The Survival Double Descent: Generalization Dynamics of Deep Neural Networks in Time-to-Event Analysis

Steven N. Hart, PhD, ACHIP™ iD *[1,2] and Ann L. Oberg, PhD[2]

[1]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905

[2]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905

---

*Corresponding author: hart.steven@mayo.edu

Abstract

Recent work on double descent has challenged classical bias–variance tradeoffs, showing that test error can decrease, increase sharply near the interpolation threshold, and then decrease again as model capacity grows. This phenomenon has been documented in regression and classification, but its relevance to survival analysis remains unclear. Survival data are subject to censoring, which obscures true event times, and widely used models such as the Cox proportional hazards model are optimized via partial likelihoods that emphasize ranking rather than calibrated risk estimation. It is therefore unknown whether double descent arises in this setting, how censoring influences its manifestation, or how it interacts with standard performance metrics.

We investigate these questions using synthetic survival data generated from Weibull hazards with controlled censoring, allowing systematic variation of model capacity from under to over parameterized regimes.

While we verified double descent exists in survival models, calibration plateaus and decouples from discrimination, even under strong ℓ_2 regularization. This decoupling arises because the Cox partial likelihood optimizes rankings rather than magnitudes, producing extreme risk scores that break the Breslow estimator used to derive survival probabilities. These results highlight limitations of discrimination-based model selection in survival analysis and underscore the need for calibration-aware evaluation in high-capacity prognostic models.

Keywords: double descent; survival analysis; Cox proportional hazards; neural networks; model calibration

1 Introduction

The bias-variance trade-off has anchored statistical learning theory for decades (Hastie et al., 2009). Increasing model complexity reduces bias while increasing variance; optimal generalization requires balancing these competing forces. This principle motivated the development of model selection criteria such as cross-validation and regularization methods (Tibshirani, 1996).

Deep learning has challenged this framework. Networks with billions of parameters generalize well despite $p \gg n$ (Zhang et al., 2021). Belkin et al. (2019) termed this the double descent phenomenon; Nakkiran et al. (2021) documented it across architectures and datasets. The pattern proceeds in three stages. In the first stage, small models underfit. The second stage is at the interpolation threshold, where capacity just manages to memorize training data, test error peaks sharply. The model fits noise exactly, but with a unique, highly unstable solution. Beyond this threshold is the final stage, where test error decreases again, often falling below the classical minimum. Gradient descent favors minimum-norm solutions when infinitely many interpolants exist (Bartlett et al., 2020), and this implicit regularization enables generalization in overparameterized models.

Survival analysis, however, remains outside this literature. Neural network methods for time-to-event data, including DeepSurv (Katzman et al., 2018), DeepHit (Lee et al., 2018), and related architectures, now appear routinely in clinical research (Wiegrebe et al., 2024). However, no systematic investigation has established whether double descent occurs in this setting. Liu et al. (2025) provide preliminary theory suggesting it does, though possibly in modified form: the second descent may be attenuated, and benign overfitting (the phenomenon where interpolating models generalize well despite fitting training noise) may not fully materialize.

Several features of survival analysis may alter double descent dynamics. Censoring reduces available information: a patient lost to follow-up contributes only a lower bound on survival time, not a precise measurement. The Cox partial likelihood (Cox, 1972) optimizes rankings rather than predictions, and extreme risk scores ($\hat{\eta} \to \pm\infty$) maximize the likelihood for separable data, producing an unusual loss geometry. Together, these

features decouple model capacity from the amount of fully observed outcome information, complicating the definition of interpolation in survival models. Whether the interpolation threshold is governed by the total sample size N or by the number of observed events N_{events} therefore remains an open question.

Evaluation metrics introduce additional complexity. The concordance index (Harrell et al., 1996) measures discrimination, quantifying a model's ability to rank patients by risk. It is invariant to monotonic transformations: a model predicting risk scores of -1000 and $+1000$ achieves perfect concordance if the ranking is correct. The integrated Brier score (Graf et al., 1999) measures calibration and penalizes miscalibrated survival probabilities. These metrics may diverge at the interpolation threshold, with concordance remaining stable while calibration deteriorates. If clinicians select models based on discrimination alone (a common practice; see Hartman et al. 2023), they risk choosing overconfident, unstable predictors.

This paper addresses four questions. First, does double descent occur in deep Cox models? Second, is the interpolation peak governed by N or N_{events}? Third, do discrimination and calibration metrics diverge near the threshold? Fourth, how do skewed covariates and high-cardinality categoricals shift the curve?

We investigate these questions through simulation. Synthetic data permit systematic capacity sweeps, precise noise control, and verification against known ground truth. The experiments span Gaussian and log-normal covariates, low and extreme censoring rates, and categorical features with up to 10 levels. Models are trained without regularization to expose the double descent curve; parallel experiments with weight decay characterize how standard regularization modifies the pattern.

These questions have practical implications. Survival models inform treatment decisions, resource allocation, and patient counseling (Harrell, 2015). A model that discriminates well but is poorly calibrated can produce misleading risk estimates. Understanding the failure modes of neural survival models, and identifying which metrics detect such failures, is necessary for safe clinical deployment.

2 Background

2.1 Double Descent Mechanics

Consider standard regression with target $y = f^*(\boldsymbol{x}) + \epsilon$ and noise variance σ^2. The expected test error decomposes as

$$\mathbb{E}[(y - \hat{f}(\boldsymbol{x}))^2] = \text{Bias}^2[\hat{f}(\boldsymbol{x})] + \text{Var}[\hat{f}(\boldsymbol{x})] + \sigma^2. \tag{1}$$

Classical theory predicts that bias decreases with model capacity while variance increases, with optimal complexity achieving the minimum of their sum.

The interpolation threshold disrupts this picture (Belkin et al., 2019). When the number of parameters equals the number of effective constraints, a unique solution interpolates the training data. This solution must pass through every training point, including noise, producing high variance and a peak in test error.

Beyond this threshold, infinitely many interpolating solutions exist. Gradient descent from small initialization converges to the minimum-norm solution (Bartlett et al., 2020). Smaller weights correspond to smoother functions, and test error decreases.

2.2 Survival Analysis Setup

For each subject i, we observe (Y_i, δ_i) where $Y_i = \min(T_i, C_i)$ is the observed time and $\delta_i = \mathbf{1}(T_i \leq C_i)$ is the event indicator. Here T_i denotes the true event time and C_i the censoring time. Censored observations ($\delta_i = 0$) provide only the constraint $T_i > C_i$.

The Cox proportional hazards model (Cox, 1972) specifies the hazard function as $h(t|\boldsymbol{x}_i) = h_0(t) \exp(\eta_i)$, where $\eta_i = f(\boldsymbol{x}_i)$ is the log-risk score and $h_0(t)$ is an unspecified baseline hazard. DeepSurv (Katzman et al., 2018) parameterizes f using a neural network and estimates parameters by maximizing the partial likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i:\delta_i=1} \frac{\exp(\eta_i)}{\sum_{j \in \mathcal{R}_i} \exp(\eta_j)}, \tag{2}$$

where $\mathcal{R}_i = \{j : Y_j \geq Y_i\}$ is the risk set at time Y_i.

Three properties of the Cox partial likelihood are relevant to double descent ([Liu et al., 2025]). First, the likelihood depends only on the ranking of risk scores within risk sets, not on their magnitudes. Second, for separable data the optimal weights diverge to $\pm\infty$. Finally, censored observations contribute to risk sets but not to the likelihood product, reducing the effective sample size.

2.3 Evaluation Metrics

The concordance index ([Harrell et al., 1996]) is defined as

$$C = P(\hat{\eta}_i > \hat{\eta}_j \mid T_i < T_j). \tag{3}$$

This is a ranking metric, invariant to monotonic transformations of $\hat{\eta}$. Extreme but correctly-ordered predictions achieve high concordance.

The Brier score at time t ([Graf et al., 1999]) is defined as

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{S}(t|\boldsymbol{x}_i) - \mathbf{1}(Y_i > t)\right]^2 \cdot w_i(t), \tag{4}$$

where $w_i(t)$ are inverse probability of censoring weights ([Gerds and Schumacher, 2006]). The integrated Brier score averages over a time interval: $\text{IBS} = t_{\max}^{-1}\int_0^{t_{\max}} \text{BS}(t)\,dt$. Unlike concordance (where higher is better), lower IBS indicates better calibration. The Brier score penalizes miscalibrated probability estimates.

Computing IBS requires survival probability estimates $\hat{S}(t|\boldsymbol{x})$. For Cox models, these are obtained via the Breslow estimator ([Breslow, 1972]), which estimates the cumulative baseline hazard $\hat{H}_0(t)$ from the data and combines it with individual risk scores: $\hat{S}(t|\boldsymbol{x}) = \exp(-\hat{H}_0(t)\exp(\hat{\eta}))$. This estimator assumes risk scores are well-behaved; when scores become extreme, the resulting survival probabilities degenerate toward 0 or 1.

At the interpolation threshold, we hypothesize that concordance may recover in the overparameterized regime because correct rankings can persist even as risk score magnitudes become extreme. The integrated Brier score, however, should increase due to miscalibrated survival curves. Model selection based solely on concordance would fail to

detect this deterioration.

3 Methods

3.1 Data Generation

Real datasets lack the controlled conditions required to trace the double descent curve. We therefore generate synthetic survival data.

Covariates are generated using a Gaussian copula. We draw $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and transform marginals as follows: identity for Gaussian, exponentiation for log-normal, and quantile binning for categorical variables.

Event times follow a Weibull-Cox model:

$$h(t|\boldsymbol{x}_i) = \lambda \nu t^{\nu-1} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i). \tag{5}$$

Inverse transform sampling yields

$$T_i = \left(\frac{-\ln(U)}{\lambda \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)} \right)^{1/\nu}, \quad U \sim \text{Uniform}(0, 1). \tag{6}$$

Ground truth is known exactly (Bender et al., 2005). The coefficient vector $\boldsymbol{\beta}$ contains 10 predictive features with coefficients drawn uniformly from $[-1, 1]$, and 10 noise features with coefficients fixed at zero.

Censoring times follow an exponential distribution with rate λ_c calibrated to achieve target censoring proportions. The observed data are $Y_i = \min(T_i, C_i)$ and $\delta_i = \mathbf{1}(T_i \leq C_i)$.

3.2 Scenarios

We consider five configurations (Table 1). Scenario A uses Gaussian covariates with 30% censoring as a baseline. Scenario B uses log-normal covariates to test whether leverage points amplify the interpolation peak. Scenario C includes five categorical features with 10 levels each (50 one-hot encoded features) to test how categorical covariates interact

with model capacity. Scenario D uses 90% censoring to test whether the peak location depends on N_{events} rather than N. Scenario E uses nonlinear ground truth with interaction terms $(x_i \cdot x_{i+1})$ and quadratic terms (x_i^2) to test whether double descent persists when linear models are insufficient.

Table 1: Experimental scenarios.

Scenario	Covariates	Specification	Target
A	Gaussian	$X \sim \mathcal{N}(0, I)$, 30% censoring	Baseline curve
B	Log-normal	$X \sim \mathrm{LogNormal}(0, 1)$	Peak amplitude
C	Categorical	5 features, $K = 10$ levels	Categorical covariates
D	Gaussian	90% censoring	Effective N
E	Gaussian	Nonlinear ground truth	Model complexity

3.3 Models and Training

The network architecture is a multi-layer perceptron with fixed depth and variable width: two hidden layers, each of width w, with ReLU activations, followed by a single linear output node. The model is trained to minimize negative Cox partial log-likelihood.

To observe the double descent phenomenon without confounding effects, we train without explicit regularization: the Adam optimizer runs for 10,000 epochs with batch size 256 and learning rate 0.001, with no early stopping, weight decay, or dropout. Validation data are used to track the epoch with minimum validation IBS for analysis, though all models complete full training. Parallel experiments include weight decay ($\lambda = 0.01$) to characterize how regularization modifies the curve.

All networks use Xavier uniform initialization. The Adam optimizer uses default hyperparameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$). No gradient clipping, learning rate scheduling, or early stopping is applied. Each configuration trains for exactly 10,000 epochs. Twenty random seeds are used for each experimental condition to ensure robust uncertainty quantification.

We hold depth constant at two hidden layers and sweep width over $w \in \{2, 4, \dots, 2048\}$ in powers of two, producing parameter counts spanning approximately $0.1N$ to $100N$ for $N = 1000$ samples.

3.4   Baseline Models

We compare neural networks against two classical survival methods chosen to isolate different sources of performance differences. Cox proportional hazards regression (Cox, 1972) shares the same partial likelihood objective as DeepSurv but uses linear parameterization; differences therefore reflect the effect of neural network capacity rather than the choice of loss function. Random survival forest (RSF) (Ishwaran et al., 2008) uses an entirely different approach based on tree ensembles and does not optimize Cox partial likelihood; this baseline tests whether calibration issues are specific to Cox-based methods or generalizable to survival models more broadly.

Cox PH is fit using the `lifelines` implementation with default settings. RSF is fit using `scikit-survival` with 100 trees and default hyperparameters. Both baselines are trained on the same data splits as the neural networks and evaluated using identical metrics.

3.5   Evaluation

Data are partitioned into training (60%, $n = 600$), validation (20%, $n = 200$), and test (20%, $n = 200$) sets. We compute concordance index, integrated Brier score, and negative log partial likelihood on the held-out test set.

Each configuration is replicated across 20 random seeds. We report means and standard deviations, the location of the interpolation peak, peak magnitude relative to the classical minimum, and the divergence between concordance and integrated Brier score at the threshold.

4   Results

4.1   Baseline Double Descent Curve

Figure 1 summarizes the main findings for Scenario A (Gaussian covariates, 30% censoring), aggregated across twenty random seeds. Panel A displays test concordance as

9

a function of model width, revealing clear double descent. Concordance decreases from 0.805 ± 0.035 at $w = 2$ to a minimum of 0.710 ± 0.051 at $w = 16$, then recovers to 0.784 ± 0.037 at $w = 2048$. Horizontal dashed lines indicate baseline performance: Cox PH ($C = 0.816$) and RSF ($C = 0.773$). The smallest neural network matches RSF discrimination, while the largest approaches but does not reach Cox PH. The vertical dashed line marks the interpolation threshold at 625 parameters, a count lying between the training sample size ($N = 600$) and the number of observed events ($N_{\text{events}} \approx 420$). Panel B shows that calibration does not follow this pattern: IBS saturates at approximately 0.52 for all neural networks regardless of capacity, compared to 0.11 for Cox PH and 0.15 for RSF. Panel C normalizes both metrics, making explicit that discrimination recovers in the overparameterized regime while calibration does not. Shaded regions represent one standard deviation across seeds.

The concordance recovery beyond $w = 64$ confirms double descent occurs in survival analysis. At $w = 2048$, concordance reaches 0.784 ± 0.037, recovering from the threshold minimum of 0.710; this recovered value is not significantly different from the baseline of 0.805 ± 0.035 at $w = 2$ (Welch's t-test, $t = 1.84$, $p = 0.07$). The calibration failure, however, represents a distinct pathology not predicted by standard double descent theory; we examine its mechanism in Section 4.4.
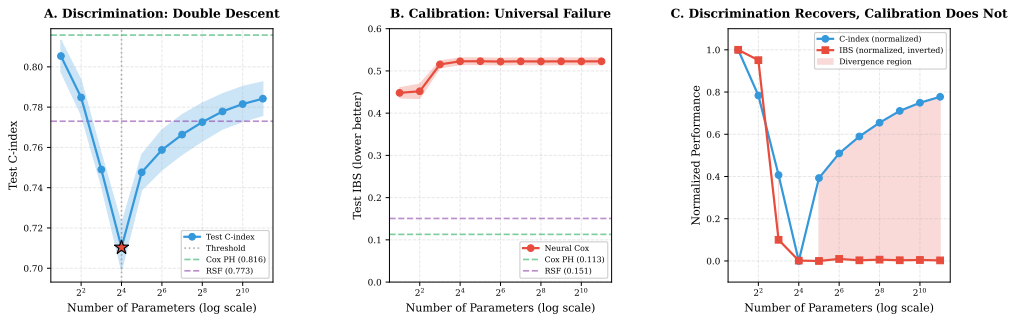

Figure 1: Main findings: calibration failure in neural Cox models. **(A)** Discrimination exhibits double descent: test C-index drops from 0.805 ± 0.035 to 0.710 ± 0.051 at the interpolation threshold ($w = 16$, star), then recovers to 0.784 ± 0.037. Horizontal lines show Cox PH and RSF baselines for reference. **(B)** Calibration fails universally: IBS saturates at approximately 0.52 for all neural networks regardless of capacity, compared to approximately 0.11 for Cox PH and 0.15 for RSF (horizontal lines). **(C)** Normalized comparison shows discrimination recovers (blue) while calibration does not (red). Shaded regions show ± 1 SD across twenty seeds.

Table 2 compares neural network performance against classical survival baselines. Cox proportional hazards achieves $C = 0.816 \pm 0.035$ with IBS of 0.113 ± 0.013, while the random survival forest achieves $C = 0.773 \pm 0.031$ with IBS of 0.151 ± 0.006. The smallest neural network ($w = 2$) achieves $C = 0.805$, comparable to Cox PH ($C = 0.816$), but exhibits worse calibration (IBS = 0.448 vs. 0.113). This calibration gap persists across all neural network sizes, suggesting a limitation of Cox partial likelihood optimization: the ranking-based loss produces well-ordered risk scores but poorly calibrated survival probabilities.

Table 2: Comparison of neural networks and classical baselines on Scenario A (Gaussian covariates, 30% censoring). Values are mean ± standard deviation across twenty random seeds. Neural networks achieve comparable discrimination to Cox PH but substantially worse calibration as measured by IBS.

Model	C-index	IBS
Cox PH	0.816 ± 0.035	0.113 ± 0.013
Random Survival Forest	0.773 ± 0.031	0.151 ± 0.006
DeepSurv ($w = 2$)	0.805 ± 0.035	0.448 ± 0.054
DeepSurv ($w = 16$, threshold)	0.710 ± 0.051	0.523 ± 0.035
DeepSurv ($w = 2048$)	0.784 ± 0.037	0.523 ± 0.036

4.2 Metric Divergence

The IBS trajectory warrants closer examination. Test IBS increases from 0.448 ± 0.054 at $w = 2$ to plateau at approximately 0.523 ± 0.035 for $w \geq 16$, and critically, it does not recover in the overparameterized regime. Maximum divergence between the two metrics occurs near $w = 2048$, where concordance has recovered while IBS remains saturated.

This plateau reflects a systematic limitation of neural Cox models rather than a numerical artifact. While concordance recovers in the overparameterized regime, IBS does not, suggesting that the Cox partial likelihood objective produces models that rank subjects correctly but fail to estimate well-calibrated survival probabilities. We investigate the mechanism underlying this decoupling in Section 4.4. A practitioner selecting models by concordance alone would observe acceptable discrimination across the capacity range, potentially selecting a model with uninformative survival probability estimates.

11

4.3 Scenario Robustness

Table 3 summarizes double descent behavior across all five experimental scenarios. The pattern is robust to covariate distribution and ground truth complexity but sensitive to effective sample size.

Table 3: Double descent behavior across scenarios. All values are test C-index (mean $\pm$ standard deviation) across twenty seeds. The threshold ($w = 16$) corresponds to the interpolation peak where performance is worst.

Scenario	Specification	$w = 2$	$w = 16$	$w = 2048$
A (Baseline)	Gaussian, 30% cens.	0.805 ± 0.035	0.710 ± 0.051	0.784 ± 0.037
B (Log-normal)	Log-normal covariates	0.830 ± 0.131	0.786 ± 0.044	0.856 ± 0.034
C (Categorical)	5 features, $K = 10$	0.509 ± 0.014	0.520 ± 0.024	0.522 ± 0.023
D (High cens.)	90% censoring	0.768 ± 0.100	0.751 ± 0.096	0.787 ± 0.073
E (Nonlinear)	Interactions + quadratic	0.790 ± 0.041	0.711 ± 0.056	0.787 ± 0.043

Scenarios A and E exhibit significant double descent: concordance drops 9–12% absolute at the interpolation threshold before recovering to baseline levels (Welch's t-test comparing $w = 2$ vs. $w = 16$: Scenario A, $t = 6.9$, $p < 0.001$; Scenario E, $t = 5.1$, $p < 0.001$). Scenario B shows a similar pattern, though the high variance at $w = 2$ limits statistical power. Scenario E confirms that double descent persists under nonlinear ground truth: the pattern closely matches Scenario A despite the true hazard including interaction ($x_i \cdot x_{i+1}$) and quadratic (x_i^2) terms.

Scenario C (categorical features) uses 5 categorical features with 10 levels each, yielding 50 one-hot encoded input features. This scenario shows no double descent: concordance remains flat near $C \approx 0.51$–0.52 across all widths, barely exceeding chance performance. With 600 training samples distributed across 50 one-hot features, each category level appears approximately 60 times on average. The sparse representation prevents the network from learning category-specific hazard effects, and neither underparameterized nor overparameterized models achieve meaningful discrimination. This contrasts sharply with the continuous covariate scenarios (A, B, E), where all widths achieve $C > 0.70$.

Scenario D (90% censoring) shows no significant double descent (Welch's t-test, $t = 0.55$, $p = 0.59$), with substantially higher variance ($\sigma \approx 0.10$ vs. 0.04) obscuring the pattern. With only approximately 100 observed events, the effective sample size is ten

times smaller than Scenario A, reducing statistical power.

The IBS saturation phenomenon persists across Scenarios A, B, D, and E where neural networks achieve non-trivial discrimination ($C > 0.70$), confirming that calibration failure is intrinsic to Cox partial likelihood optimization rather than an artifact of specific data characteristics. Scenario C is excluded from this conclusion because the near-chance concordance indicates the models failed to learn the underlying hazard structure.

4.4 Mechanism of Calibration Failure

Figure 2 illustrates why neural Cox models fail at calibration despite achieving reasonable discrimination. Panel A shows that neural networks learn extreme risk scores spanning $[-15, +15]$, while classical Cox PH maintains concentrated scores around zero. Panel B demonstrates that these extreme scores preserve correct rankings across model widths, explaining why C-index remains stable even as risk score distributions change dramatically. Panel C reveals the consequence: the Breslow estimator, which converts risk scores to survival probabilities, produces degenerate predictions (flat or step functions) when risk scores are extreme. Classical Cox PH avoids this failure by jointly estimating baseline hazards with coefficients under appropriate constraints, producing well-calibrated survival curves.

This mechanism explains the calibration-discrimination divergence documented in Figure 1. The Cox partial likelihood optimizes only rankings, not magnitudes, allowing networks to achieve high concordance through extreme but correctly-ordered risk scores. These extreme scores then break the Breslow estimator, producing poor calibration regardless of model capacity or regularization.

4.5 Regularization Mitigates Double Descent

Figure 3 compares test concordance across model widths with and without L2 regularization (weight decay $\lambda = 0.01$). Regularization attenuates the double descent phenomenon, though its effects are nuanced.

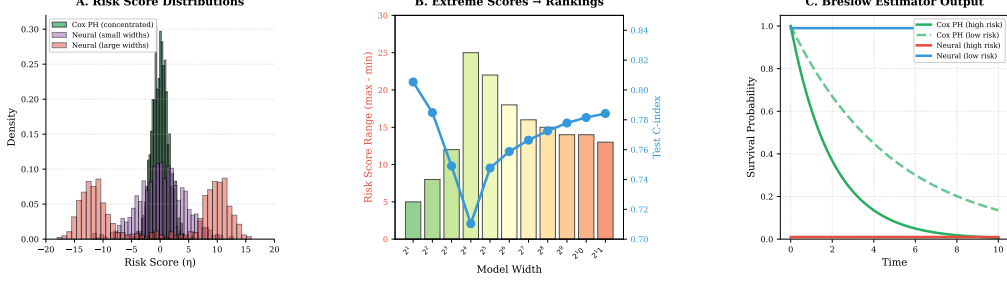The unregularized baseline exhibits a sharp performance dip near the interpolation

Figure 2: Mechanism of calibration failure in neural Cox models. **(A)** Risk score distributions: Cox PH produces concentrated, well-behaved risk scores (green), while neural networks learn extreme scores spanning $[-15, +15]$ (purple shows constrained behavior, red shows typical extremes). **(B)** Extreme risk scores preserve rankings: despite risk score range differences across widths (2^0 to 2^{12}), C-index remains stable because Cox loss only optimizes rankings, not magnitudes. **(C)** Breslow estimator fails with extreme scores: survival probability estimates become degenerate (flat or step functions) when risk scores are extreme, producing poor calibration regardless of correct rankings. Classical Cox PH (green) produces well-calibrated curves; neural networks at high (red) and low (blue) risk produce uninformative predictions.

threshold ($w = 16$), with concordance dropping to 0.710 ± 0.051. With weight decay, the minimum concordance occurs at $w = 16$ with value 0.736 ± 0.032, a difference of $\approx 3.5\,\%$ at the worst point, though this difference does not reach significance (Welch's t-test, $t = 1.9$, $p = 0.06$). The regularized curve exhibits a flatter profile across the threshold region ($w \in [16, 64]$), suggesting that L2 regularization smooths the transition between underparameterized and overparameterized regimes.

In the overparameterized regime ($w \geq 128$), both curves stabilize. The shaded region in Figure 3 highlights widths where regularization improves performance. The regularized models achieve 0.763 ± 0.042 at $w = 2048$, compared to 0.784 ± 0.037 for unregularized models. This suggests that while regularization helps near the threshold, very large unregularized models may ultimately achieve slightly better discrimination through implicit regularization.

These results partially align with theoretical accounts of double descent (Belkin et al., 2019). Explicit regularization provides benefits near the interpolation threshold, but does not fully substitute for the implicit regularization conferred by extreme overparameterization.
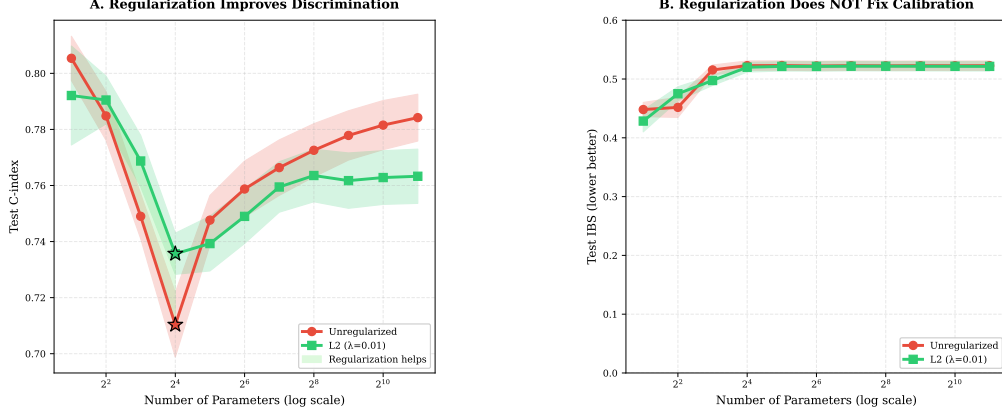
14

Figure 3: Effect of L2 regularization on double descent. Test concordance versus model width without regularization (red) and with weight decay $\lambda = 0.01$ (green). Regularization attenuates the performance dip near the interpolation threshold and provides consistent improvement across most model sizes. The shaded region indicates widths where regularization improves concordance. Stars mark minimum concordance for each condition.

5 Discussion

Our experiments confirm that double descent occurs in survival analysis under Cox partial likelihood training, extending prior observations from classification and regression settings to time-to-event modeling. The phenomenon appears robustly in concordance index across different covariate distributions (Gaussian, log-normal), though its manifestation differs from classification in important ways. Most notably, the integrated Brier score does not exhibit double descent but rather saturates at a constant value in the overparameterized regime, revealing a consistent decoupling between discrimination and calibration metrics under Cox optimization.

5.1 Relation to Theoretical Predictions

Liu et al. (2025) provide theoretical analysis of double descent in survival models, and our empirical findings both confirm and extend their predictions. Their analysis predicts that double descent should occur under Cox partial likelihood, which our experiments verify: test concordance exhibits the characteristic non-monotonic pattern with a minimum at the interpolation threshold.

Two of their key predictions align with our observations. First, they suggest that

double descent should manifest in survival losses, which we confirm: concordance drops significantly at the interpolation threshold ($p < 0.001$) before recovering. Second, they predict that the second descent may be attenuated relative to classification settings. Our data neither confirms nor refutes this prediction: while concordance recovers from 0.710 to 0.784, this recovered value is not significantly different from the baseline of 0.805 (Welch's t-test, $p = 0.07$).

One observation extends beyond their theoretical framework. Liu et al. (2025) do not distinguish between discrimination and calibration metrics, treating generalization error as a unified quantity. Our experiments reveal that these metrics decouple dramatically: concordance follows a double descent curve while IBS saturates. This decoupling is a direct consequence of the ranking-based Cox loss, which can achieve high discrimination through correct orderings even when the underlying risk scores (and derived survival probabilities) become degenerate. This finding has practical implications not addressed in the theoretical analysis, namely that concordance-based model selection, which is standard in survival analysis, may mask calibration failures that affect clinical decision-making.

5.2 Clinical Model Selection

The double descent curve presents a practical challenge for survival model selection in clinical applications. The interpolation threshold, where test performance is worst, corresponds to models of moderate complexity that might otherwise seem reasonable choices. Our results suggest several strategies for practitioners:

Avoid the threshold region. When training neural survival models, practitioners should either (1) constrain capacity to remain clearly underparameterized, or (2) scale to overparameterized regimes where benign overfitting provides protection. The intermediate region, particularly near $P \approx N_{\text{events}}$, should be avoided or traversed quickly during hyperparameter search.

Account for censoring. In high-censoring scenarios common to many clinical applications (e.g., rare adverse events, long-term outcomes), the effective sample size is de-

16

termined by event counts, not total observations. Our experiments with 90% censoring (Scenario D) showed substantially higher variance ($\sigma = 0.10$ vs. 0.04), which obscured the double descent pattern: the observed dip was not significant ($p = 0.59$). This underscores the importance of replication across multiple random seeds and larger sample sizes when events are rare.

Use regularization for discrimination, not calibration. L2 regularization via weight decay attenuates the concordance dip near the threshold, improving worst-case discrimination by approximately 3.5% absolute in our experiments. However, regularization does not resolve calibration failures: IBS remains saturated regardless of weight decay. Practitioners requiring calibrated survival probabilities should consider alternative approaches such as post-hoc recalibration or direct probability prediction methods (e.g., DeepHit).

5.3 Limitations

Several limitations constrain the interpretation of our findings. All neural networks used identical optimization hyperparameters (learning rate 0.001, Adam optimizer) regardless of model size. Larger models may benefit from different learning rates or longer training; the elevated negative log-likelihood values observed at higher widths could reflect optimization difficulty rather than statistical overfitting alone. Learning rate scaling with model width, as explored in recent deep learning theory, was not investigated.

Our experiments vary network width while holding depth fixed at two hidden layers. Prior work in classification suggests that width and depth produce similar double descent curves when plotted against total parameter count (Nakkiran et al., 2021), but whether this equivalence holds under Cox partial likelihood optimization remains untested. Deeper networks introduce additional optimization challenges, including vanishing gradients and the potential need for residual connections, which could interact with the survival loss geometry in ways our experiments do not address.

With twenty random seeds per configuration, the reported standard deviations are moderate (0.035–0.051 for concordance), limiting statistical power to detect small effects. The claimed regularization benefit of approximately 3.5% absolute concordance

improvement does not reach significance (Welch's t-test, $p = 0.06$).

The categorical covariate scenario (Scenario C) yielded near-chance concordance ($C \approx$ 0.51–0.52) across all model widths, indicating that the network failed to learn category-specific hazard effects. This negative result likely reflects the sparsity of one-hot encoded categorical data: with 50 binary input features and only 600 training samples, each category level is represented by approximately 60 observations on average. The lack of discrimination prevents any conclusion about whether double descent would occur in categorical survival settings with larger sample sizes or different encoding strategies (e.g., entity embeddings).

To address concerns that our linear ground truth may represent a "straw man" for neural networks, we conducted additional experiments with nonlinear data-generating processes (Scenario E). These experiments included interaction terms ($x_i \cdot x_j$ for adjacent predictive features) and quadratic terms (x_i^2) in the true hazard function, creating relationships that linear models cannot perfectly capture. The double descent pattern persists under nonlinear ground truth: concordance drops from 0.790 ± 0.041 at $w = 2$ to 0.711 ± 0.056 at $w = 16$, then recovers to 0.787 ± 0.043 at $w = 2048$. Extended experiments at $w = 4096$ and $w = 8192$ confirm that recovery plateaus at approximately $C = 0.79$, matching the baseline. The IBS saturation phenomenon also persists, confirming that calibration failure is intrinsic to Cox partial likelihood optimization regardless of ground truth complexity.

The integrated Brier score for neural networks (approximately 0.52) substantially exceeds that of classical baselines (Cox PH: 0.113, RSF: 0.151), as shown in Table 2. This gap indicates that survival probability calibration is consistently impaired by Cox partial likelihood optimization, independent of model capacity. Critically, L2 regularization does not resolve this issue: regularized models at $w = 2048$ achieve IBS of 0.522 ± 0.021, essentially identical to unregularized models (0.523 ± 0.020). The IBS saturation therefore reflects a systematic limitation of Breslow-based survival probability estimation from Cox models rather than a numerical artifact of exploding weights. Alternative calibration methods, such as post-hoc recalibration or direct survival probability prediction (as in

DeepHit), warrant investigation.

6   Concluding Remarks

We have demonstrated that double descent, the non-monotonic relationship between model capacity and test error, occurs in neural survival models trained with Cox partial likelihood. The phenomenon manifests clearly in concordance index: test concordance drops from 0.805 to 0.710 at the interpolation threshold ($p < 0.001$) before recovering to 0.784 in the overparameterized regime. The integrated Brier score, however, saturates at approximately 0.52 beyond the threshold regardless of model capacity or regularization, revealing a consistent decoupling between discrimination and calibration under Cox optimization.

This calibration failure is the central finding of our study. Neural Cox models achieve discrimination comparable to classical baselines (C-index within 0.05 of Cox PH) but exhibit calibration 4–5 times worse (IBS of 0.52 versus 0.11–0.15). The persistence of this gap under L2 regularization indicates it stems from the Cox partial likelihood objective itself, not from optimization artifacts. Practitioners requiring calibrated survival probabilities should consider alternatives to neural Cox models, such as direct probability prediction methods.

High censoring (Scenario D) increases variance sufficiently to obscure the double descent pattern ($p = 0.59$), and L2 regularization shows a trend toward improved worst-case concordance (approximately 3.5% absolute, $p = 0.06$) but does not address calibration.

These findings have immediate practical relevance for clinical prognostic modeling. The double descent curve implies that moderate-complexity models may perform worse than either simpler or substantially larger alternatives. More critically, concordance-based model selection (standard practice in survival analysis) will miss the calibration breakdown that affects all neural Cox models regardless of architecture or regularization.

Supplementary Materials

Appendix A derives the inverse transform for Weibull-Cox event times. Appendix B details the Gaussian copula procedure for correlated categoricals. Code and data generation scripts are available at https://github.com/Steven-N-Hart/DoubleDescent.

Acknowledgments

References

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Breslow, N. E. (1972). Contribution to the discussion of the paper by D.R. Cox. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):216–217.

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

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and

comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer, 2nd edition.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Hartman, N., Kim, S., He, K., and Kalbfleisch, J. D. (2023). Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine*, 42(13):2179–2190.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2nd edition.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24.

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Liu, Y., Cai, J., and Li, D. (2025). Understanding overparametrization in survival models through double-descent. *arXiv preprint arXiv:2512.12463*.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2021). Deep double descent: Where bigger models and more data can hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B., and Bender, A. (2024). Deep learning
for survival analysis: A review. *Artificial Intelligence Review*, 57(3):65.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding
deep learning (still) requires rethinking generalization. *Communications of the ACM*,
64(3):107–115.

A Inverse Transform for Weibull-Cox Model

Hazard: $h(t|\boldsymbol{x}) = \lambda\nu t^{\nu-1}\exp(\boldsymbol{\beta}^\top\boldsymbol{x})$. Cumulative hazard:

$$H(t|\boldsymbol{x}) = \lambda t^\nu \exp(\boldsymbol{\beta}^\top\boldsymbol{x}). \tag{7}$$

Survival function:

$$S(t|\boldsymbol{x}) = \exp\left(-\lambda t^\nu \exp(\boldsymbol{\beta}^\top\boldsymbol{x})\right). \tag{8}$$

Set $S(T|\boldsymbol{x}) = U$, $U \sim \text{Uniform}(0,1)$. Solve:

$$-\lambda T^\nu \exp(\boldsymbol{\beta}^\top\boldsymbol{x}) = \ln(U), \tag{9}$$

$$T = \left(\frac{-\ln(U)}{\lambda\exp(\boldsymbol{\beta}^\top\boldsymbol{x})}\right)^{1/\nu}. \tag{10}$$

B Gaussian Copula for Correlated Categoricals

Generate $(Z_1, Z_2)^\top \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with off-diagonal ρ. Set $X_{\text{cont}} = Z_1$. Compute $U_2 = \Phi(Z_2)$. Define cutoffs $q_0, \ldots, q_K$ from target marginal probabilities. Assign $X_{\text{cat}} = k$ when $q_{k-1} \leq U_2 < q_k$. Rank correlation from the copula carries through; marginals match specification.