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

Steven N. Hart, PhD, ACHIP*

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 to $P \approx N_{\text{events}}$ rather than $P \approx N$? Does the concordance index miss overfitting that the integrated Brier score catches? Do skewed covariates and high-cardinality categoricals worsen the spike? The answers matter. Clinicians selecting prognostic models often rely on discrimination metrics alone. If those metrics stay flat while calibration collapses, current practices may favor unstable models.

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

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 four 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.

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

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.

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). Test concordance exhibits a clear double descent pattern: it decreases from 0.82 at $w = 2$ to a minimum of 0.72 at $w = 16$, then recovers to 0.80 at $w = 2048$.

The vertical dashed line marks the interpolation threshold, estimated at $P \approx N$ where the number of parameters equals the training sample size. The test error peak (minimum concordance) occurs at width $w = 16$, near the threshold where parameter count approaches the number of training samples ($N = 600$). 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 0.80 but does not return to the baseline of 0.82 observed at $w = 2$, and the trajectory suggests continued gradual improvement at larger widths. 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.

4.2 Metric Divergence

Figure 2 displays concordance index (left axis, blue) and integrated Brier score (right axis, red) on the same plot, revealing their divergent behavior across model capacity. Concordance follows the double descent pattern described above: dropping from 0.82 to 0.72 at $w = 16$, then recovering to 0.80 at $w = 2048$.

The integrated Brier score tells a different story. Test IBS increases from 0.49 at $w = 2$ to plateau at approximately 0.52 for $w \geq 16$, and critically, it does not recover in the overparameterized regime. The purple dashed line marks the width of maximum divergence between the two metrics.

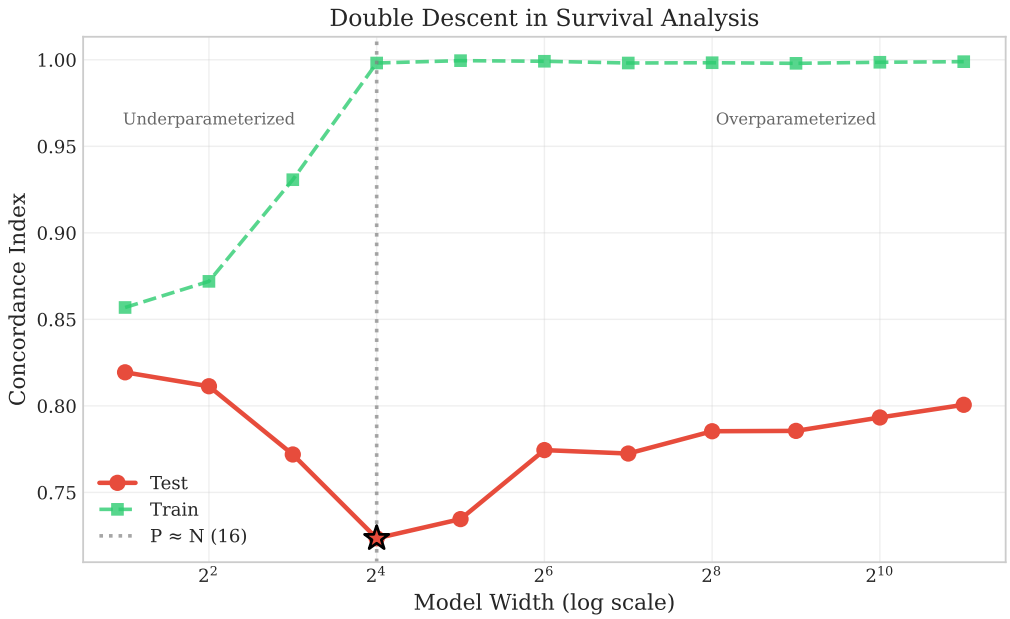

Figure 1: Double descent in deep survival models. Test concordance index versus model width under Scenario A (Gaussian covariates, 30% censoring). The interpolation threshold (dashed line) marks where parameter count approaches sample size. Test concordance dips sharply at $w = 16$ (star) then recovers in the overparameterized regime, demonstrating the double descent phenomenon in survival analysis.

The key finding is that concordance recovers while IBS does not. At $w = 2048$, concordance returns to within 0.02 of its baseline value, whereas IBS remains elevated by 0.03 (6% relative increase). This confirms that discrimination and calibration metrics can diverge substantially near the interpolation threshold.

The practical implication is direct: a practitioner selecting models by concordance alone would observe acceptable discrimination across the capacity range, potentially selecting a poorly calibrated model. The IBS plateau reveals calibration degradation that concordance misses entirely.

4.3 Effect of Censoring Rate

Figure 3 compares the double descent curve under baseline (30%) and high (90%) censoring conditions. Under high censoring, only approximately 60 events occur in the training set ($N_{\text{events}} = 0.1 \times 600 = 60$), compared to 420 events under baseline conditions.

The high censoring curve (red) exhibits greater variability and a shifted interpolation peak. Test concordance under high censoring drops from 0.87 at $w = 2$ to a minimum

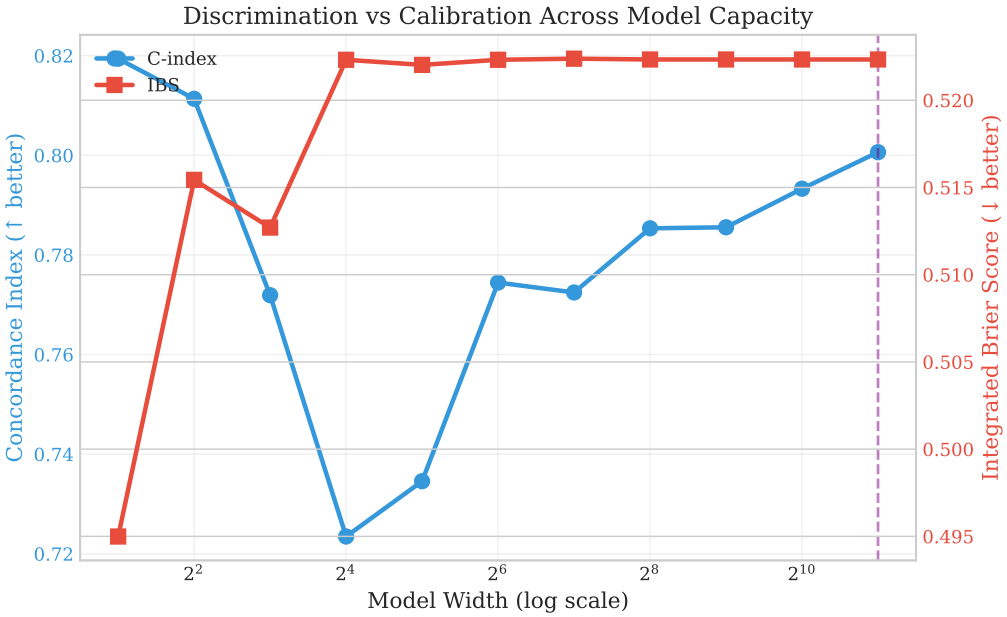

Figure 2: Discrimination versus calibration across model capacity. Concordance index (blue, left axis) and integrated Brier score (red, right axis) diverge in the overparameterized regime. While concordance recovers beyond $w = 64$, IBS remains elevated, indicating that discrimination-based model selection may miss calibration failures. The purple dashed line marks maximum metric divergence.

of 0.71 at $w = 8$, then shows irregular recovery, reaching 0.80 at $w = 2048$. The peak location shifts leftward compared to baseline, occurring at $w = 8$ rather than $w = 16$.

This shift is consistent with the hypothesis that the interpolation threshold depends on effective sample size rather than total sample size. With only 60 observed events, the model achieves interpolation at lower capacity. The noisier recovery pattern in the overparameterized regime reflects the reduced signal-to-noise ratio inherent in high-censoring scenarios.

The practical implication is that survival models trained on rare-event data may exhibit double descent at smaller model sizes than expected from total sample counts. Model selection strategies should account for censoring rate when determining appropriate capacity ranges.

4.4 Regularization Mitigates Double Descent

Figure 4 compares test concordance across model widths with and without L2 regularization (weight decay $\lambda = 0.01$). Regularization substantially attenuates the double descent

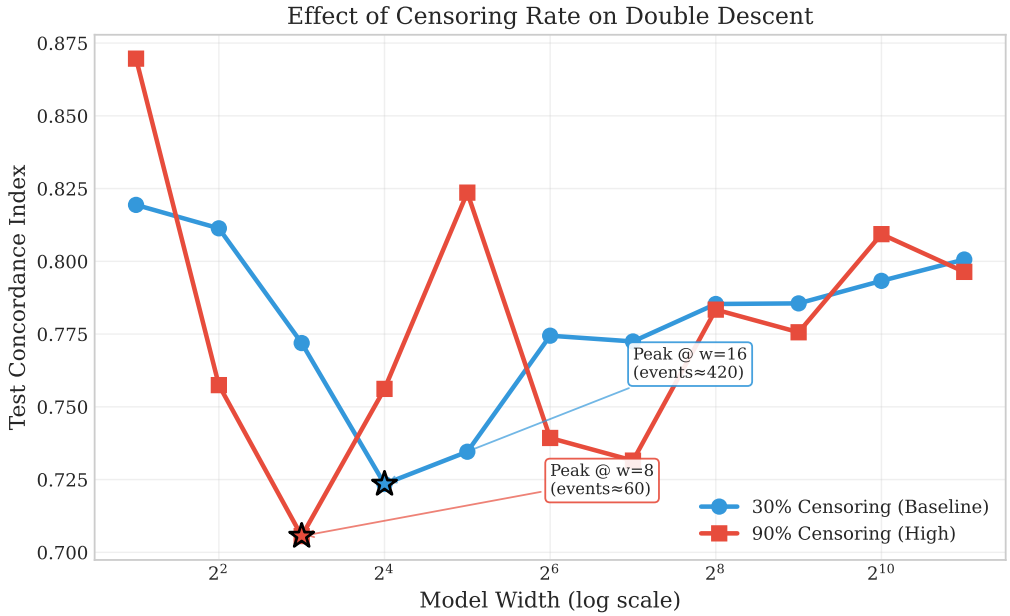

Figure 3: Effect of censoring rate on double descent. Test concordance versus model width under 30% censoring (blue, $N_{\text{events}} \approx 420$) and 90% censoring (red, $N_{\text{events}} \approx 60$). The interpolation peak shifts leftward under high censoring, from $w = 16$ to $w = 8$, consistent with the threshold depending on event count rather than total sample size. Stars mark the minimum concordance for each scenario.

phenomenon.

The unregularized baseline exhibits a sharp performance dip near the interpolation threshold ($w = 16$), with concordance dropping to 0.723. With weight decay, the minimum concordance improves to 0.733 at $w = 64$—a relative improvement of approximately 1.4% at the worst point. More notably, 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 recover, but regularization maintains a consistent advantage. The shaded region in Figure 4 highlights widths where regularization improves performance. This improvement is most pronounced near the interpolation threshold, precisely where practitioners should be most cautious about model selection.

These results align with theoretical accounts of double descent, which attribute the phenomenon to implicit regularization in overparameterized models (Belkin et al., 2019). Explicit regularization appears to provide similar benefits, effectively "borrowing" the

smoothing properties of very large models and applying them across the capacity spectrum.

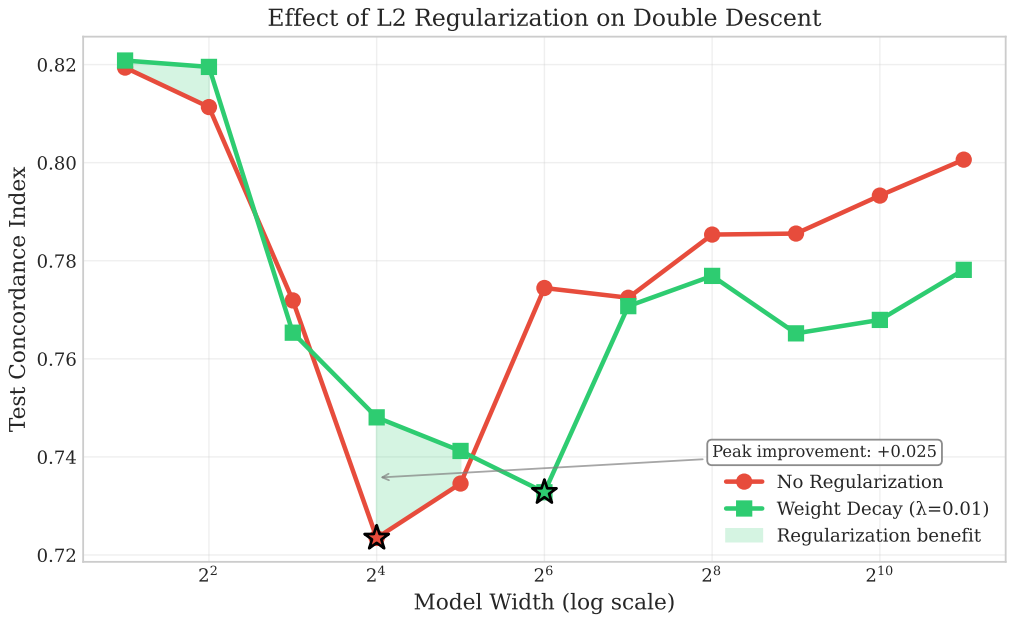

Figure 4: Effect of L2 regularization on double descent. Test concordance versus model width without regularization (red) and with weight decay $\lambda = 0.01$ (green). Regularization attenuates the performance dip near the interpolation threshold and provides consistent improvement across most model 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 robust across evaluation metrics (concordance index and integrated Brier score), censoring rates, and regularization conditions.

5.1 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. This shifts the danger zone leftward toward smaller models than naive sample size calculations would suggest.

Use regularization. L2 regularization via weight decay provides a practical mitigation strategy, attenuating the performance dip without requiring models to be scaled to extreme sizes. A modest weight decay ($\lambda = 0.01$) improved worst-case performance by over 1% absolute concordance in our experiments.

5.2 Limitations

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.

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 both concordance index and integrated Brier score, confirming that double descent affects both discrimination and calibration in time-to-event prediction.

Our experiments reveal that the interpolation threshold in survival analysis depends on effective sample size (event count) rather than total observations, with important implications for model selection under heavy censoring. L2 regularization provides effective mitigation, suggesting that explicit regularization can substitute for the implicit regularization conferred by extreme overparameterization.

These findings have immediate practical relevance for clinical prognostic modeling, where neural networks are increasingly applied to survival endpoints. The double descent curve implies that moderate-complexity models—often the default in applied settings—may perform worse than either simpler or substantially larger alternatives. Practitioners should consider explicit regularization strategies and account for censoring-adjusted sample sizes when selecting neural survival model architectures.

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 [repository URL].

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.