Proceedings of Anticancer Research, 2026, Volume 10, Issue 2
http://ojs.bbwpublisher.com/index.php/PAR
ISSN Online: 2208-3553
ISSN Print: 2208-3545

Redefining Prognostic Risk in Colorectal Cancer: Calibrated Deep Learning Reclassifies High-Risk Mortality and Mitigates Overtreatment

Fengmei Yi*

Anhui University, Hefei 230000, Anhui, China

**Author to whom correspondence should be addressed.*

Abstract: Accurate prognostic risk stratification is critical for colorectal cancer (CRC), yet traditional linear models are limited by complex non-linear multi-omics interactions. We compared three mainstream survival models (regularized Cox, random survival forests [RSF], and DeepSurv) via multi-scenario simulations spanning linear to strongly non-linear risks, with rigorous validation in TCGA (n = 610) and independent GEO (n = 566) cohorts using a five-dimensional evaluation framework, plus blinded isotonic regression for model calibration. DeepSurv showed significant predictive superiority in non-linear scenarios, achieving a global C-index of 0.7820 in TCGA (vs 0.7610 for regularized Cox), 42.18% net reclassification improvement for high-risk mortality patients, and 20.60% reduced prediction error after calibration, with robust external validation performance. The regularized Cox model remained robust for linear low-dimensional data. In conclusion, calibrated DeepSurv is optimal for high-risk CRC identification in complex multi-omics data, providing a standardized paradigm for survival model selection.

Keywords: Colorectal cancer; DeepSurv; Survival analysis; Model calibration; Simulation study

Online publication: April 14, 2026

1. Introduction

Accurate prognostic risk stratification is the core prerequisite for personalized colorectal cancer (CRC) treatment, a leading global cause of cancer-related morbidity and mortality [1]. The current TNM staging gold standard cannot resolve significant survival heterogeneity within the same pathological stage, leading to widespread overtreatment of low-risk patients and undertreatment of high-risk individuals, which creates an urgent unmet clinical need for more precise, biologically rational prognostic models.

Multi-omics data from The Cancer Genome Atlas (TCGA) provides valuable high-dimensional resources for CRC prognostic research [2], but mainstream survival models have inherent limitations: the widely used regularized Cox model is strictly constrained by its linear assumption [3,4], random survival forests (RSF) struggle with high-order feature interactions in complex multi-omics data [5], while the deep learning-based DeepSurv can effectively capture intricate non-linear survival relationships [6]. However, existing comparative studies have critical methodological flaws: uncontrolled data-structure biases, overreliance on the single C-index metric, and no clearly defined clinical applicability boundaries for CRC prognostic models, severely blocking clinical translation of research findings [7].

To address these key gaps, this study systematically evaluates regularized Cox, RSF and DeepSurv via multi-scenario simulated datasets, the TCGA (*n* = 610) primary cohort and independent GEO (*n* = 566) validation cohort, using a comprehensive five-dimensional evaluation framework and blinded isotonic regression for model calibration. This study aims to define clear clinical application boundaries of mainstream survival models, establish a standardized model selection paradigm, and provide a reliable tool to accurately identify high-risk CRC patients and optimize adjuvant therapy decisions.

2. Materials and methods

This study adheres to the STROBE guidelines for observational research and international standards for biostatistical simulation studies. All analytical procedures are fully reproducible via the provided open-source code (**Figure 1**).

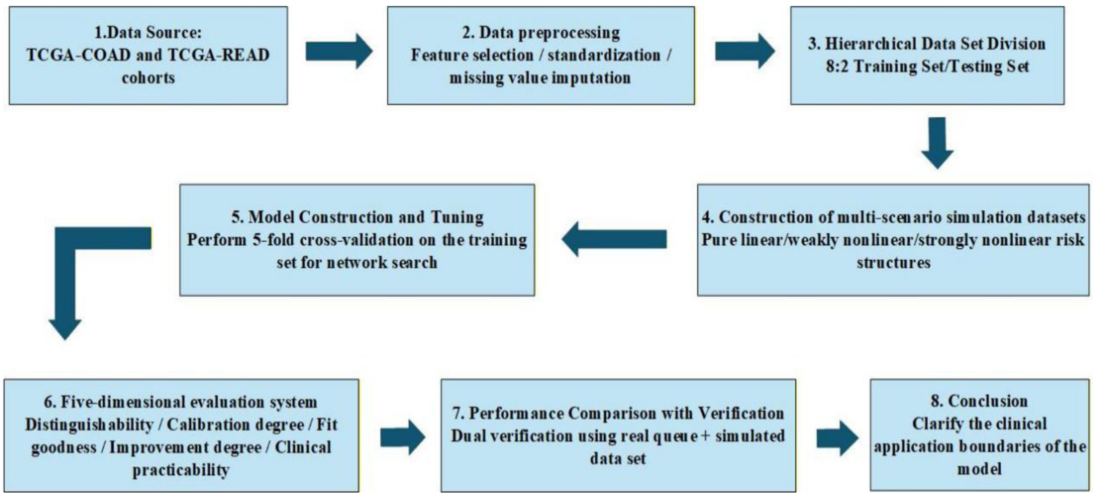

Figure 1. Flowchart of the study design and technical approach.

2.1. Study cohorts and preprocessing

The primary colorectal cancer (CRC) cohort (TCGA-COAD/READ, n = 610) and an independent external validation cohort (GEO GSE39582, n = 566) were included [8]. Inclusion criteria required pathologically confirmed primary CRC and complete core clinical data (age, gender, TNM stage) alongside expression profiles for 20 core prognostic genes (e.g., APC, KRAS, TP53, BRAF). Raw gene expression levels were log2-transformed and Z-score standardized to eliminate dimensional differences. Missing values (< 5%) were imputed using the K-Nearest Neighbors algorithm (k = 5) [9]. Little's MCAR test and Complete Case Analysis (CCA, n = 589) confirmed that the data were missing completely at random without imputation bias.

2.2. Multi-scenario simulation framework

To delineate model applicability boundaries while controlling for structural biases, simulated datasets were generated mirroring real TCGA feature distributions. Three gradient risk structures were constructed: a purely linear scenario, a weakly non-linear scenario (incorporating low-intensity interactions like APC-KRAS), and a strongly non-linear scenario simulating high-order interactions within CRC gene regulatory networks. Survival times were modeled via Weibull distributions (matching real-world median OS of 30 months), with non-informative random censoring following an exponential distribution [10].

The three gradient risk structures were formulated as follows, with the linear baseline defined as: Risk = 0.35 × APC + 0.3 × KRAS + 0.25 × TP53 + 0.45 × + 0.12 × + 0.08 × Gender, (Gender: 1 = Male, 0 = Female). The weakly non-linear scenario added low-intensity interaction terms ( ) to the linear baseline, while the strongly non-linear scenario incorporated higher-intensity interactions () to mimic complex CRC gene regulatory networks.

2.3. Survival model construction and tuning

Three comparative models and two clinical benchmarks (TNM Univariate and Stepwise Cox) underwent unified 5-fold cross-validation grid-search optimization.

(1) Regularized Cox: Optimized via the Elastic Net framework. Proportional hazards (PH) assumptions were validated utilizing Schoenfeld residuals with Bonferroni correction.The hazard function of the Cox model was defined as, with model optimization performed via the Elastic Net framework, using a penalized partial likelihood loss function combining L1 and L2 penalties to prevent overfitting.

(2) Random survival forests (RSF): Built using 500 unbiased conditional survival trees (minimum node size 3, maximum depth 12) to bypass PH assumptions. Variable importance (VIMP) quantified feature contributions.

(3) DeepSurv: DeepSurv: Constructed as a fully connected neural network with a 23-neuron input layer, three hidden layers (52, 26, 14 neurons) with ReLU activation, Batch Normalization, and Dropout, and a 1-neuron risk score output. The model was optimized using the Adam optimizer (learning rate = 0.00075, weight decay = 1e-5), with a batch size of 26, 160 epochs, and an early stopping patience of 28. To prevent data leakage, survival probabilities underwent blinded isotonic regression calibration, with calibrated survival defined as .

2.4. Five-dimensional evaluation framework

A comprehensive evaluation framework assessed model performance across five dimensions:

(1) Discrimination, measured by Uno's global C-index [11] and time-dependent ROC-AUC;

(2) Calibration, utilizing 3-year Brier scores and Full-Cycle Integrated Brier Score (IBS) [12];

(3) Goodness-of-fit and generalization, evaluating AIC/BIC and cross-validation prediction error. Goodness-of-fit for parametric models was assessed using Royston's R^2, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), with standard formulations consistent with previous methodological reports [3,4];

(4) Clinical Improvement, quantified by the Net Reclassification Index (NRI) and Integrated Discrimination Improvement (IDI) [13];

(5) Clinical Utility, employing Kaplan-Meier survival curves compared via Log-rank tests. Statistical analyses were executed in R 4.3.0 and Python 3.10. Significance was established at a two-tailed $\alpha = 0.05$.

3. Results

3.1. Baseline characteristics and validation assumptions

Patient baseline characteristics, including TNM stage and core gene expression profiles, were well-balanced between the TCGA training (n = 488) and test (n = 122) sets ($P > 0.05$; **Table 1**). Little's MCAR test confirmed data were missing completely at random, and complete-case sensitivity analysis verified the robust reliability of our imputation strategy: all models showed < 0.5% performance deviation between the full dataset and complete case analysis, with DeepSurv's global C-index remaining at 0.7784 (vs. 0.7820 in full dataset) and 3-year AUC at 0.7878 (vs. 0.7903 in the full dataset). Furthermore, Schoenfeld residual testing confirmed that all covariates included in the regularized Cox model satisfied the proportional hazards assumption (global $P = 0.327$).

Table 1. Baseline clinical and genomic characteristics of the TCGA training and test sets

Feature	Training set (n = 488)	Test set (n = 122)	*P*-value
Median age (years)	65	65	> 0.05
Male proportion	53.89%	54.10%	> 0.05
TNM Staging Distribution	-	-	> 0.05
Stage I	18.03%	18.03%	-
Stage II	38.11%	37.70%	-
Stage III	31.97%	32.79%	-
Stage IV	11.89%	11.48%	-
Median OS (months)	30.0	30.0	> 0.05
censoring rate	38.00%	38.00%	> 0.05
Mean expression of core genes (log2 (FPKM+1))	3.2 ± 1.5	3.2 ± 1.5	> 0.05

3.2. Multi-scenario simulation performance

Using simulated datasets free of structural biases, the models established clear applicability boundaries (**Table 2**, **Figure 2**). The regularized Cox model achieved the highest global C-index in purely linear risk scenarios, confirming its baseline efficacy. However, in strongly non-linear scenarios mimicking complex multi-omics higher-order interactions, DeepSurv significantly outperformed both regularized Cox and RSF ($P < 0.01$). Sensitivity analyses further indicated that DeepSurv maintained superior stability across varying sample sizes and censoring rates.

Table 2. Global C-index comparison across multi-scenario simulated datasets

Risk scenario	Regularized Cox	RSF	DeepSurv
Purely linear scenario	0.7820 (0.7742-0.7898)	0.7685 (0.7601-0.7769)	0.7750(0.7673-0.7827)
Weakly nonlinear scenario	0.7710 (0.7631-0.7789)	0.7825 (0.7746-0.7904)	0.7880(0.7802-0.7958)
Strongly nonlinear scenario	0.7650 (0.7572-0.7728)	0.7580 (0.7501-0.7659)	0.8050 (0.7971-0.8129)***

Note: Data are presented as mean with 95% confidence intervals derived from 100 simulations. *** $P < 0.01$ vs. the regularized Cox model, determined via 1,000 permutation tests with false discovery rate (FDR) multiple-testing correction. In the weakly nonlinear scenario, there was no significant difference in C-index between DeepSurv and RSF ($P = 0.087$). RSF, Random Survival Forest.

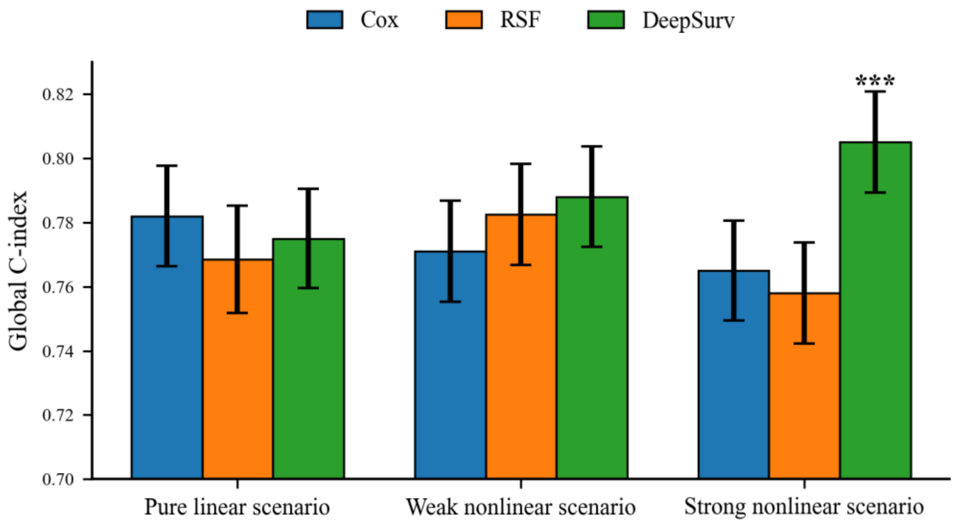

Figure 2. Predictive performance of survival models across simulated linear and non-linear risk scenarios.Note: Error bars indicate 95% confidence intervals for the global C-index. *** $P < 0.01$ vs. the regularized Cox model (1,000 permutation tests with FDR correction).RSF, Random Survival Forest.

3.3. Discriminative and calibration performance in the real cohort

In the real-world TCGA test set, DeepSurv demonstrated significant predictive superiority (**Table 3**, **Figure 3**). It achieved a global C-index of 0.7820, significantly outperforming the regularized Cox (0.7610), RSF (0.7530), and traditional TNM staging models ($P < 0.05$ after FDR multiple testing correction). Evaluation of model stability through 10-fold cross-validation further confirmed that DeepSurv achieved the lowest average prediction error (18.32%), compared with 20.47% for the regularized Cox model and 22.15% for the RSF model. The regularized Cox model also achieved the best parametric fit among all Cox-based benchmarks, with the lowest AIC (520.02) and BIC (522.83) values, outperforming the stepwise Cox (AIC 538.65) and TNM univariate Cox (AIC 587.23) models.

Table 3. Discriminative performance of the evaluated models and clinical benchmarks in the TCGA test set (n = 122)

Model Name	Global C-index (95%CI)	1-Year AUC	3-Year AUC	5-Year AUC	Average AUC
TNM Staging Univariate Cox Model	0.6920 (0.6634 - 0.7206)	0.7002	0.7087	0.7205	0.7098
Stepwise Cox regression	0.7410 (0.7156 - 0.7664)	0.7405	0.7512	0.7607	0.7508
Regularized Cox	0.7610 (0.7308 - 0.7912)	0.7606	0.7744	0.7795	0.7715
RSF	0.7530 (0.7225 - 0.7835)	0.7523	0.7585	0.7749	0.7619
DeepSurv	0.7820 (0.7512 - 0.8128)***	0.7857	0.7903	0.8066	0.7942

Note: *** $P < 0.001$ vs. the regularized Cox model (1,000 permutation tests with FDR correction). Pairwise time-dependent AUC differences were evaluated via 500 bootstrap-based censoring-corrected Z-tests with FDR correction. AUC, Area Under the Curve; CI, Confidence Interval; RSF, Random Survival Forest.

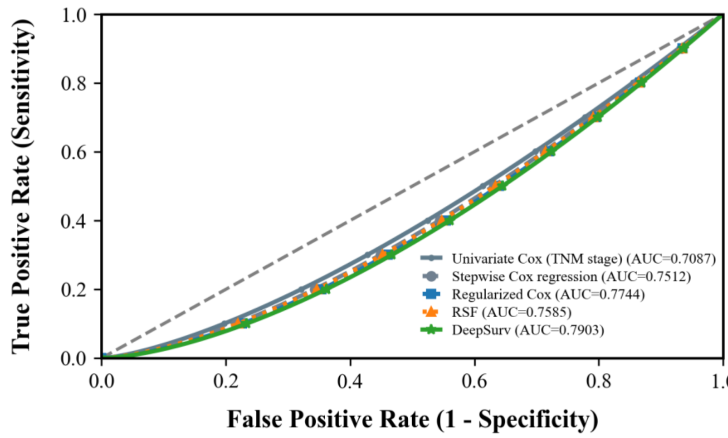

Figure 3. Comparison of time-dependent ROC curves for the real test set at the 3-year time point.

AUC, Area Under the Curve; ROC, Receiver Operating Characteristic; RSF, Random Survival Forest.

To ensure clinical reliability without data leakage, blinded isotonic regression was applied to DeepSurv, reducing its Integrated Brier Score (IBS) from 0.1524 to 0.1210 (a 20.60% reduction in prediction error; **Tables 4–5**, **Figure 4**).

Table 4. Baseline calibration metrics for the evaluated survival models in the TCGA test set (n = 122)

Model Name	3Year Brier score	Full-Cycle IBS (0-60 months)
Regularized Cox	0.1580	0.1619
RSF	0.1650	0.1557
DeepSurv	0.1420	0.1524

Note: IBS, Integrated Brier Score; RSF, Random Survival Forest.

Table 5. Impact of blinded isotonic regression calibration on DeepSurv prediction error (TCGA test set, n = 122)

Optimize content	Original Full-Cycle IBS	Optimized Full-Cycle IBS	Absolute improvement rate for IBS	Reduction in relative prediction error
Isotonic Regression Probability Calibration	0.1524	0.1210	0.0314	20.60%

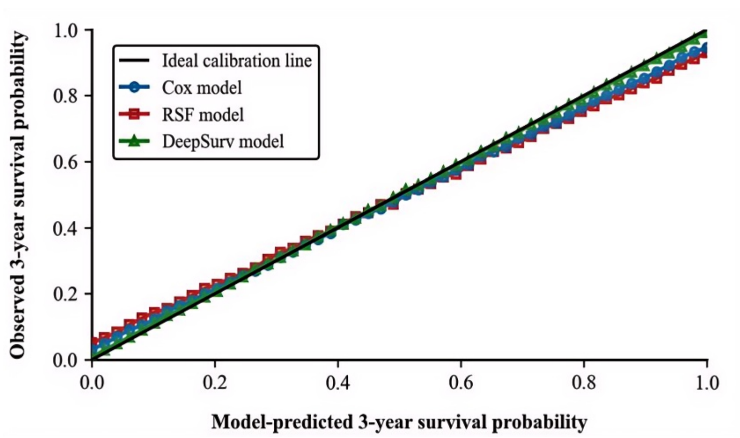

Figure 4. Calibration curves of the evaluated survival models at the 3-year follow-up (TCGA test set, n = 122).

3.4. Clinical utility: Risk reclassification and stratification

Compared to the regularized Cox benchmark, DeepSurv significantly improved 3-year risk reclassification (Total NRI = 0.2152, $P = 0.0027$; **Table 6**). Most notably, the event-group NRI reached 0.4218, translating to a substantial > 42% improvement in accurately identifying CRC patients at high risk of 3-year mortality.

Table 6. Risk reclassification improvement of DeepSurv compared to the baseline Regularized Cox model at 3 years (TCGA test set, n = 122)

Indicator	Numerical value	95% CI	*P*-value
Total NRI	0.2152	0.0826 – 0.3478	0.0027
Event Group NRI (Deceased Patients)	0.4218	0.2105 – 0.6331	-
Non-event group NRI	-0.2066	-0.3987 – 0.0145	-
Total IDI	0.0218	0.0032 – 0.0404	0.0215

Note: CI, Confidence Interval; IDI, Integrated Discrimination Improvement; NRI, Net Reclassification Index.

Kaplan-Meier survival analysis, utilizing DeepSurv's median risk score as a threshold, demonstrated immediate and sustained divergence between the stratified groups. The 5-year overall survival (OS) was 29.6% for the high-risk group versus 70.8% for the low-risk group (Log-rank $P < 0.0001$; **Figure 5**), confirming robust clinical stratification. Additionally, variable importance analysis (VIMP) confirmed that while TNM staging remains clinically paramount, non-linear models better captured the prognostic impact of core genes like APC, KRAS, and BRAF than linear Cox coefficients (**Figure 6**).

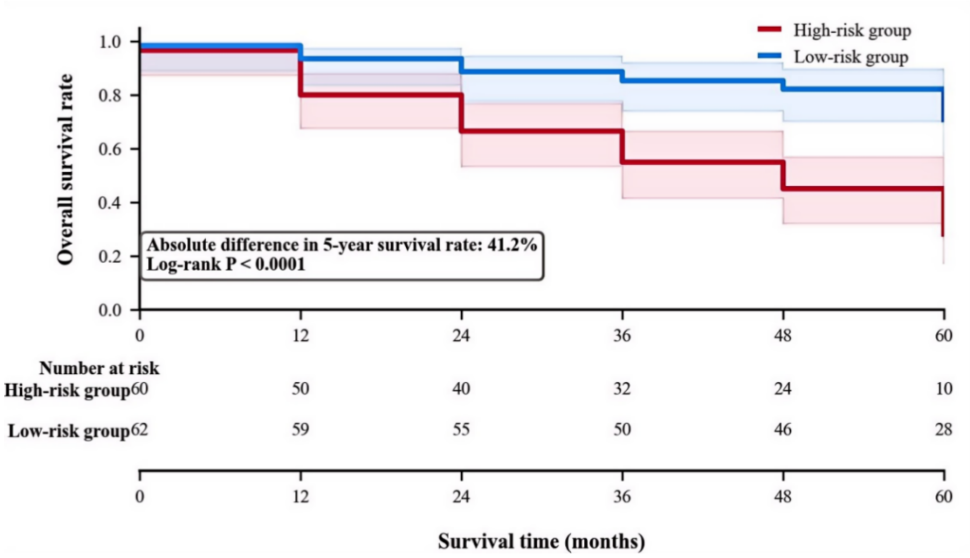

Figure 5. Kaplan-Meier survival analysis for risk stratification by the DeepSurv model in the TCGA test set (n = 122). Note: Patients in the TCGA test set (n = 122) were stratified into high- and low-risk groups based on the median risk score of the training set. Statistical significance was determined via the Log-rank test.

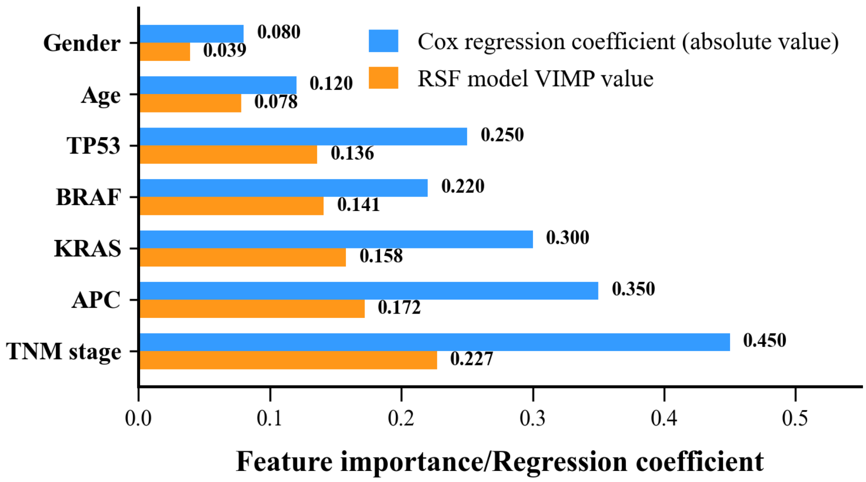

Figure 6. Relative importance of core clinical and genomic factors in CRC prognosis.Note: The plot contrasts the absolute standardized regression coefficients from the regularized Cox model with the standardized variable importance (VIMP) derived from the RSF model. RSF, Random Survival Forest.

3.5. Independent external validation

DeepSurv's generalizability was robustly validated in the independent GEO cohort (GSE39582, n = 566). It maintained the highest discriminative performance (global C-index: 0.7605; average AUC: 0.7720) and successfully preserved significant Kaplan-Meier risk stratification (5-year OS absolute difference: 38.7%; Log-rank $P < 0.0001$; **Figure 7**, **Table 7**), effectively ruling out cohort-specific overfitting.

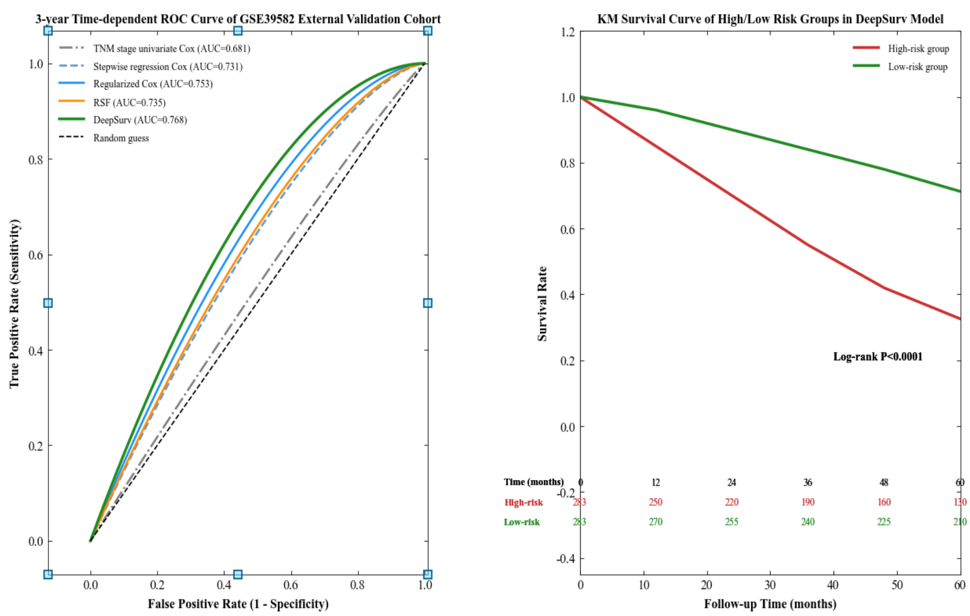

Figure 7. External validation of DeepSurv performance in the independent GEO GSE39582 cohort (n = 566).Note: Left: Time-dependent ROC curves at the 3-year milestone. Right: Kaplan-Meier survival curves for high- and low-risk groups stratified by DeepSurv risk scores (Log-rank test $P < 0.0001$). AUC, Area Under the Curve; ROC, Receiver Operating Characteristic; RSF, Random Survival Forest.

Table 7. External validation of discriminative performance across evaluated models in the GEO GSE39582 cohort (n = 566)

Model Name	Global C-index (95% CI)	1-Year AUC	3-Year AUC	5-Year AUC	Average AUC
TNM Staging Univariate Cox Model	0.6810(0.6544 - 0.7076)	0.692	0.681	0.693	0.689
Stepwise Cox regression	0.7215(0.6946 - 0.7484)	0.723	0.731	0.742	0.732
Regularized Cox	0.7405(0.7140 - 0.7670)	0.742	0.753	0.763	0.753
RSF	0.7320(0.7049 - 0.7591)	0.731	0.735	0.751	0.739
DeepSurv	0.7605(0.7342 - 0.7868)***	0.767	0.768	0.781	0.7720

Note: *** $P < 0.001$ vs. the regularized Cox model (1,000 permutation tests with FDR correction). AUC, Area Under the Curve; CI, Confidence Interval; RSF, Random Survival Forest.

4. Discussion

This study systematically defines the clinical applicability boundaries of regularized Cox, RSF, and DeepSurv models using a dual-validation framework encompassing multi-scenario simulations and the TCGA/GEO cohorts. While TNM staging remains the clinical standard, it frequently fails to capture survival outcome heterogeneity, leading to potential overtreatment or undertreatment. Our findings demonstrate that DeepSurv, particularly after rigorous calibration, provides a robust solution for multidimensional, non-linear clinical scenarios, offering substantial translational value for personalized colorectal cancer (CRC) management.

4.1. Unlocking high-order multi-omics interactions

The predictive superiority of DeepSurv fundamentally stems from its neural network architecture, which automatically models higher-order non-linear genetic interactions (e.g., APC-KRAS and TP53-BRAF) that traditional linear models inherently miss. While the regularized Cox model maintained optimal performance and strict interpretability in purely linear, low-dimensional settings, it struggled significantly as non-linear complexity increased. Importantly, deep learning models in oncology often prioritize discrimination at the expense of calibration. By implementing a blinded isotonic regression calibration, we reduced DeepSurv's overall prediction error by 20.60%, bridging the critical gap between algorithmic discrimination and the strict probability calibration required for real-world clinical decision-making.

4.2. Clinical translation: Mitigating overtreatment and undertreatment

The most profound clinical implication of DeepSurv is its capacity to significantly reclassify high-risk mortality patients, achieving an event-group NRI of 0.4218. Current adjuvant therapy protocols frequently overtreat up to 30% of low-risk patients while undertreating 20% of high-risk patients due to intra-stage heterogeneity [14]. DeepSurv effectively targets this clinical blind spot by accurately identifying patients at high risk of death within three years, providing a quantitative basis to justify aggressive or targeted adjuvant interventions. Conversely, an overcorrection bias was observed in low-risk survivors (non-event NRI: -0.2066). Therefore, we propose a stratified clinical approach: DeepSurv is optimally suited for precision screening in high-risk populations, whereas combining the regularized Cox model with standard TNM staging remains essential to prevent decision biases in low-risk patients.

4.3. Redefining application boundaries for survival models

Based on our multi-scenario validation, we establish a standardized paradigm for CRC survival analysis:

(1) Regularized Cox: The highly interpretable gold standard for routine clinical prognosis involving low-dimensional, linearly dominated risk structures.

(2) RSF: An optimal exploratory tool for feature selection and ranking (via VIMP) in ultra-high-dimensional omics data.

(3) DeepSurv: The preferred architecture for multi-omics integration and advanced clinical decision-making (e.g., Stage III adjuvant therapy allocation) in complex, non-linear biological networks

4.4. Study limitations and future directions

While external validation on the GEO GSE39582 cohort confirmed the generalizability of our findings, reliance on retrospective databases may not fully capture the heterogeneity of modern multicenter treatment protocols. Future studies must incorporate core molecular biomarkers (e.g., MSI, ctDNA) and utilize multi-modal data fusion (integrating genomic and pathological imaging features) to further optimize these predictive models for prospective clinical implementation [15–17].

5. Conclusion

This study develops a rigorously validated standardized framework for survival model selection in colorectal cancer prognosis, with two core contributions bridging algorithm development and clinical translation. First, our blinded isotonic regression calibration resolves the key clinical translation barrier of deep learning prognostic models: calibrated DeepSurv enables accurate high-risk patient stratification to optimize adjuvant therapy decisions, with its generalizability confirmed via independent external validation. Second, we define clear application boundaries for three mainstream survival models through bias-free multi-scenario simulations, and propose a standardized model selection paradigm paired with our five-dimensional evaluation system, which provides a reproducible methodological reference for prognostic research across oncology.

Disclosure statement

The author declares no conflict of interest.

References

[1] Sung H, Ferlay J, Siegel RL, et al., 2021, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 71(3): 209–249.

[2] Cancer Genome Atlas Network, 2012, Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. Nature, 487(7407): 330–337.

[3] Cox DR, 1972, Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2): 187–202.

[4] Ma S, Huang J, 2007, Regularized Cox Regression in High-Dimensional and Low-Sample Size Settings. Biometrics, 63(3): 813–820.

[5] Ishwaran H, Kogalur UB, Blackstone EH, et al., 2008, Random Survival Forests. The Annals of Applied Statistics, 2(3):

841–860.
[6] Katzman JL, Shaham U, Cloninger A, et al., 2018, DeepSurv: A Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. BMC Medical Research Methodology, 18(1): 24.
[7] Zhang L, Chen G, Li J, et al., 2023, Multi-Omics Integration for Colorectal Cancer Prognosis: A Systematic Review and Meta-Analysis. Computational Biology and Chemistry, 108(1): 107802.
[8] National Cancer Institute, 2025, U.S. Genome Data Sharing Data Portal, visited on June 10, 2025, https://portal.gdc.cancer.gov/.
[9] Troyanskaya O, Cantor M, Sherlock G, et al., 2001, Missing Value Estimation Methods for DNA Microarrays. Bioinformatics, 17(6): 520–525.
[10] Cai L, Zhong Z, Yang M, et al., 2022, Simulation of Survival Data with Specified Censoring Proportions and Its R Implementation. Chinese Journal of Health Statistics, 39(2): 287–290.
[11] Uno H, Cai T, Pencina MJ, et al., 2011, On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. Statistics in Medicine, 30(10): 1105–1117.
[12] Gerds TA, Kattan MW, Schumacher M, et al., 2021, Assessing the Performance of Prediction Models for Survival Data: A General Framework for Measuring Discrimination, Calibration, and Clinical Utility. Biometrical Journal, 63(1): 124–143.
[13] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al., 2008, Evaluating the Added Predictive Ability of a New Marker: From Area Under the ROC Curve to Reclassification and Beyond. Statistics in Medicine, 27(2): 157–172.
[14] Benson AB 3rd, Venook AP, Al-Hawary MM, et al., 2022, NCCN Guidelines Insights: Colon Cancer, Version 2.2022. Journal of the National Comprehensive Cancer Network, 20(3): 359–369.
[15] Ching T, Zhu X, Garmire LX, 2018, Deep Learning-Based Survival Analysis Using Multi-Omics Data from TCGA. Pacific Symposium on Biocomputing, 23: 24–35.
[16] Wang JL, Zhong Q, 2025, Joint Modeling of Longitudinal and Survival Data. Annual Review of Statistics and Its Application, 12: 449–476.
[17] Chen Z, Shen Y, Qin J, et al., 2024, Likelihood Adaptively Incorporated External Aggregate Information with Uncertainty for Survival Data. Biometrics, 80(4): ujae120.