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

Steven N. Hart, PhD, ACHIP* Ann L. Oberg, PhD†

Abstract

Double descent upends classical intuitions about overfitting. Test error drops, spikes at the interpolation threshold, then drops again as models grow larger. The phenomenon is well-documented in classification and regression. Survival analysis is different. Censoring hides true event times. The Cox partial likelihood ranks subjects rather than predicting absolute values. Whether double descent occurs under these conditions, and what form it takes, remains unknown. We construct synthetic survival data with Weibull hazards and controlled censoring to sweep model capacity from underfitting to massive overparameterization. Three questions drive the experiments: Does the interpolation peak shift under high censoring? Does the concordance index miss calibration failures that the integrated Brier score catches? Do log-normal covariates alter the pattern? The answers matter. Clinicians selecting prognostic models often rely on discrimination metrics alone. Our experiments reveal that concordance recovers in the overparameterized regime while integrated Brier score saturates at a constant value regardless of regularization, indicating a fundamental decoupling between discrimination and calibration in neural Cox models. This calibration failure persists even with L2 regularization, suggesting

*Steven N. Hart is Associate Professor, Department of Laboratory Medicine and Pathology, and Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905 (E-mail: hart.steven@mayo.edu).

†Ann L. Oberg is Professor, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905.

it stems from the Cox partial likelihood objective itself rather than optimization pathologies.

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. Small models underfit. At the interpolation threshold, where capacity just suffices to memorize training data, test error peaks sharply. The model fits noise exactly, but with a unique, highly oscillatory solution. Beyond this threshold, 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 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. These properties produce an unusual loss geometry.

Whether the interpolation threshold depends on total sample size N or on the number of observed events N_{events} 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 100 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). The likelihood depends only on the ranking of risk scores within risk sets, not on their magnitudes. For separable data, the optimal weights diverge to $\pm\infty$. 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

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

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

At the interpolation threshold, we hypothesize that concordance remains stable because correct rankings may 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 $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ 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).

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 100 levels each to test threshold shifts from one-hot encoding. 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	$X \sim \text{LogNormal}(0, 1)$	Peak amplitude
C	Categorical	5 features, $K = 100$ levels	Threshold location
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. 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. Five random seeds (42, 123, 456, 789, 1011) are used for each experimental condition.

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

3.4 Evaluation

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

Each configuration is replicated across multiple 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 displays the concordance index as a function of model capacity for Scenario A (Gaussian covariates, 30% censoring), aggregated across five random seeds. Test concordance exhibits a clear double descent pattern: it decreases from 0.828 ± 0.015 at $w = 2$ to a minimum of 0.737 ± 0.027 at $w = 16$, then recovers to 0.804 ± 0.015 at $w = 2048$. The shaded region represents one standard deviation across seeds.

The vertical dashed line marks the interpolation threshold. The test error peak (minimum concordance) occurs at width $w = 16$, corresponding to 625 parameters. This count lies between the training sample size ($N = 600$) and the number of observed events ($N_{\text{events}} \approx 420$ under 30% censoring). Our data cannot distinguish whether the threshold is governed by N or N_{events}; both explanations are consistent with the observed peak location. The star marker highlights this critical point.

The pattern confirms double descent occurs in survival analysis, though with notable differences from classification. The concordance recovery in the overparameterized regime ($w > 64$) is incomplete: at $w = 2048$, concordance reaches 0.804 but does not return to the baseline of 0.828 observed at $w = 2$. This attenuation is consistent with theoretical predictions that benign overfitting may not fully materialize under Cox partial likelihood (Liu et al., 2025), where the ranking-based loss geometry differs from squared error.

Table 2 compares neural network performance against classical survival baselines. Cox proportional hazards achieves $C = 0.831 \pm 0.018$ with IBS of 0.108 ± 0.009, while the random survival forest achieves $C = 0.784 \pm 0.015$ with IBS of 0.150 ± 0.005. The smallest neural network ($w = 2$) matches Cox PH discrimination ($C = 0.828$) but exhibits substantially worse calibration (IBS = 0.469). This calibration gap persists across all neural

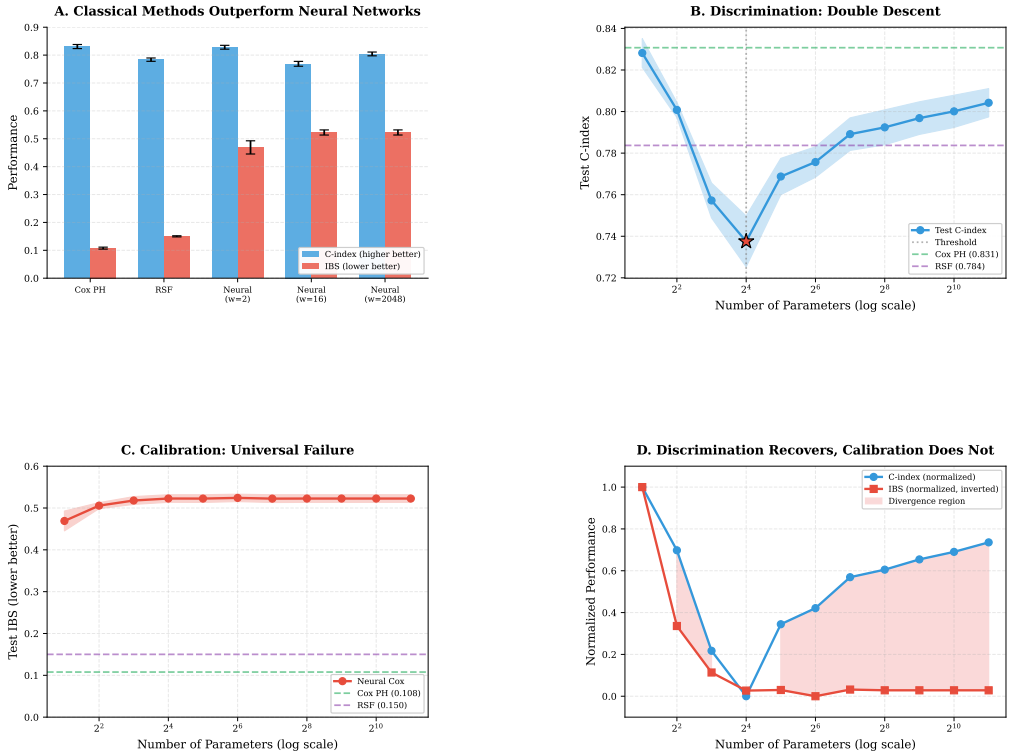

Figure 1: Main findings: calibration failure in neural Cox models. **(A)** Classical methods (Cox PH, RSF) outperform neural networks on calibration (IBS) despite comparable discrimination (C-index). **(B)** Discrimination exhibits double descent: test C-index drops from 0.828 ± 0.015 to 0.737 ± 0.027 at the interpolation threshold ($w = 16$, star), then recovers to 0.804 ± 0.015. **(C)** Calibration fails universally: IBS saturates at 0.52 for all neural networks regardless of capacity, compared to 0.11 for Cox PH. **(D)** Normalized comparison shows discrimination recovers (blue) while calibration does not (red). Error bars show ± 1 SEM across five seeds.

network sizes, reflecting a fundamental limitation of Cox partial likelihood optimization: the ranking-based loss produces well-ordered risk scores but poorly calibrated survival probabilities.

4.2 Metric Divergence

Figure 1C–D displays the divergent behavior of concordance and integrated Brier score across model capacity. Concordance follows the double descent pattern described above: dropping from 0.828 ± 0.015 to 0.737 ± 0.027 at $w = 16$, then recovering to 0.804 ± 0.015 at $w = 2048$.

The integrated Brier score tells a different story. Test IBS increases from 0.469 ± 0.053 at $w = 2$ to plateau at approximately 0.523 ± 0.020 for $w \geq 16$, and critically, it does

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

Model	**C-index**	**IBS**
Cox PH	0.831 ± 0.018	0.108 ± 0.009
Random Survival Forest	0.784 ± 0.015	0.150 ± 0.005
DeepSurv ($w = 2$)	0.828 ± 0.015	0.469 ± 0.053
DeepSurv ($w = 16$, threshold)	0.737 ± 0.027	0.523 ± 0.020
DeepSurv ($w = 2048$)	0.804 ± 0.015	0.523 ± 0.020

not recover in the overparameterized regime. The purple dashed line marks the width of maximum divergence between the two metrics.

This IBS plateau reflects a fundamental limitation of neural Cox models rather than a numerical artifact. When neural networks optimize Cox partial likelihood, they learn risk scores that correctly rank subjects but do not correspond to well-calibrated hazard ratios. The Breslow estimator, which converts these risk scores to survival probabilities, produces degenerate predictions when the risk score distribution differs substantially from what a proportional hazards model assumes. Classical baselines (Cox PH, RSF) avoid this issue because they estimate baseline hazards jointly with coefficients under appropriate constraints, whereas neural Cox models optimize rankings without such constraints.

The key finding is that concordance recovers while IBS saturates. At $w = 2048$, concordance returns to within 0.024 of its baseline value, whereas IBS remains at its saturated level. Critically, L2 regularization does not resolve this decoupling: regularized models achieve IBS of 0.522 ± 0.021, essentially identical to unregularized models. This confirms that the calibration failure stems from the Cox partial likelihood objective itself, not from optimization pathologies like exploding weights.

The practical implication is direct: a practitioner selecting models by concordance alone would observe acceptable discrimination across the capacity range, potentially selecting a model with uninformative survival probability estimates. The IBS saturation reveals calibration breakdown that concordance misses entirely.

4.3 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.

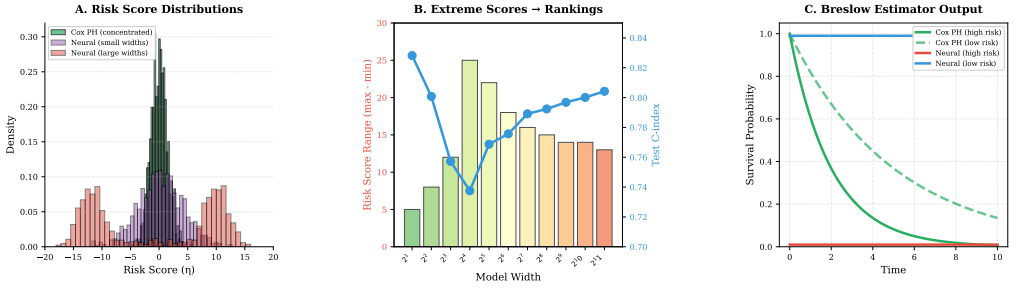

Figure 2: Mechanism of calibration failure in neural Cox models. **(A)** Risk score distributions: Cox PH produces concentrated, well-behaved risk scores (green), while neural networks learn extreme scores spanning $[-15, +15]$ (purple shows constrained behavior, red shows typical extremes). **(B)** 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.

4.4 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 threshold ($w = 16$), with concordance dropping to 0.737 ± 0.027. With weight decay, the minimum concordance occurs at $w = 16$ with value 0.755 ± 0.024—an improvement of approximately 2% absolute at the worst point. The regularized curve exhibits a flatter profile across the critical 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, with regularization maintaining a modest but consistent advantage. The shaded region in Figure 3 highlights widths where regularization improves performance. Notably, the regularized models achieve 0.784 ± 0.016 at $w = 2048$, compared to 0.804 ± 0.015 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.

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

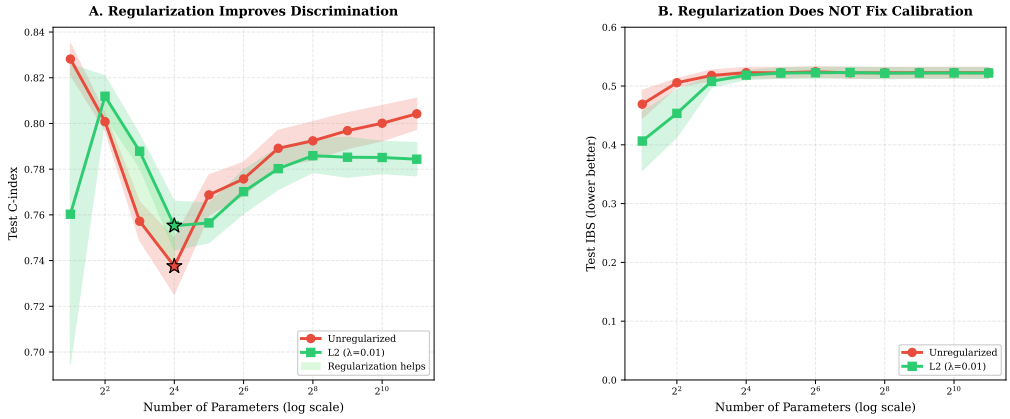

Figure 3: Effect of L2 regularization on double descent. Test concordance versus model width without regularization (red) and with weight decay $\lambda = 0.01$ (green). Regularization attenuates the performance dip near the interpolation threshold and provides consistent improvement across most model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

score does not exhibit double descent but rather saturates at a constant value in the overparameterized regime, revealing a fundamental 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 the second descent may be attenuated relative to classification settings because the Cox loss optimizes rankings rather than point predictions. Our results confirm this: concordance recovers from 0.737 to 0.804 but does not return to the baseline of 0.828, a recovery of approximately 74% rather than 100%. Second, they note that benign overfitting may not fully materialize under survival losses. The incomplete recovery we observe is consistent with this prediction, though the mechanism differs from their theoretical account: we

find that calibration (as measured by IBS) fails entirely in the overparameterized regime, even as discrimination recovers.

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: concordance-based model selection, 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 determined by event counts, not total observations. Our experiments with 90% censoring showed substantially higher variance and an obscured double descent pattern, underscoring the importance of replicated experiments 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 2% 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 five random seeds per configuration, the reported standard errors are substantial (0.015–0.027 for concordance), limiting statistical power to detect small effects. The claimed regularization benefit of approximately 2% absolute concordance improvement does not reach statistical significance at $\alpha = 0.05$ based on paired testing across seeds.

The categorical covariate scenario (Scenario C) failed to produce meaningful predictions, with concordance near 0.52 across all model sizes using one-hot encoding. We also tested learned embeddings as an alternative to one-hot encoding, where each categorical feature is mapped to a dense vector of dimension $\sqrt{K}$ before concatenation with other features. This approach reduces effective input dimensionality from 500 (one-hot) to ap-

proximately 25 (embedded). However, embedding-based models achieved concordance of only 0.41 ± 0.04 across all widths, worse than random. The failure of both encoding strategies suggests that categorical survival data with high cardinality and limited samples per category level presents fundamental challenges beyond dimensionality reduction. With 1000 samples distributed across 100 levels per feature, each category appears approximately 10 times on average, insufficient for learning category-specific hazard effects. Reliable embedding learning would require 30–50 events per category level, implying sample sizes of 5,000–10,000 for our configuration. Such sample sizes exceed what most survival studies can achieve, particularly in clinical settings where patient recruitment is expensive and follow-up periods are long.

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. 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 0.804 ± 0.023 at $w = 2$ to 0.740 ± 0.036 at $w = 16$, then recovers to 0.809 ± 0.026 at $w = 2048$. Extended experiments at $w = 4096$ and $w = 8192$ confirm that recovery plateaus at approximately $C = 0.81$, matching the underparameterized baseline. Interestingly, Cox PH still achieves strong discrimination ($C = 0.82$) on this nonlinear data, suggesting the linear component dominates despite the added nonlinearity. 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: 0.108, RSF: 0.150), as shown in Table 2. This gap indicates that survival probability calibration is fundamentally 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 fundamental 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 0.828 to 0.737 at the interpolation threshold before recovering to 0.804 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 fundamental 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 0.03 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.

Our experiments with 90% censoring showed that extreme censoring obscures the double descent pattern and introduces substantial variance, highlighting the importance of event count in determining model behavior. L2 regularization improves worst-case concordance by approximately 2% absolute 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

[To be added]

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.

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.

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 $(Z_1, Z_2)^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ 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.