-like comparison (Table 12). A paired-bootstrap C-index difference test ($B = 1000$) with BH-FDR does not distinguish SA-BN from a size-matched LASSO-Cox or RSF-VI set on either cohort (all $q > 0.5$). The only significant difference is a larger cross-validated 12-feature LASSO-Cox model, which exceeds the 6-feature radiation blanket ($q = 0.04$). The advantage is attributable to the larger feature set at a fixed estimator.

Table 12: Feature-selection parity. Test-set C-index (Cox PH on each selected set) and paired-bootstrap BH-FDR q -value versus the SA-BN blanket. “–” = not applicable.

Selection method	k (rad/surg)	Rad. C	Surg. C	q vs. SA-BN (rad/surg)
SA-BN blanket (Cox)	9 / 5	0.791	0.725	–
LASSO-Cox (size-matched)	6 / 4	0.769	0.699	0.64 / 0.57
LASSO-Cox (CV-best)	12 / –	0.795	–	0.04 / –
RSF-VI (top- k)	6 / 5	0.769	0.674	0.64 / 0.57

Appendix K. Computational Cost

We report wall-clock times on a single CPU thread. A single ensemble pass is dominated by the PC constraint-based step at large n (Table 13). The 50-bootstrap stability analysis is a one-time offline step and is not required at deployment. A single pass at $n \approx 100$ (~ 2.5 s) is a fast route for privacy- or size-constrained cohorts, and a single SA-BN pass remains cheaper than RSF permutation importance.

Table 13: Runtime, single CPU thread. Reference selectors (RSF variable importance, LASSO-Cox) shown for comparison.

Operation	Radiation	Surgical
1 ensemble pass	~ 137 s	~ 13 s
single pass, $n \approx 100$	~ 2.5 s	~ 2.5 s
50-bootstrap stability (offline)	~ 114 min	~ 11 min
RSF-VI (reference)	~ 16.6 min	–
LASSO-Cox (reference)	< 1 s	< 1 s