KAPLAN: KOLMOGOROV-ARNOLD PROGNOSTIC LEARNABLE ACTIVATION NETWORKS FOR SURVIVAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Survival analysis aims to model how covariates and time jointly shape the time-to-event distribution under right censoring. Classical methods such as the Cox model and generalised additive models (GAMs) require interactions and time-varying effects to be manually specified, which is increasingly impractical on rich clinical datasets. We introduce KAPLAN-HR, a B-spline Kolmogorov-Arnold Network (KAN) for nonparametric estimation of the conditional hazard as a joint function of covariates and time. A single-layer KAPLAN-HR model recovers a GAM, while deeper architectures capture interactions and time-varying effects through composition. We establish a convergence rate for the nonparametric KAN hazard estimator whose exponent depends only on the smoothness of the underlying KAN representation and not on the covariate dimension, thereby mitigating the curse of dimensionality for KAN-representable targets. In evaluations over six clinical benchmark datasets, KAPLAN-HR attains predictive performance competitive with established statistical and deep learning survival methods.

1 INTRODUCTION

Time-to-event data is ubiquitous in biostatistics, engineering, and business, but analysing it is complicated because the events of interest are often only partially observed. Patients may leave a study before their disease progresses, or a machine may be replaced before it fails, so only a lower bound on the true event time is ever observed for these subjects. Survival analysis addresses this by modelling the time-to-event distribution under this incomplete observation, called right censoring.

The Cox proportional hazards model is the de facto standard for survival regression (Cox, 1972), but it requires careful manual work to accommodate interactions between covariates and restricts their relationship with time. Among the more flexible alternatives, spline-based approaches for modelling the hazard have been studied for over three decades. For example, generalised additive models (GAMs) fit a spline for each covariate's effect on the hazard (Hastie & Tibshirani, 1995; Wood, 2025). HARE models the log-hazard as linear splines and selects main effects and pairwise interaction terms by stepwise search (Kooperberg et al., 1995). However, these models are structurally limited in how they represent interactions, as GAMs include them only when specified manually, and HARE cannot represent three-way or higher interactions.

This limitation is increasingly consequential, particularly in medicine and biostatistics, where modern datasets such as UK Biobank (Sudlow et al., 2015) and MIMIC (Johnson et al., 2016) are far richer than those available before, motivating methods that capture interactions and time-varying effects without manual specification. Deep learning approaches (Wiegrebe et al., 2024; Chen, 2024) relax these structural restrictions and perform well on high-dimensional data such as imaging, but depart from the additive structure that makes classical models simple to analyse and interpret.

Kolmogorov-Arnold Networks (KANs) (Liu et al., 2025b) occupy a structural middle ground. Each edge of a KAN applies a learnable univariate function and each node sums its inputs. When the edge functions are splines, a one-layer KAN is an additive model of its inputs, and each term is a univariate function that can be inspected directly. Deeper architectures introduce interactions through composition, at which point the first-layer edges remain univariate functions of the inputs but no longer represent marginal effects. Depth therefore trades structural transparency for flexibility. Recent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

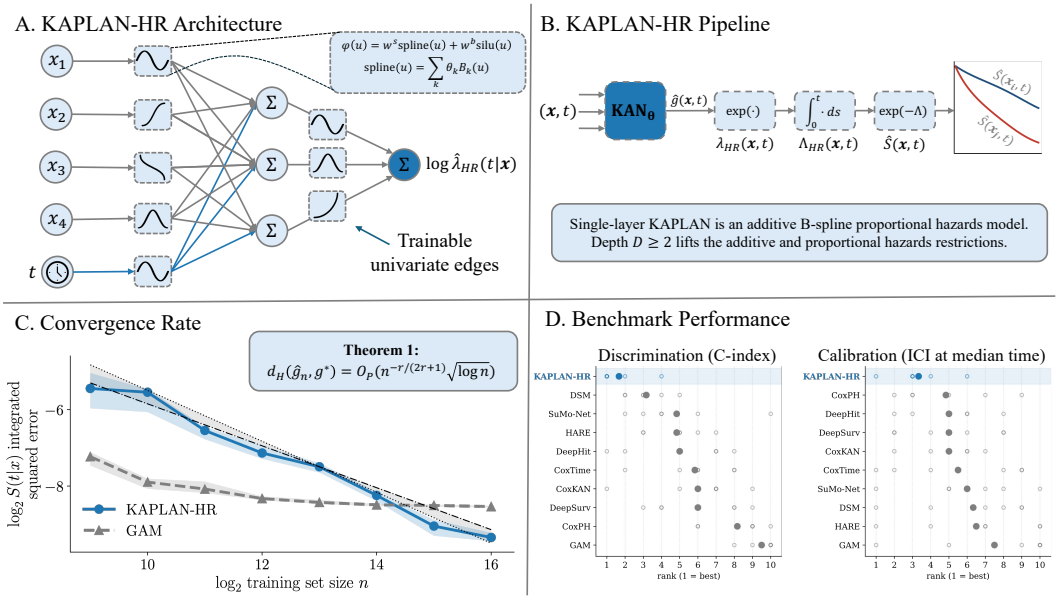


Figure 1: **Overview of KAPLAN-HR. (A)** KANs compose layers of learnable univariate functions. We propose using a KAN to estimate the log-hazard as a joint function of covariates and time. **(B)** Numerical integration transforms the predicted log-hazard into predicted survival curves. A single-layer KAPLAN-HR is an additive B-spline proportional hazards model, while deeper models capture covariate interactions and time-varying effects through composition. **(C)** We establish a nonparametric convergence rate of $O_P(n^{-r/(2r+1)}\sqrt{\log n})$ for the KAPLAN-HR estimator at component smoothness r, corroborated by simulation studies. **(D)** Across six clinical benchmarks, KAPLAN-HR attains the best mean rank of ten models on discrimination and calibration.

work has applied KANs to survival analysis via different parameterisations, including proportional hazards (Knottenbelt et al., 2025; Cheng & Hu, 2025), accelerated failure time (Jose et al., 2025), and nonparametric hazard regression (Mastroleo et al., 2026). These studies report encouraging results, but they focus on empirical performance and on interpretability via edge visualisation and symbolic regression, and none provides a theoretical analysis. Analysing the statistical properties of KAN-based survival estimators is therefore of substantive interest.

We introduce KAPLAN-HR, a B-spline KAN model that estimates the conditional hazard as a joint function of covariates and time. A single-layer KAPLAN-HR is a B-spline GAM under the proportional hazards (PH) restriction, and deeper architectures capture covariate interactions and time-varying effects through composition. We make the following contributions: (i) We propose KAPLAN-HR as a bridge between additive spline-based hazard models and deep survival models, where depth controls the level of structural relaxation, trading transparency for flexibility. (ii) We establish a convergence rate of the proposed nonparametric estimator to the true KAN-representable log-hazard whose exponent is independent of the input dimension. (iii) We empirically evaluate KAPLAN-HR on six clinical benchmark datasets, where it attains predictive performance competitive with established statistical and deep learning survival models.

2 Preliminaries

Denote by $\mathcal{D}$ a time-to-event dataset $\{(\mathbf{x}_i, Y_i, \Delta_i)\}_{i=1}^n$, comprising n i.i.d. observations of continuous covariates $\mathbf{x}_i \in [0,1]^d$, observed times $Y_i \in [0,1]$, and event indicators $\Delta_i \in \{0,1\}$. We assume the covariates are scaled to the unit cube, and the times are scaled by a finite study horizon $\tau_{\max} \in (0, \infty)$, so $Y_i = 1$ corresponds to subject i surviving to or beyond $\tau_{\max}$. Let T_i denote the (possibly unobserved) time the event of interest occurred, and C_i the censoring time. Under right censoring, we only observe the earlier of the two: $Y_i := \min\{T_i, C_i\}$ and $\Delta_i := \mathbb{1}\{T_i \leq C_i\}$.

We adopt the common assumption that T and C are independent conditional on X. We consider the censoring distribution $\mathbb{P}_{C|\mathbf{x}}$ to combine a continuous density $f_C(\cdot \mid \mathbf{x})$ on $[0, 1)$ with a point mass $\alpha(\mathbf{x}) \in [0, 1]$ at $t = 1$ (with $f_C \geq 0$ and $\int_0^1 f_C(s \mid \mathbf{x})\, ds + \alpha(\mathbf{x}) = 1$). This accommodates administrative censoring, where every subject still at risk at $\tau_{\max}$ is censored simultaneously, which is a standard feature of clinical survival datasets.

Hazard model The ubiquitous Cox PH model is defined in terms of the hazard function $\lambda(t \mid \mathbf{x}) = -\frac{d}{dt} \ln S(t \mid \mathbf{x})$, which it decomposes into a covariate-invariant baseline hazard function $\lambda_0(t)$ and a time-invariant log-partial hazard function $g(\mathbf{x}; \boldsymbol{\theta}) : \lambda_{\text{PH}}(t \mid \mathbf{x}) = \lambda_0(t) \exp(g(\mathbf{x}; \boldsymbol{\theta}))$ (Cox, 1972). This constrains the predicted hazard to the same shape as the baseline $\lambda_0(t)$ and the covariate effects to be constant across time. We use a more flexible hazard regression (HR) model where $g(\cdot)$ expresses both the relationship with time and the covariate effects, similar to (Kooperberg et al., 1995):

$$\lambda_{\text{HR}}(t \mid \mathbf{x}) = \exp(g(\mathbf{x}, t; \boldsymbol{\theta})) \tag{1}$$

We assume the true log-hazard g^* is bounded, $\|g^*\|_\infty \leq M_0 < \infty$, $M_0 \geq 1$.

We fit the model by minimising the negative log-likelihood (NLL) of the observed data:

$$\mathcal{L}_{\text{HR}}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i \log \lambda_{\text{HR}}(Y_i \mid \mathbf{x}_i) - \Lambda_{\text{HR}}(Y_i \mid \mathbf{x}_i)\right] \tag{2}$$

The cumulative hazard is $\Lambda_{\text{HR}}(t \mid \mathbf{x}) = \int_0^t \lambda_{\text{HR}}(s \mid \mathbf{x})\, ds$, and the survival $S(t \mid \mathbf{x}) = e^{(-\Lambda_{\text{HR}}(t|\mathbf{x}))}$.

3 METHODOLOGY

KAPLAN-HR represents the learnable function $g(\cdot)$ in Eq. (1) as a KAN. KANs are neural networks in which the edges carry trainable univariate functions and the nodes sum their inputs (Liu et al., 2025b). Each edge is parameterised as the weighted sum of a B-spline and a residual SiLU term:

$$\varphi(u) = w^{\text{s}} \sum_k \theta_k B_k(u) + w^{\text{b}} \operatorname{silu}(u) \tag{3}$$

where $\{B_k\}$ is a B-spline basis, $\boldsymbol{\theta}$ its coefficient vector, and $w^{\text{s}}, w^{\text{b}}$ are trainable scalars.

Consider a KAN consisting of D layers Φ_ℓ having d_ℓ nodes, and trainable edges $\varphi_{\ell,i,j}$ of the form Eq. (3), with coefficient vector $\boldsymbol{\theta}^{(\ell,i,j)}$, that are aggregated by the nodes via summation:

$$z_j^{(\ell)} = \left[\Phi_\ell\left(\mathbf{z}^{(\ell-1)}\right)\right]_j = \sum_{i=1}^{d_{\ell-1}} \varphi_{\ell,i,j}\left(z_i^{(\ell-1)}\right), \qquad \ell = 1, \ldots, D, \quad j \in \{1, \ldots, d_\ell\} \tag{4}$$

The full network is the composition $\text{KAN}_{\boldsymbol{\theta}}(\mathbf{z}) = (\Phi_D \circ \cdots \circ \Phi_1)(\mathbf{z}) = z_1^{(D)}$. We refer readers to Noorizadegan et al. (2026) for further details on KAN architecture, training, and basis functions.

Let $\text{KAN}_{\boldsymbol{\theta}}$ be defined as above, with input $\mathbf{z}^{(0)} = (\mathbf{x}, t) \in \mathcal{Z} := [0, 1]^{d_0}$ where $d_0 = d + 1$ (covariate dimension d plus time). We propose the KAPLAN-HR survival model:

$$\mathbf{z}^{(0)} = (\mathbf{x},\, t) \in \mathcal{Z}, \quad \lambda_{\text{HR}}(t \mid \mathbf{x}) = \exp(\text{KAN}_{\boldsymbol{\theta}}(\mathbf{x}, t)) \tag{5}$$

Observe that a single-layer KAPLAN-HR model is structurally an additive B-spline PH model, with a univariate spline for each covariate and for time:

$$\log \lambda_{\text{HR}}(t \mid \mathbf{x}) = \sum_{i=1}^{d} \varphi_i(x^{(i)}) + \varphi_t(t) \implies \lambda_{\text{HR}}(t \mid \mathbf{x}) = \underbrace{\exp(\varphi_t(t))}_{\lambda_0(t)} \cdot \exp\left(\sum_{i=1}^{d} \varphi_i(x^{(i)})\right) \tag{6}$$

For $D \geq 2$, the hidden representations combine time and covariate effects. For example, for $D = 2$:

$$\log \lambda_{\text{HR}}(t \mid \mathbf{x}) = \sum_{j=1}^{d_1} \varphi_{2,j,1}\left(\sum_{i=1}^{d} \varphi_{1,i,j}(x^{(i)}) + \varphi_{1,d_0,j}(t)\right) \tag{7}$$

Network depth therefore removes both the additive and proportional-hazards restrictions, admitting covariate interactions and time-varying effects. The cumulative hazard Λ_{HR} has no closed form for KAN-modelled λ_{HR}, so we approximate it by numerical integration. We define a grid of time points $\tau_0 < \tau_1 < \cdots < \tau_K$ (with $\tau_0 \equiv 0$). Then, on each interval $[\tau_{k-1}, \tau_k]$ we apply the trapezoidal rule:

$$\Delta\hat{\Lambda}[k \mid \mathbf{x}] = \tfrac{1}{2}\left(\lambda_{\text{HR}}(\tau_{k-1} \mid \mathbf{x}) + \lambda_{\text{HR}}(\tau_k \mid \mathbf{x})\right)(\tau_k - \tau_{k-1}) \;\approx\; \int_{\tau_{k-1}}^{\tau_k} \lambda_{\text{HR}}(u \mid \mathbf{x})\, du \tag{8}$$

We use square bracket notation, such as $\hat{\Lambda}[k \mid \mathbf{x}]$, to distinguish discrete-time survival quantities. Accumulating these increments yields the discrete survival estimate:

$$\hat{S}[k \mid \mathbf{x}] := \exp\Big(-\hat{\Lambda}_{\text{HR}}[k \mid \mathbf{x}]\Big) = \exp\left(-\sum_{m=1}^{k} \Delta\hat{\Lambda}[m \mid \mathbf{x}]\right) \tag{9}$$

The KAPLAN-HR model is trained by minimising the NLL $\mathcal{L}_{\text{HR}}$ (Eq. (2)). We apply an ℓ_1 penalty on the B-spline coefficients and their first differences:

$$\mathcal{L}_{\ell_1} = \lambda_{\ell_1} \cdot \sum_{\ell=1}^{D} \frac{1}{d_\ell} \sum_{i=1}^{d_{\ell-1}} \sum_{j=1}^{d_\ell} \left[\underbrace{\|\boldsymbol{\theta}^{(\ell,i,j)}\|_1}_{\rho_{\text{magnitude}}} + \underbrace{\|\Delta\boldsymbol{\theta}^{(\ell,i,j)}\|_1}_{\rho_{\text{difference}}} \right] \tag{10}$$

where Δ is the forward difference on the coefficient index. We detail the implementation, configuration, and training procedure in Appendix B.

4 Theoretical Analysis

We establish a rate of convergence of KAPLAN-HR to the ground truth log-hazard function using the method of sieves (Shen & Wong, 1994) as evidence of our method's statistical correctness. Under the structural assumption that the true log-hazard admits a representation as a layered composition of smooth univariate functions, the rate's exponent depends only on the smoothness of those univariate components, with the input dimension affecting only the constant. KAPLAN-HR therefore mitigates the curse of dimensionality for true log-hazards in this problem class.

Denote by $\mathcal{X} := [0,1]^d$ the covariate domain and $\mathcal{Z} := \mathcal{X} \times [0,1] = [0,1]^{d_0}$ the input domain including time. Further denote integer smoothness $r \geq 1$. We define $\mathcal{F}_r^{\text{KAN}}$ as the class of functions $g : \mathcal{Z} \to \mathbb{R}$ admitting a representation as an additive KAN composition of univariate $W^{r,\infty}$ functions. Fix this composition's depth $D \geq 1$, input dimension d_0, and uniform hidden layer widths $d_1 = \cdots = d_{D-1} = Q$. Each edge's $W^{r,\infty}$-norm is at most M and every composition $g \in \mathcal{F}_r^{\text{KAN}}$ satisfies $\|g\|_{\infty,\mathcal{Z}} \leq M_0 < M$ for some $M_0 \in [1, \infty)$.

Define the sieve $\mathcal{G}_n$ consisting of all KANs with edges carrying B-splines of order $m \geq \max(2, r)$ on $J_n \asymp n^{1/(2r+1)}$ uniform interior knots. The splines' coefficients are bounded by C_θ, and every $g \in \mathcal{G}_n$ satisfies $\|g\|_{\infty,\mathcal{Z}} \leq M' := 2M_0$.

We measure the distance between g^* and g_n by the covariate-averaged Hellinger distance:

$$d_H(\hat{g}_n, g^*) = \sqrt{\mathbb{E}_{X \sim P_X}[H^2(\mathbb{P}_{\hat{g}_n,X} \parallel \mathbb{P}_{g^*,X})]} \tag{11}$$

between conditional distributions $\mathbb{P}_{g,X}$ of $(Y, \Delta) \mid X$ under log-hazard g, where P_X is the covariate distribution and $H^2(p,q) = \int(\sqrt{p} - \sqrt{q})^2$.

Assumptions

- (C1) $T \perp C \mid X$ independent censoring.
- (C2) Conditional on X, T has (true) hazard $\lambda(t \mid \mathbf{x}) = e^{g^*(\mathbf{x},t)}$ on $[0,1]$ with $\|g^*\|_{\infty,\mathcal{Z}} \leq M_0$ for some $M_0 \in [1, \infty)$.
- (C3) Conditional on X, $C \in [0,1]$ has a continuous density $f_C(\cdot \mid \mathbf{x})$ on $[0,1)$ together with a point mass $\alpha(\mathbf{x}) \in [0,1]$ at $t = 1$, with $\int_0^1 f_C(s \mid \mathbf{x})\, ds + \alpha(\mathbf{x}) = 1$.

(C4) The true log-hazard g^* lies in $\mathcal{F}_r^{\mathrm{KAN}}(d_0, D, Q, M_0, M)$ with $M_0 < M$.

(C5) The estimator lies in the sieve: $\hat{g}_n = g_{\hat{\boldsymbol{\theta}}_n}$ for a coefficient vector $\hat{\boldsymbol{\theta}}_n$ with entries bounded by C_θ, and $\|\hat{g}_n\|_{\infty,\mathcal{Z}} \le M'$.

(C6) $\hat{g}_n$ falls short of the log-likelihood of the sieve approximant $\bar{g}_n$ of Proposition 1 by at most $\eta_n := C_\eta\, n^{-2r/(2r+1)} \log n$ for an n-independent constant $C_\eta \ge 0$.

Assumption (C2) positions the true log-hazard g^* (the HR parameterisation) and caps $\|g^*\|_{\infty,\mathcal{Z}}$. (C3) formalises the censoring mechanism and explicitly accommodates administrative censoring at $t = 1$. (C4) is the structural assumption on g^* that the sieve targets. (C5) puts $\hat{g}_n$ inside the sieve and (C6) limits the slack between $\hat{g}_n$ and the sieve approximant that the rate theorem can absorb.

Theorem 1 (KAPLAN-HR Hellinger rate)**.** *Under (C1)-(C6):*

$$d_H(\hat{g}_n,\, g^*) = O_P\Big(n^{-r/(2r+1)} \sqrt{\log n}\Big)$$

Denote by $\|f\|_{L^2}$ the L^2 norm with P_X-weighted covariates and uniform time on $[0, \tau], \tau \in (0, 1)$: $\|f\|_{L^2}^2 := \mathbb{E}_{X\sim P_X} \int_0^\tau f(\mathbf{x}, t)^2\, dt$. Further, assume censoring positivity $\inf_{\mathbf{x},\, t\in[0,\tau]} S_C(t \mid \mathbf{x}) \ge c, c > 0$, which is not required for Theorem 1.

Corollary 1 (P_X-weighted L^2 log-hazard rate)**.** *Under (C1)-(C6) and censoring positivity:*

$$\|\hat{g}_n - g^*\|_{L^2} = O_P\Big(n^{-r/(2r+1)} \sqrt{\log n}\Big)$$

This follows from the curvature inequality $\|g - g^*\|_{L^2}^2 \lesssim d_H^2(g, g^*)$ on $\|g\|_{\infty,\mathcal{Z}} \le M'$ under the additional censoring positivity hypothesis, applied to Theorem 1.

Corollary 2 (Survival function L^2 rate)**.** *Under the hypotheses of Corollary 1, the survival estimator* $\hat{S}_n(t \mid \mathbf{x}) = \exp\Big(-\int_0^t e^{\hat{g}_n(\mathbf{x},s)}\, ds\Big)$ *satisfies:*

$$\|\hat{S}_n - S^*\|_{L^2} = O_P\Big(n^{-r/(2r+1)} \sqrt{\log n}\Big)$$

This follows from the Lipschitz stability of $g \mapsto \exp(-\int_0^\cdot e^{g(\mathbf{x},s)}\, ds)$ on $\|g\|_{\infty,\mathcal{Z}} \le M'$ applied to Corollary 1.

Proof outline Theorem 1 is proved by sieve maximum-likelihood estimation. The full proof is given in Appendix A. We prove four propositions that supply the inputs that the rate theorem (van der Vaart & Wellner, 2023, Thm. 3.4.12) requires for an O_P rate in Hellinger distance (Appendix A.3).

1. Proposition 1: Approximation error bound. Every target $g^* \in \mathcal{F}_r^{\mathrm{KAN}}$ has a sieve approximant $\bar{g}_n \in \mathcal{G}_n$ within $O(J_n^{-r})$.
2. Proposition 2: Bracketing entropy bound. We quantify the complexity of the sieve $\mathcal{G}_n$ by upper-bounding its bracketing entropy at scale ε by $O(p_n \log(p_n/\varepsilon))$, where p_n is the sieve's parameter count.
3. Proposition 3: Likelihood ratio bound. We transfer the proximity of g^* and $\bar{g}_n$ to a bounded $\|p_0/p_{\bar{g}_n}\|_\infty$.
4. Proposition 4: Bracketing integral bound. We show the Hellinger bracketing integral is dominated by an explicit envelope $\widetilde{J}_n(\delta) \asymp \sqrt{J_n}\delta\sqrt{\log(J_n/\delta)}$, matching the rate theorem's shape requirements.

Balancing these four bounds at $J_n \asymp n^{1/(2r+1)}$ yields the rate of Theorem 1 for the exact sieve maximiser. The estimator $\hat{g}_n$ is not an exact maximiser but falls short of the approximant's log-likelihood by at most η_n per (C6). To accommodate this, we re-trace the derivation of the rate theorem through (van der Vaart & Wellner, 2023, Thm. 3.4.1). The parent theorem's slack-tolerance hypothesis absorbs η_n, transferring the rate to $\hat{g}_n$.

Rate transfer to the implemented estimator The KAN implementation differs from the theory in four ways. Each either enlarges the sieve, thus the approximant is unchanged and only the entropy count must be updated, or it perturbs the criterion, in which case it is absorbed into the slack of (C6).

- **SiLU residual.** Each edge of the KAN implementation has the form of Eq. (3), a B-spline with a SiLU residual, with trainable weights w^{s} on the spline and w^{b} on the residual. Setting $w^{\mathrm{b}} = 0$ and $w^{\mathrm{s}} = 1$ recovers the pure-spline sieve $\mathcal{G}_n$, so the approximant $\bar{g}_n$ of Proposition 1 lies in the implementation's sieve and its approximation rate transfers. The two weights per edge contribute $O(\log n)$ to the bracketing entropy of Proposition 2 at scale $\varepsilon \asymp \delta_n$, which is dominated by the $O(J_n \log n)$ contribution of the spline coefficients, so the rate exponent of Theorem 1 is preserved.
- **Knot placement.** The sieve $\mathcal{G}_n$ assumes uniform spline knots, while the KAN implementation places them at a blend of uniform positions and sample quantiles of the edge's inputs. Knots remain at least a fixed fraction γ of the uniform spacing apart. Enlarging the sieve to all knot vectors with this minimum gap contains the uniform sieve as a special case, so $\bar{g}_n$ and its approximation rate transfer. The J_n knots per edge become parameters of the enlarged sieve and contribute $O(J_n \log n)$ to the bracketing entropy, the same order as the $J_n + m$ spline coefficients per edge, so the argument of Proposition 2 with the knots as parameters gives the same order. The dependence of the sieve on the knots is controlled by a Lipschitz bound for splines with sufficiently separated knots (Belitser & Serra, 2014, Lem. 2), applied at γ times the uniform spacing. The rate exponent of Theorem 1 is preserved.
- **Penalisation.** Training maximises the penalised criterion $\mathcal{L}_n(g) - \lambda_{\ell_1}\rho(\boldsymbol{\theta})$, with the penalty of Eq. (10) comprising the magnitude term $\rho_{\text{magnitude}}$ and the difference term $\rho_{\text{difference}}$. Let $\bar{\boldsymbol{\theta}}_n$ be the coefficient vector of $\bar{g}_n$. Since $\rho \geq 0$, $\mathcal{L}_n(\hat{g}_n) - \mathcal{L}_n(\bar{g}_n)$ is at least the difference of the penalised criterion at $\hat{g}_n$ and at $\bar{g}_n$, minus $\lambda_{\ell_1}\rho(\bar{\boldsymbol{\theta}}_n)$. The penalty therefore affects Theorem 1 only through η_n in (C6), which it increases by $\lambda_{\ell_1}\rho(\bar{\boldsymbol{\theta}}_n)$. The value $\rho(\bar{\boldsymbol{\theta}}_n)$ is $O(J_n)$, since $\rho_{\text{magnitude}}$ sums $J_n + m$ bounded coefficients per edge and $\rho_{\text{difference}}$ is bounded by the total variation of the approximant's edges, which is $O(1)$. The bound on η_n in (C6) is therefore met whenever $\lambda_{\ell_1} \lesssim n^{-1}\log n$.
- **Quadrature.** The integral $\int_0^{Y_i} e^{g(X_i,s)}\,ds$ in $\mathcal{L}_n$ is evaluated by the trapezoid rule of Eq. (8) on a mesh of K intervals. The resulting criterion differs from $\mathcal{L}_n$ by at most $3e^{M'}[c_1(J_n + 1)]^D/K$ uniformly over the sieve, the integrand's time-Lipschitz constant $e^{M'}[c_1(J_n+1)]^D$ bounding each of three error terms, where $c_1(J_n + 1)$ is the per-layer Lipschitz constant as in Proposition 2. Thus the quadrature error affects Theorem 1 only through η_n in (C6). The bound on η_n in (C6) is therefore met whenever $K \gtrsim n^{1+(D-1)/(2r+1)}$.

In practice we hold λ_{ℓ_1} and K fixed over the finite range of n used in the simulation experiments (Table 2). We directly measure (C5) and (C6) at the fitted estimator for those experiments in Appendix D. To train, the implementation uses Adam optimisation, early stopping and periodic knot re-fitting (Appendix B). Conditions (C5) and (C6) do not restrict training, so long as the final estimator lies in the enlarged sieve and its log-likelihood is within η_n of the approximant's.

5 Related Work

Deep survival analysis Deep learning has been widely applied to survival analysis (Wiegrebe et al., 2024; Chen, 2024), prominently by substituting multi-layer perceptrons (MLPs) into classical parameterisations (Katzman et al., 2018; Kvamme et al., 2019; Lee et al., 2018), but also with neural ODEs (Tang et al., 2022) and transformers (Wang & Sun, 2022). Influential models include DeepSurv (Katzman et al., 2018), DeepHit (Lee et al., 2018), and deep survival machines (DSM) (Nagpal et al., 2022). We refer readers to Chen (2024) and Wiegrebe et al. (2024) for comprehensive reviews. Recent work has also applied KANs to survival analysis (Knottenbelt et al., 2025; Jose et al., 2025; Cheng & Hu, 2025; Mastroleo et al., 2026). A subset of deep learning survival analysis studies provides theoretical analyses of their proposed estimators' convergence and asymptotic normality. These include Deep Extended Hazards (Zhong et al., 2021), Neural Frailty Machines (Wu et al., 2023), and Deep Partially Linear Transformations (Yin et al., 2025). Closest to our parameterisation, Su et al. (2026) model the conditional hazard as a joint function of covariates and time using a class of ReLU networks, establishing a nonasymptotic error bound and functional asymptotic normality.

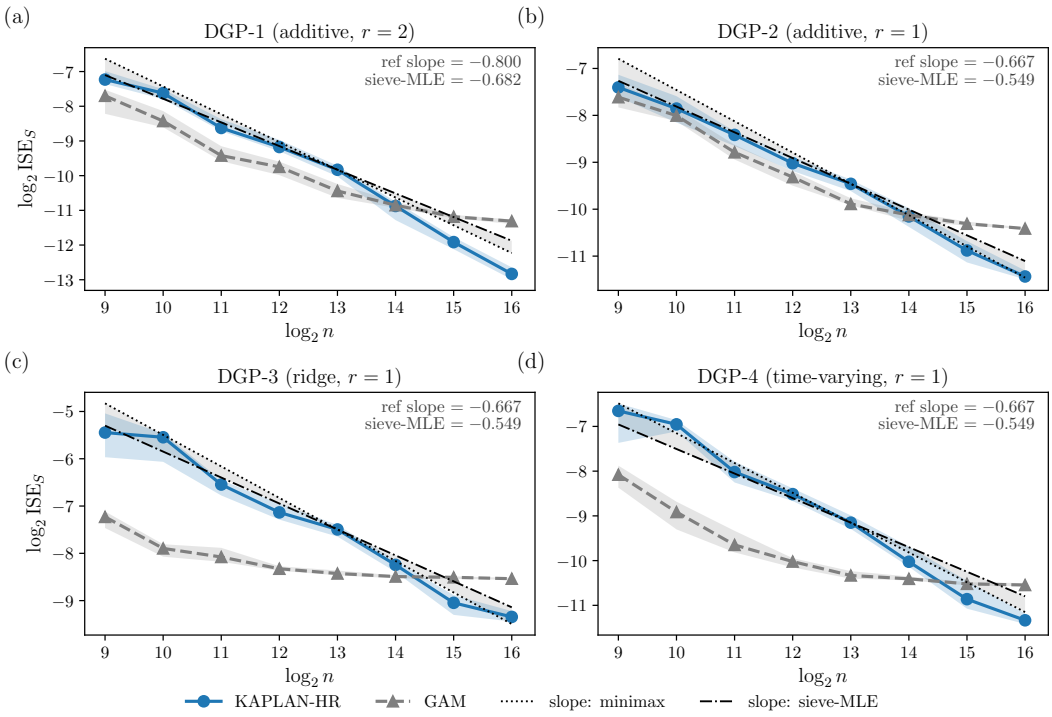


Figure 2: Log-log convergence of ISE_S vs. n for KAPLAN-HR and GAM. Median and IQR over 20 repetitions, compared to the minimax slope $-2r/(2r+1)$ and the Corollary 2 slope.

Kolmogorov-Arnold Network approximation guarantees Liu et al. (2025b)'s initial KAN paper establishes a dimension-free approximation upper bound for B-spline KANs when the target admits a representation as an additive composition of smooth univariate edges. Liu et al. (2025a) formalise this analysis for two-layer KANs, deriving a matching lower bound and thus establishing minimax optimality. Kratsios & Furuya (2025) introduce an augmented KAN architecture with residual connections between layers and show that it is an optimal approximator for Besov targets. Complementary lines of work in KAN theory include comparisons and mutual representation between KANs and MLPs (Shukla et al., 2024; Wang et al., 2025), guarantees on gradient descent optimisation (Gao & Tan, 2025), and generalisation bounds (Zhang & Zhou, 2025; Li et al., 2025).

6 Experiments

6.1 Simulations

We conduct simulations to test the finite-sample validity of the convergence rates in Section 4. We define four data-generating processes (DGPs) whose log-hazards are built from a multiscale univariate function ζ_r with smoothness exponent r. DGP-1 and DGP-2 are additive in $(\mathbf{x}, t)$ and thus representable by a 1-layer KAN, at $r = 2$ and $r = 1$ respectively. DGP-3 composes ζ_1 with a linear combination of x_1 and x_2, requiring at least two KAN layers to represent. DGP-4 multiplies $\zeta_1(x_1)$ by a function of time, so its hazards are not proportional. The DGP formulae and experimental details are given in Appendix C.2.

We generate $R = 20$ independent training sets per DGP at sample sizes $n \in \{2^9, 2^{10}, \ldots, 2^{16}\}$, each with its own test set. We draw i.i.d. covariates $x_1, \ldots, x_5 \sim \mathrm{Uniform}(0, 1)$ (x_3, x_4, x_5 are intentionally left as noise) and generate continuous event times T by inverting the cumulative hazard. Censoring times are drawn independent of T at an overall censoring rate of 30%, including administrative censoring at a fixed horizon.

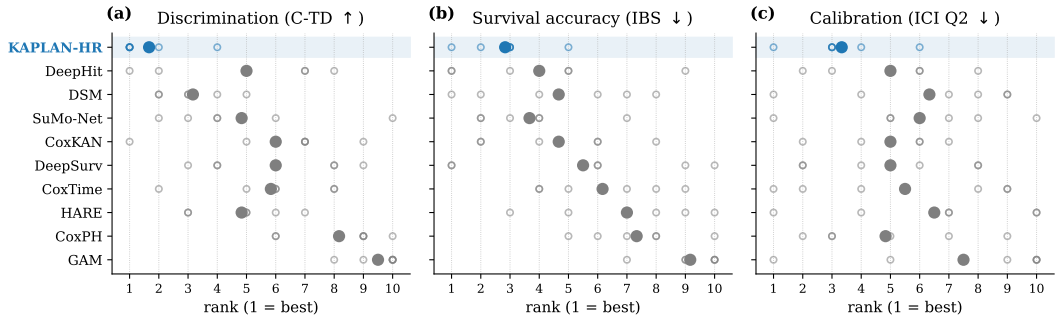


Figure 3: Ranking of KAPLAN-HR vs. baselines across datasets for time-dependent concordance C-TD (↑), IBS (↓), and ICI at the median event time (↓). D-cal is a pass count and is not ranked. Small markers show per-dataset ranks, and large markers show the mean across datasets.

For each draw we measure the integrated squared error (ISE) of the predicted survival:

$$\mathrm{ISE}_S = \|\hat{S} - S^*\|_{L^2}^2 \tag{12}$$

Corollary 2 gives $\mathrm{ISE}_S = O_P(n^{-2r/(2r+1)} \log n)$, the numerator carrying $-2r$ rather than $-r$ because the error is squared. We also fit GAMs as a finite-sample baseline.

Results Figure 2 reports the median $\log_2 \mathrm{ISE}_S$ against $\log_2 n$ on the held-out test set for KAPLAN-HR and GAM. Each panel marks two reference slopes, the minimax slope $-2r/(2r+1)$ and the slope predicted by Corollary 2. Over this range KAPLAN-HR descends at least as fast as the predicted slope on all four DGPs and continues to descend at the largest sample sizes, corroborating the rate analysis in Section 4. On these fits (C6) holds across the grid, and (C5) holds except for a minority on DGP-3 and DGP-4 at the smaller sample sizes (Appendix D). The GAM attains lower error at the smallest sample sizes on all four DGPs and plateaus on all four, including DGP-1 and DGP-2 where it is correctly specified. We note that these DGPs were chosen to stress the convergence rate, and the clinical data benchmarks present a better comparison of the two models' performance in real-world situations. Appendix E reports experiments varying the ambient dimension and the censoring rate. Both perturbations increase the absolute error while the convergence curves continue to descend, except at 90% censoring, where the descent slows on three of the four DGPs.

6.2 Real-World Clinical Datasets

We benchmark KAPLAN-HR on six standard clinical survival analysis datasets, for comparability with prior studies and to test the model's performance on real data, where it is unknown if the log-hazard admits the KAN representation the theory assumes. We evaluate performance in terms of discrimination, via Antolini's C-TD (Antolini et al., 2005); survival accuracy, via the IBS (Graf et al., 1999); and calibration, via the median event time ICI (Austin et al., 2020) and the D-cal statistic (Haider et al., 2020). Performance is compared against nine classical and deep learning baselines.

Appendix C.3 details our experimental protocol. Each dataset is split ten times into training, validation, and testing sets $64:16:20$. We report the mean and standard deviation across splits for C-TD, IBS, and ICI, and the number of splits passing the D-cal test at $p > 0.05$. We tune KAPLAN-HR and all tunable baseline models by hyperparameter search to minimise the validation IBS.

Results Table 1 reports KAPLAN-HR's scores against the highest-scoring baseline per dataset, and Fig. 3 shows KAPLAN-HR's mean rank on each metric against the baselines. KAPLAN-HR attains the best C-TD on 4 of 6 datasets, the best IBS on 1 of 6 datasets, and the best ICI on 1 of 6 datasets. KAPLAN-HR attains the best mean rank of the ten models on all metrics, 1.7 on discrimination, 2.8 on survival accuracy and 3.3 on calibration. KAPLAN-HR passes the D-cal test on all ten splits of every dataset except SUPPORT, where it passes 5 of 10. We check comparisons' significance via two-sided Wilcoxon signed-rank test on the paired per-split scores, Holm-Bonferroni corrected across the six datasets (table's column p). No per-dataset comparison reaches significance. KAPLAN-HR is therefore competitive with the highest-scoring baseline per dataset on all three metrics. Appendix F

presents the complete benchmark results. Appendix E.3 reports an ablation removing the SiLU residual, finding that removing it degrades calibration.

Table 1: KAPLAN-HR vs. the best-scoring baseline per dataset, for discrimination (C-TD, ↑), survival accuracy (IBS, ↓) and calibration (ICI at the median event time, ↓). p: Wilcoxon signed-rank, Holm-corrected across datasets within metric.

	KAPLAN-HR	**Best-Scoring Baseline**			
Dataset	Score	Model	Score	Δ	p
		C-TD (↑)			
METABRIC	0.6690 ± 0.0162	DeepHit	0.6736 ± 0.0195	-0.0046	1.0000
MIMIC-III	0.8136 ± 0.0101	DeepHit	0.8084 ± 0.0119	$+0.0052$	0.2227
RotGBSG	0.6769 ± 0.0153	DSM	0.6750 ± 0.0144	$+0.0019$	1.0000
FLCHAIN	0.7935 ± 0.0129	CoxKAN	0.7966 ± 0.0123	-0.0031	0.2441
NWTCO	0.7223 ± 0.0220	SuMo-Net	0.7186 ± 0.0154	$+0.0037$	1.0000
SUPPORT	0.6665 ± 0.0094	DSM	0.6654 ± 0.0121	$+0.0010$	1.0000
		IBS (↓)			
METABRIC	0.1603 ± 0.0056	DeepHit	0.1595 ± 0.0066	$+0.0008$	1.0000
MIMIC-III	0.1742 ± 0.0241	SuMo-Net	0.1810 ± 0.0362	-0.0069	1.0000
RotGBSG	0.1800 ± 0.0077	DeepSurv	0.1798 ± 0.0076	$+0.0002$	1.0000
FLCHAIN	0.0983 ± 0.0025	DSM	0.0981 ± 0.0028	$+0.0003$	1.0000
NWTCO	0.1024 ± 0.0039	DeepHit	0.0959 ± 0.0101	$+0.0065$	0.1641
SUPPORT	0.1837 ± 0.0064	DeepSurv	0.1832 ± 0.0062	$+0.0005$	1.0000
		ICI (↓)			
METABRIC	0.0231 ± 0.0059	CoxTime	0.0233 ± 0.0080	-0.0002	1.0000
MIMIC-III	0.0086 ± 0.0022	HARE	0.0079 ± 0.0020	$+0.0007$	1.0000
RotGBSG	0.0280 ± 0.0068	CoxTime	0.0279 ± 0.0088	$+0.0001$	1.0000
FLCHAIN	0.0186 ± 0.0056	DSM	0.0145 ± 0.0061	$+0.0042$	1.0000
NWTCO	0.0334 ± 0.0084	GAM	0.0154 ± 0.0154	$+0.0179$	0.2930
SUPPORT	0.0251 ± 0.0059	SuMo-Net	0.0215 ± 0.0040	$+0.0036$	1.0000

7 Conclusion

We introduced KAPLAN-HR, a survival modelling approach that generalises established spline-based models with KANs, able to capture covariate interactions and time-varying effects. Our theoretical analysis shows that the KAPLAN-HR estimator attains the rate $n^{-r/(2r+1)}$, up to a logarithmic factor, when the true log-hazard follows the KAN structural assumption. The exponent depends only on the smoothness r of the representation's univariate components, not on the covariate dimension, so the estimator mitigates the curse of dimensionality. To our knowledge, this is the first formal convergence guarantee for a KAN-based survival model. Simulation studies corroborate the theory's predicted convergence rate. Together with competitive performance on six clinical benchmarks, these results support that the KAN hazard parameterisation is well-suited for survival analysis.

Limitations Our convergence results are asymptotic and do not characterise finite-sample behaviour. Increasing the ambient dimension in simulations (Appendix E.1) illustrates this, as the added noise raises the absolute error over the tested n range. The results also rest on a structural assumption about the true log-hazard, consistent with prior KAN theory (Liu et al., 2025b) but not guaranteed to hold in all situations. In the simulations, a correctly specified GAM attains lower error up to large sample sizes on all four DGPs. However, these DGPs are built to hold the approximation error at the class rate rather than to represent typical hazard functions. Training and prediction evaluate the network over sample and time pairs, so cost grows with the time-grid resolution as well as the sample size, which we partly mitigate by factoring the first KAN layer (Eq. (47)). Finally, benchmarking on medium-scale clinical datasets has known shortcomings (Wu et al., 2023), and no per-dataset comparison reaches significance in either direction, so these datasets support competitive performance rather than an advantage.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks: formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, and implementing methods. We have not used generative AI tools for the following tasks: generating synthetic data sets, proposing or refining hypotheses, cleaning and reformatting datasets, or interpreting results. The remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools for the following tasks: suggesting experimental parameters, creating or editing software code, drafting parts of a research paper, summarising or analysing existing literature, editing a research paper to improve readability, identifying relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors, and all LLM-assisted proofs were checked line by line by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions behind our theoretical results are stated in full as conditions (C1)-(C6) in Section 4, and complete proofs of Theorem 1 and the four supporting propositions are given in Appendix A. The model, its likelihood and its numerical integration scheme are specified in Section 3, and the training procedure is given as pseudocode in Algorithm 1. The implementation, including the version of the `pykan` library used, is described in Appendix B, with its configuration listed in Table 2. For the simulation study, Appendix C.2 gives the four data generating processes, the sample sizes and the replication count. For the benchmark, every dataset is publicly available and cited at first use, and Appendix C.3 gives the preprocessing and the tuning protocol, with the hyperparameters selected for KAPLAN-HR reported per dataset in Table 10. The source code for the model and experiments is submitted as supplementary material.

REFERENCES

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis*. Studies in Economic Theory. Springer Berlin, Heidelberg, 1999. doi: 10.1007/978-3-662-03961-8. URL https://doi.org/10.1007/978-3-662-03961-8.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, December 2005. ISSN 0277-6715. doi: 10.1002/sim.2427. URL https://doi.org/10.1002/sim.2427.

Peter C. Austin, Frank E. Harrell Jr, and David van Klaveren. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Statistics in Medicine*, 39(21):2714–2742, September 2020. ISSN 0277-6715. doi: 10.1002/sim.8570. URL https://doi.org/10.1002/sim.8570.

Eduard Belitser and Paulo Serra. Adaptive Priors Based on Splines with Random Knots. *Bayesian Analysis*, 9(4):859–882, December 2014. doi: 10.1214/14-BA879. URL https://doi.org/10.1214/14-BA879.

N. E. Breslow and N. Chatterjee. Design and Analysis of Two-Phase Studies with Binary Outcome Applied to Wilms Tumour Prognosis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 48(4):457–468, December 1999. ISSN 0035-9254. doi: 10.1111/1467-9876.00165. URL https://doi.org/10.1111/1467-9876.00165.

George H Chen. An Introduction to Deep Survival Analysis Models for Predicting Time-to-Event Outcomes, 2024. URL https://doi.org/10.48550/arXiv.2410.01086.

Jiaxiang Cheng and Guoqiang Hu. Extending cox proportional hazards model with symbolic non-linear log-risk functions for survival analysis, 2025. URL https://doi.org/10.48550/arXiv.2504.04353. arXiv: 2504.04353 [cs.LG].

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246. URL http://www.jstor.org/stable/2985181.

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, June 2012. ISSN 1476-4687. doi: 10.1038/nature10983. URL https://doi.org/10.1038/nature10983.

Cameron Davidson-Pilon. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019. doi: 10.21105/joss.01317. URL https://doi.org/10.21105/joss.01317.

Angela Dispenzieri, Jerry A. Katzmann, Robert A. Kyle, Dirk R. Larson, Terry M. Therneau, Colin L. Colby, Raynell J. Clark, Graham P. Mead, Shaji Kumar, L. Joseph Melton, III, and S. Vincent Rajkumar. Use of Nonclonal Serum Immunoglobulin Free Light Chains to Predict Overall Survival in the General Population. *Mayo Clinic Proceedings*, 87(6):517–523, June 2012. ISSN 0025-6196. doi: 10.1016/j.mayocp.2012.03.009. URL https://doi.org/10.1016/j.mayocp.2012.03.009.

John A. Foekens, Harry A. Peters, Maxime P. Look, Henk Portengen, Manfred Schmitt, Michael D. Kramer, Nils Brünner, Fritz Jänicke, Marion E. Meijer-van Gelder, Sonja C. Henzen-Logmans, Wim L. J. van Putten, and Jan G. M. Klijn. The Urokinase System of Plasminogen Activation and Prognosis in 2780 Breast Cancer Patients. *Cancer Research*, 60(3):636–643, February 2000. ISSN 0008-5472.

Yihang Gao and Vincent Y. F. Tan. On the convergence of (stochastic) gradient descent for kolmogorov–arnold networks. *IEEE Transactions on Information Theory*, 71(9):7270–7291, 2025. doi: 10.1109/TIT.2025.3588401.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5. URL https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(1), January 2020. ISSN 1532-4435.

Trevor Hastie and Robert Tibshirani. Generalized additive models for medical research. *Statistical Methods in Medical Research*, 4(3):187–196, 1995. doi: 10.1177/096228029500400302. URL https://doi.org/10.1177/096228029500400302.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III clinical database. *PhysioNet*, September 2016. doi: 10.13026/C2XW26. URL https://doi.org/10.13026/C2XW26.

Mebin Jose, Jisha Francis, and Sudheesh Kumar Kattumannil. KAN-AFT: An interpretable nonlinear survival model integrating kolmogorov-arnold networks with accelerated failure time analysis, 2025. URL https://doi.org/10.48550/arXiv.2512.20305.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, February 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1. URL https://doi.org/10.1186/s12874-018-0482-1.

William A. Knaus, Frank E. Harrell, Joanne Lynn, Lee Goldman, Russell S. Phillips, Alfred F. Connors, Neal V. Dawson, William J. Fulkerson, Robert M. Califf, Norman Desbiens, Peter Layde, Robert K. Oye, Paul E. Bellamy, Rosemarie B. Hakim, and Douglas P. Wagner.

The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults. *Annals of Internal Medicine*, 122(3):191–203, February 1995. ISSN 0003-4819. doi: 10.7326/0003-4819-122-3-199502010-00007. URL https://doi.org/10.7326/0003-4819-122-3-199502010-00007.

William Knottenbelt, William McGough, Rebecca Wray, Woody Zhidong Zhang, Jiashuai Liu, Ines Prata Machado, Zeyu Gao, and Mireia Crispin-Ortuzar. CoxKAN: Kolmogorov-Arnold networks for interpretable, high-performance survival analysis. *Bioinformatics*, 41(8):btaf413, August 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf413. URL https://doi.org/10.1093/bioinformatics/btaf413.

Charles Kooperberg, Charles J. Stone, and Young K. Truong. Hazard Regression. *Journal of the American Statistical Association*, 90(429):78–94, March 1995. ISSN 0162-1459. doi: 10.1080/01621459.1995.10476491. URL https://doi.org/10.1080/01621459.1995.10476491.

Anastasis Kratsios and Takashi Furuya. Approximation rates in besov norms and sample-complexity of kolmogorov-arnold networks with residual connections, 2025. URL https://doi.org/10.48550/arXiv.2504.15110.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, October 2021. ISSN 1572-9249. doi: 10.1007/s10985-021-09532-6. URL https://doi.org/10.1007/s10985-021-09532-6.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. doi: 10.48550/arXiv.1907.00825. URL http://jmlr.org/papers/v20/18-424.html.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11842. URL https://doi.org/10.1609/aaai.v32i1.11842.

Pengqi Li, Lizhong Ding, Jiarun Fu, ChunhuiZhang, Guoren Wang, and Ye Yuan. Generalization bounds for Kolmogorov-Arnold Networks (KANs) and enhanced KANs with lower Lipschitz complexity. In *The thirty-ninth annual conference on neural information processing systems*, 2025. URL https://openreview.net/forum?id=b387eWFV3V.

Wei Liu, Eleni Chatzi, and Zhilu Lai. On the Rate of Convergence of Kolmogorov-Arnold Network Regression Estimators, 2025a. URL https://doi.org/10.48550/arXiv.2509.19830.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov-arnold networks. In *Proceedings of the 2025 International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Ozo7qJ5vZi.

Marina Mastroleo, Alberto Archetti, Federico Mastroleo, and Matteo Matteucci. SurvKAN: a fully parametric survival model based on kolmogorov-arnold networks, 2026. URL https://doi.org/10.48550/arXiv.2602.02179.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, August 2021. ISSN 2168-2208. doi: 10.1109/JBHI.2021.3052441.

Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. auton-survival: an Open-Source Package for Regression, Counterfactual Estimation, Evaluation and Phenotyping with Censored Time-to-Event Data, 2022. URL https://doi.org/10.48550/arXiv.2204.07276.

Amir Noorizadegan, Sifan Wang, Leevan Ling, and Juan P. Dominguez-Morales. A Practitioner's Guide to Kolmogorov-Arnold Networks, 2026. URL https://doi.org/10.48550/arXiv.2510.25781.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112–134, July 2018. ISSN 1532-0464. doi: 10.1016/j.jbi.2018.04.007. URL https://www.sciencedirect.com/science/article/pii/S1532046418300716.

Sebastian Pölsterl. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL http://jmlr.org/papers/v21/20-729.html.

Shi-ang Qi, Weijie Sun, and Russell Greiner. SurvivalEVAL: a comprehensive open-source python package for evaluating individual survival distributions. *Proceedings of the AAAI Symposium Series*, 2:453–457, January 2024. doi: 10.1609/aaaiss.v2i1.27713.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of the 25th international conference on artificial intelligence and statistics*, volume 151 of *Proceedings of machine learning research*, pp. 1190–1205. PMLR, March 2022. doi: 10.48550/arXiv.2103.14755. URL https://proceedings.mlr.press/v151/rindt22a.html.

M Schumacher, G Bastert, H Bojar, K Hübner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, R L Neumann, and H F Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *Journal of Clinical Oncology*, 12(10):2086–2093, October 1994. ISSN 0732-183X. doi: 10.1200/JCO.1994.12.10.2086. URL https://doi.org/10.1200/JCO.1994.12.10.2086.

Larry Schumaker. *Spline Functions: Basic Theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 3rd ed edition, January 2007. ISBN 978-0-521-70512-7. URL https://research.ebsco.com/linkprocessor/plink?id=7b136a48-6a99-3fe0-a0a8-2b0fc23984fc.

Xiaotong Shen and Wing Hung Wong. Convergence Rate of Sieve Estimates. *The Annals of Statistics*, 22(2):580–615, June 1994. doi: 10.1214/aos/1176325486. URL https://doi.org/10.1214/aos/1176325486.

Khemraj Shukla, Juan Diego Toscano, Zhicheng Wang, Zongren Zou, and George Em Karniadakis. A comprehensive and FAIR comparison between MLP and KAN representations for differential equations and operator networks. *Computer Methods in Applied Mechanics and Engineering*, 431:117290, November 2024. ISSN 0045-7825. doi: 10.1016/j.cma.2024.117290. URL https://www.sciencedirect.com/science/article/pii/S0045782524005462.

Wen Su, Qiang Wu, Kin-Yat Liu, Guosheng Yin, Jian Huang, and Xingqiu Zhao. Nonparametric inference for censored data using deep neural networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, April 2026. ISSN 1369-7412. doi: 10.1093/jrsssb/qkag060. URL https://doi.org/10.1093/jrsssb/qkag060.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, March 2015. doi: 10.1371/journal.pmed.1001779. URL https://doi.org/10.1371/journal.pmed.1001779.

Weijing Tang, Jiaqi Ma, Qiaozhu Mei, and Ji Zhu. SODEN: a scalable continuous-time survival model through ordinary differential equation networks. *Journal of Machine Learning Research*, 23 (34):1–29, 2022. doi: 10.48550/arXiv.2008.08637. URL http://jmlr.org/papers/v23/20-900.html.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, New York, 2 edition, 2023. ISBN 978-3-031-29038-1. doi: 10.1007/978-1-4757-2545-2. URL https://doi.org/10.1007/978-3-031-29040-4.

Yixuan Wang, Jonathan W. Siegel, Ziming Liu, and Thomas Y. Hou. On the expressiveness and spectral bias of KANs. In *The thirteenth international conference on learning representations*, 2025. URL https://openreview.net/forum?id=ydlDRUuGm9.

Zifeng Wang and Jimeng Sun. SurvTRACE: transformers for survival analysis with competing events. In *Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics*, Bcb '22, Northbrook, Illinois, 2022. Association for Computing Machinery. ISBN 978-1-4503-9386-7. doi: 10.1145/3535508.3545521. URL https://doi.org/10.1145/3535508.3545521.

Weights and Biases. Sweeps: Tune hyperparameters, 2026. URL https://docs.wandb.ai/guides/sweeps/.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, February 2024. ISSN 1573-7462. doi: 10.1007/s10462-023-10681-3. URL https://doi.org/10.1007/s10462-023-10681-3.

Simon N. Wood. Generalized additive models. *Annual Review of Statistics and Its Application*, 12(Volume 12, 2025):497–526, 2025. ISSN 2326-831X. doi: https://doi.org/10.1146/annurev-statistics-112723-034249. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-112723-034249.

Simon N. Wood, Natalya Pya, and Benjamin Säfken. Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, 111(516):1548–1563, October 2016. ISSN 0162-1459. doi: 10.1080/01621459.2016.1180986. URL https://doi.org/10.1080/01621459.2016.1180986.

Ruofan Wu, Jiawei Qiao, Mingzhe Wu, Wen Yu, Ming Zheng, Tengfei LIU, Tianyi Zhang, and Weiqiang Wang. Neural Frailty Machine: Beyond proportional hazard assumption in neural survival regressions. In *Thirty-seventh conference on neural information processing systems*, 2023. URL https://openreview.net/forum?id=3Fc9gnR0fa.

Junkai Yin, Yue Zhang, and Zhangsheng Yu. Deep partially linear transformation model for right-censored survival data. *Biometrics*, 81(4):ujaf126, December 2025. ISSN 0006-341X. doi: 10.1093/biomtc/ujaf126. URL https://doi.org/10.1093/biomtc/ujaf126.

Xianyang Zhang and Huijuan Zhou. Generalization bounds and model complexity for kolmogorov–arnold networks. In *The thirteenth international conference on learning representations*, 2025. URL https://openreview.net/forum?id=q5zMyAUhGx.

Qixian Zhong, Jonas Mueller, and Jane-Ling Wang. Deep extended hazard models for survival analysis. In *Proceedings of the 35th international conference on neural information processing systems*, Nips '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 978-1-7138-4539-3. doi: 10.5555/3540261.3541419. URL https://doi.org/10.5555/3540261.3541419.

A Proofs

A.1 Notation

We introduce some notation that features frequently in the proof: For a function class $\mathcal{F}$, denote the covering number $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ and the bracketing number $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ at scale ε with respect to norm $\|\cdot\|$. Denote also the bracketing integral:

$$\mathcal{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)}\, d\varepsilon. \tag{13}$$

We use the notation $a \lesssim b$ to mean $a \leq Cb$ for constant $C > 0$, and $a \asymp b$ to mean $c_1 b \leq a \leq c_2 b$.

A.2 SETUP

As defined in Section 4, let the class $\mathcal{F}_r^{\mathrm{KAN}}$ with architecture parameters (d_0, D, Q, M_0, M) with smoothness $r \in \mathbb{N}_{\geq 1}$, spline order $m \geq \max(2, r)$, bounding constants $M_0 \in [1, \infty)$, $M_0 < M$, and depth $D \geq 1$. Let also the sieve $\mathcal{G}_n$ with knot count $J_n \asymp n^{1/(2r+1)}$ and sup-norm cap $M' := 2M_0$. We assume conditions (C1)-(C6) and the censoring law f_C, α, S_C from (C3). Finally, we use the covariate-averaged Hellinger metric d_H from Eq. (11).

$\mathcal{X} := [0,1]^d$ is the covariate domain, and $\mathcal{Z} := \mathcal{X} \times [0,1] = [0,1]^{d_0}$ is the model input domain (covariates and time). Each KAN layer $\Phi_\ell, \ell \in \{1, \ldots, D\}$ has domain $\mathcal{Z}_\ell$ with $\mathcal{Z}_1 = \mathcal{Z}$.

The sieve's total parameter count is $p_n = E(J_n + m)$ where $E := \sum_{\ell=1}^{D} d_{\ell-1} d_\ell$ is the total edge count. Each sieve edge is an order-m B-spline on J_n uniform interior knots (m-fold boundary knots) parameterised by a coefficient vector in $\mathbb{R}^{J_n+m}$ with $\|\cdot\|_\infty \leq C_\theta$. Denote the corresponding coefficient box $\Theta_n := [-C_\theta, C_\theta]^{p_n}$.

Define the dominating measure on $\mathcal{Z}$:

$$\mu = P_X \otimes \mathrm{Leb}_{[0,1]} \otimes (\delta_0 + \delta_1) \; + \; P_X \otimes \delta_1 \otimes \delta_0$$

where the left term corresponds to events and censoring before $Y = 1$, and the right term corresponds to administrative censoring. Under (C1) and (C3), the joint density p_g of (X, Y, Δ) with respect to μ is

$$\begin{cases} p_g(\mathbf{x}, y, 1) = S_C(y \mid \mathbf{x})\, e^{g(\mathbf{x},y)}\, S_g(y \mid \mathbf{x}) & \Delta = 1,\ y \in [0,1) \\ p_g(\mathbf{x}, y, 0) = f_C(y \mid \mathbf{x})\, S_g(y \mid \mathbf{x}) & \Delta = 0,\ y \in [0,1) \\ p_g(\mathbf{x}, 1, 0) = \alpha(\mathbf{x})\, S_g(1 \mid \mathbf{x}) & \Delta = 0,\ y = 1 \\ p_g(\mathbf{x}, 1, 1) := 0 & \Delta = 1,\ y = 1 \end{cases} \tag{14}$$

where S_g denotes the survival function induced by log-hazard g.

The empirical log-likelihood decomposes into the g-dependent empirical criterion $\mathcal{L}_n(g)$ and the g-invariant remainder R_n which depends only on the censoring mechanism and is finite.

$$\frac{1}{n} \sum_i \log p_g(\mathbf{x}_i, y_i, \Delta_i) = \mathcal{L}_n(g) + R_n \tag{15}$$

The joint Hellinger distance on densities w.r.t. μ is

$$H(p, q) = \left(\int (\sqrt{p} - \sqrt{q})^2 \, d\mu \right)^{1/2}, \qquad H^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2 \, d\mu. \tag{16}$$

The covariate-averaged Hellinger d_H from Eq. (11) relates to H via $\mu = P_X \otimes \nu$ (with $\nu = \mathrm{Leb}_{[0,1]} \otimes (\delta_0 + \delta_1) + \delta_1 \otimes \delta_0$): both densities $p_g, p_{g'}$ have covariate marginal P_X, so:

$$d_H^2(g, g') = \mathbb{E}_{X \sim P_X} \left[H^2(\mathbb{P}_{g,X} \parallel \mathbb{P}_{g',X}) \right] = \int (\sqrt{p_g} - \sqrt{p_{g'}})^2 \, d\mu = H^2(p_g, p_{g'}). \tag{17}$$

Hence $H(p_g, p_{g'}) = d_H(g, g')$ throughout the proof.

A.3 PROPOSITIONS

The following propositions are used in the main theorem proof. The proposition proofs are given in Appendix A.5.

Proposition 1 (Approximation bound)**.** *There exist constants $J_0, C_{\mathrm{app}} \in (0, \infty)$ independent of n such that, for every $g^* \in \mathcal{F}_r^{\mathrm{KAN}}$ and every $J_n \geq J_0$, there exists a deterministic sieve approximant $\bar{g}_n \in \mathcal{G}_n$ with:*

$$\|\bar{g}_n - g^*\|_{\infty, \mathcal{Z}} \leq C_{\mathrm{app}}\, J_n^{-r}. \tag{18}$$

Proposition 2 (Bracketing entropy bound)**.** *There exist constants $p_{\min}, C_{\mathrm{ent}} \in (0, \infty)$ independent of n such that for every $\varepsilon \in (0, 2M']$ and every n for which $p_n \geq p_{\min}$, the bracketing entropy:*

$$\log N_{[]}(\varepsilon, \mathcal{G}_n, \|\cdot\|_{\infty, \mathcal{Z}}) \leq C_{\mathrm{ent}} p_n \log(p_n/\varepsilon) \tag{19}$$

Proposition 3 (Likelihood ratio bound)**.** *Let constant $B \in [0, \infty)$. For every pair of measurable log-hazards $g_1, g_2 : \mathcal{Z} \to \mathbb{R}$ with $\|g_i\|_{\infty,\mathcal{Z}} \leq B$ ($i = 1, 2$):*

$$\left|\log \frac{p_{g_1}(\mathbf{x}, y, \Delta)}{p_{g_2}(\mathbf{x}, y, \Delta)}\right| \leq (1 + e^B)\|g_1 - g_2\|_{\infty,\mathcal{Z}} \tag{20}$$

for every $(\mathbf{x}, y, \Delta)$ in the μ support.

Proposition 4 (Bracketing integral bound)**.** *There exist constants $K, J^* \in (0, \infty)$ independent of n such that for every n with $J_n \geq J^*$, the function:*

$$\widetilde{J}_n(\delta) := K\sqrt{J_n}\delta\sqrt{\log(J_n/\delta) + 1}, \qquad \delta \in (0, \sqrt{2}] \tag{21}$$

satisfies:

(i) $\widetilde{J}_n$ is strictly increasing

(ii) $\widetilde{J}_n(\delta)/\delta^{3/2}$ is strictly decreasing

(iii) $\mathcal{J}_{[]}(\delta, \mathcal{P}_n, H) \leq \widetilde{J}_n(\delta)$

A.4 Proof of Theorem 1

Let the sieve density class $\mathcal{P}_n := \{p_g : g \in \mathcal{G}_n\}$. For notational ease denote $\hat{p}_n = p_{\hat{g}_n}$, and $p_0 = p_{g^*}$. Theorem 1 is proven in two parts. The first part is invoking (van der Vaart & Wellner, 2023, Thm. 3.4.12): For the exact sieve maximum-likelihood estimator (MLE) over $\mathcal{P}_n$, $H(\cdot, p_0) = O_P(\delta_n)$ conditional on:

(i) *Bounded ratio.* A sieve approximant density $p_n^\circ \in \mathcal{P}_n$ and constant $M_{\text{ratio}} \in [1, \infty)$ such that $\|p_0/p_n^\circ\|_\infty \leq M_{\text{ratio}}$.

(ii) *Bounded bracketing integral* ((van der Vaart & Wellner, 2023, Eq. 3.4.11) clause 1). $\mathcal{J}_{[]}(\delta_n, \mathcal{P}_n, H) \lesssim \sqrt{n}\,\delta_n^2$ at some $\delta_n \in (0, \sqrt{2}\,]$.

(iii) *Discrepancy threshold* ((van der Vaart & Wellner, 2023, Eq. 3.4.11) clause 2). $\delta_n \gtrsim H(p_n^\circ, p_0)$.

Per (C6), however, $\hat{p}_n$ is not the exact MLE but must come within η_n of the log-likelihood of the sieve approximant $\bar{g}_n$. Therefore the second part of the proof is to extend Thm. 3.4.12 to accommodate this slack by carrying it through the theorem's original derivation from (van der Vaart & Wellner, 2023, Thm. 3.4.1) and (van der Vaart & Wellner, 2023, Lem. 3.4.10).

Let measurable selector $\hat{g}_n \in \mathcal{G}_n$ such that $H(\hat{p}_n, p_0)$ is Borel measurable (given by Lemma 4). This lets us collapse Thm. 3.4.1's O_P^* bound to O_P. Constrain $J_n \geq \max(J_0, J^*) =: J^*_{\text{rate}}$ where J_0 is the minimum J_n from Proposition 1 and J^* is the same from Proposition 4.

Step 1 - Sieve approximant Given target $g^* \in \mathcal{F}_r^{\text{KAN}}$ define the sieve approximant $\bar{g}_n \in \mathcal{G}_n$ such that $\|\bar{g}_n - g^*\|_{\infty,\mathcal{Z}} \leq C_{\text{app}} J_n^{-r}$ via Proposition 1. Denote the sieve approximant density $p_n^\circ = p_{\bar{g}_n} \in \mathcal{P}_n$. Both g^* and $\bar{g}_n$ are sup-norm bounded by M' from their definition. Invoke Lemma 3 at $B = M'$:

$$H^2(p_n^\circ, p_0) \leq C_h(M')\,\|\bar{g}_n - g^*\|^2_{\infty,\mathcal{Z}} \leq C_h(M')\,C_{\text{app}}^2\,J_n^{-2r} \tag{22}$$

Where $C_h(\cdot)$ is a constant defined in Lemma 3. Setting $\delta_n^{(a)} = \sqrt{C_h(M')}\,C_{\text{app}}\,J_n^{-r}$ gives $H(p_n^\circ, p_0) \leq \delta_n^{(a)}$.

Step 2 - Bounded ratio Proposition 3 at $B = M'$ gives:

$$\left|\log \frac{p_0}{p_n^\circ}\right| \leq (1 + e^{M'})\|\bar{g}_n - g^*\|_{\infty,\mathcal{Z}} \leq (1 + e^{M'})\,C_{\text{app}}\,J_n^{-r}$$

on the μ support. Exponentiate and monotonise via $J_n \geq J^*_{\text{rate}}$:

$$\|p_0/p_n^\circ\|_\infty \leq M_{\text{ratio}} := e^{\alpha_0}, \qquad \alpha_0 = (1 + e^{M'})\,C_{\text{app}}\,(J^*_{\text{rate}})^{-r} \tag{23}$$

Step 3 - Bracketing integral For $\delta \in (0, \sqrt{2}\,]$, apply Proposition 4:

$$\mathcal{J}_{[]}(\delta, \mathcal{P}_n, H) \le \widetilde{J}_n(\delta) := K\sqrt{J_n}\delta\sqrt{\log(J_n/\delta) + 1}, \tag{24}$$

with $\widetilde{J}_n$ strictly increasing on $(0, \sqrt{2}\,]$ and $\widetilde{J}_n(\delta)/\delta^{3/2}$ strictly decreasing.

Step 4 - Rate balance By the definition of $\mathcal{F}_r^{\mathrm{KAN}}$, $J_n \asymp n^{1/(2r+1)}$. By (C6), $\eta_n = O(n^{-2r/(2r+1)}\log n)$. Set $\delta_n = C_\delta n^{-r/(2r+1)}\sqrt{\log n}$ with C_δ large enough (depending on M_{ratio}, $C_h(M')$, C_{app}, K, and the (a) asymptotic constants) that, for $n \ge n_*$, the following conditions hold:

1. $\delta_n \ge 64\, M_{\mathrm{ratio}}\, H(p_n^\circ, p_0)$. By equation 22:
$$H(p_n^\circ, p_0) \lesssim J_n^{-r} \lesssim n^{-r/(2r+1)} \tag{25}$$
2. $2\sqrt{M_{\mathrm{ratio}}}\,\widetilde{J}_n(\delta_n) \le \sqrt{n}\delta_n^2$. By equation 24 with $J_n \lesssim n^{1/(2r+1)}$ and $\log(J_n/\delta_n) + 1 \le 2\log n$:
$$\widetilde{J}_n(\delta_n) \lesssim n^{1/(2(2r+1))}\sqrt{\log n}\delta_n \tag{26}$$
matching $\sqrt{n}\delta_n^2 = C_\delta n^{1/(2(2r+1))}\sqrt{\log n}\delta_n$.
3. $$\eta_n \le \delta_n^2 : \eta_n \lesssim n^{-2r/(2r+1)}\log n, \qquad \delta_n^2 = C_\delta^2\, n^{-2r/(2r+1)}\log n \tag{27}$$

Step 5 - Thm. 3.4.12 verification. Steps 1-4 supply the required conditions: (i) bounded ratio by Step 2; (ii) bracketing integral by Step 3 and Step 4 condition 2 ($\widetilde{J}_n(\delta_n) \le \sqrt{n}\,\delta_n^2$ via $M_{\mathrm{ratio}} \ge 1$, hence $\mathcal{J}_{[]}(\delta_n, \mathcal{P}_n, H) \le \sqrt{n}\delta_n^2$); and (iii) discrepancy threshold by Step 4 condition 1 ($\delta_n \ge 64\, M_{\mathrm{ratio}}\, H(p_n^\circ, p_0) \ge H(p_n^\circ, p_0)$).

Step 6 - Rate transfer Thm. 3.4.12 is derived from (van der Vaart & Wellner, 2023, Thm. 3.4.1) via (van der Vaart & Wellner, 2023, Lem. 3.4.10), which supplies the latter theorem's curvature and modulus hypotheses from our bounded ratio (Proposition 3) and bracketing integral (Proposition 4). The comparison-discrepancy hypothesis is made trivial by the matched choice $\theta_n = \theta_{n,0} = p_n^\circ$. The only modification to accommodate the slack of (C6) is to verify Thm. 3.4.1's slack tolerance condition.

Thm. 3.4.1 includes a slack tolerance that allows the estimator to fall short of the comparison-point value $\mathbb{M}_n(\theta_n)$ by $O_P(\delta_n^2)$:

$$\mathbb{M}_n(\hat{\theta}_n) \ge \mathbb{M}_n(\theta_n) - O_P(\delta_n^2)$$

where $\mathbb{M}_n(p) = \mathbb{P}_n m_{n,p}$ for $m_{n,p} = \log\frac{p+p_n^\circ}{2p_n^\circ}$, $\hat{\theta}_n = \hat{p}_n$, $\theta_n = p_n^\circ$.

Concavity of $\log$ gives $m_{n,\hat{p}_n} \ge \frac{1}{2}\log(\hat{p}_n/p_n^\circ)$. Eq. (15) and assumption (C6) yield:

$$\mathbb{P}_n \log(\hat{p}_n/p_n^\circ) = \mathcal{L}_n(\hat{g}_n) - \mathcal{L}_n(\bar{g}_n) \ge -\eta_n$$

Hence $\mathbb{P}_n m_{n,\hat{p}_n} \ge -\eta_n/2 \ge -\delta_n^2/2$ via Eq. (27). Note $\mathbb{M}_n(\theta_n) = \mathbb{P}_n m_{n,p_n^\circ} \equiv 0$ since $m_{n,p_n^\circ} = \log 1 = 0$. Thus we have Thm. 3.4.1's slack tolerance at scale $\delta_n^2/2$.

Thm. 3.4.1 yields $H(\hat{p}_n, p_n^\circ) = O_P^*(\delta_n)$. Triangle inequality $H(\hat{p}_n, p_0) \le H(\hat{p}_n, p_n^\circ) + H(p_n^\circ, p_0)$ with Eq. (25) and collapsing O_P^* to O_P via Lemma 4 as mentioned:

$$H(\hat{p}_n, p_0) = O_P(\delta_n) = O_P\Big(n^{-r/(2r+1)}\sqrt{\log n}\Big)$$

Since $H(p_g, p_{g^*}) = d_H(g, g^*)$ per Eq. (17), $d_H(\hat{g}_n, g^*) = O_P(n^{-r/(2r+1)}\sqrt{\log n})$.

A.5 Proofs of propositions

Proof of Proposition 1 Denote the target's representation $g^* = \Phi_D^* \circ \cdots \circ \Phi_1^*$ with scalar edges $\varphi_{\ell,i,j}^* : D_\ell \to \mathbb{R}$ in the $W^{r,\infty}$-ball of radius M, on closed and bounded edge domains D_ℓ and layer domains $\mathcal{Z}_\ell = D_\ell^{d_{\ell-1}}$. Construct the approximant $\bar{g}_n$, with edges defined as B-splines on the slightly

enlarged domain $\tilde{D}_\ell \supseteq D_\ell$ (corresponding $\tilde{\mathcal{Z}}_\ell \supseteq \mathcal{Z}_\ell$). We derive the stated bound and the conditions under which $\bar{g}_n \in \mathcal{G}_n$.

To approximate each "true" edge $\varphi^*_{\ell,i,j}$ of g^*, apply Stein extension to $\varphi^*_{\ell,i,j} \in W^{r,\infty}(D_\ell)$, obtaining a $W^{r,\infty}(\mathbb{R})$ intermediate function with norm inflation $C_{\text{ext}}(r, |D_\ell|)$. Then, restrict to $\tilde{D}_\ell$ and apply Schumaker's order-m quasi-interpolant (Schumaker, 2007, Thm. 6.18) with J_n uniform knots, yielding B-splines $\tilde{\varphi}_{\ell,i,j}$. It follows from (Schumaker, 2007, Cor. 6.21) (which requires $1 \le r \le m$) that:

$$\|\tilde{\varphi}_{\ell,i,j} - \varphi^*_{\ell,i,j}\|_{\infty,D_\ell} \le \eta J_n^{-r} \tag{28}$$

Further, the derivative $\|\tilde{\varphi}'_{\ell,i,j}\|_\infty$ is bounded by L_{edge} ((Schumaker, 2007, Thm. 6.25), which requires $m \ge 2$). It follows that each $\tilde{\varphi}_{\ell,i,j}$ is L_{edge}-Lipschitz on $\tilde{\mathcal{Z}}_\ell$. Finally, $\|\boldsymbol{\theta}^{\tilde{\varphi}_{\ell,i,j}}\|_\infty \le C_\theta$ for constant C_θ if J_n exceeds a threshold $J_{\min}$ ((Schumaker, 2007, Thm. 4.42)).

Assemble the representation layers $\tilde{\Phi}_\ell$. Each layer coordinate consists of $d_{\ell-1} \le \max(d_0, Q)$ edges. Per the edges' Lipschitz property, the summation $\tilde{\Phi}_\ell$ is $L_{\text{edge}} \max(d_0, Q)$-Lipschitz on $\tilde{\mathcal{Z}}_\ell$. Further, summing Eq. (28) over $d_{\ell-1}$ gives $\|\tilde{\Phi}_\ell(\mathbf{u}) - \Phi^*_\ell(\mathbf{u})\|_\infty \le \eta J_n^{-r} \max(d_0, Q)$ for every $\mathbf{u} \in \mathcal{Z}_\ell$. By Lemma 2(i) at coefficient cap C_θ and summation over $d_{\ell-1}$ inputs, $\tilde{\Phi}_\ell$ maps $\tilde{\mathcal{Z}}_\ell$ into $\tilde{\mathcal{Z}}_{\ell+1}$ and the composition $\tilde{\Phi}_D \circ \cdots \circ \tilde{\Phi}_1$ is well-defined.

Finally, assemble the approximant $\bar{g}_n = \tilde{\Phi}_D \circ \cdots \circ \tilde{\Phi}_1$. Invoking Lemma 1 with $L_0 = L_{\text{edge}} \max(d_0, Q), \eta_\ell = \eta J_n^{-r} \max(d_0, Q)$:

$$\|\bar{g}_n - g^*\|_{\infty,\mathcal{Z}} \le D L_{\text{edge}}^{D-1} \eta \max(d_0, Q)^D J_n^{-r} =: C_{\text{app}} J_n^{-r} \tag{29}$$

It remains to enforce $\bar{g}_n \in \mathcal{G}_n$. From the assumption $\|g^*\|_{\infty,\mathcal{Z}} \le M_0$ and Eq. (29) it follows $\|\bar{g}_n\|_\infty \le M_0 + C_{\text{app}} J_n^{-r}$. Thus define constant:

$$J_0 = \max\left(J_{\min}, \left\lceil (C_{\text{app}}/M_0)^{1/r} \right\rceil\right)$$

Then for $J_n \ge J_0$, $C_{\text{app}} J_n^{-r} \le M_0$ thus $\|\bar{g}_n\|_\infty \le 2M_0 = M'$. Coefficient bound $\|\boldsymbol{\theta}^{\tilde{\varphi}_{\ell,i,j}}\|_\infty \le C_\theta$ holds since $J_0 \ge J_{\min}$, so each edge of $\bar{g}_n$ satisfies $\mathcal{G}_n$'s coefficient cap. Thus, $\bar{g}_n \in \mathcal{G}_n$.

Proof of Proposition 2 Apply Lemma 2 per edge yielding $\|\varphi^{(\boldsymbol{\theta})}_{\ell,i,j}\|_{\infty,I_\ell} \le C_\theta$, and since edges are piecewise polynomial on the per-edge interval I_ℓ, with Lemma 2 (ii) they are Lipschitz of constant $\le c_{\text{slope}}(J_n + 1)C_\theta$. Aggregate edges into each node of layer Φ_ℓ, first obtaining per-layer Lipschitz constant:

$$L_n^{(\boldsymbol{\theta})} \le c_1(J_n + 1) \tag{30}$$

for $c_1 = \max(d_0, Q) c_{\text{slope}} C_\theta \ge 1$. Then linearity of B-splines and Lemma 2 (i) on $\varphi^{(\boldsymbol{\theta}-\boldsymbol{\theta}')}_{\ell,i,j}$ gives per-edge perturbation for every $\mathbf{u} \in \tilde{\mathcal{Z}}_\ell$. Aggregating:

$$\|\Phi^{(\boldsymbol{\theta})}_\ell(\mathbf{u}) - \Phi^{(\boldsymbol{\theta}')}_\ell(\mathbf{u})\|_\infty \le \max(d_0, Q)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty$$

Applying Lemma 1, each layer's perturbation telescopes through the Lipschitz property of the layers:

$$\|g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}'}\|_{\infty,\mathcal{Z}} \le D\left(L_n^{(\boldsymbol{\theta})}\right)^{D-1} \max(d_0, Q)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty = L_n\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty \tag{31}$$

Finally bound L_n in terms of p_n. Since $J_n + m \le p_n$, Eq. (30) implies $L_n^{(\boldsymbol{\theta})} \le c_1 p_n$, substituting into the definition of L_n:

$$L_n \le C_L p_n^{D-1}$$

for an architecture-dependent $C_L \ge 1$.

Convert bracketing in $\mathcal{G}_n$ to covering in Θ_n via (van der Vaart & Wellner, 2023, Thm. 2.7.17). From Eq. (31) we have measurable envelope L_n on $\mathcal{Z}$. For every $\rho > 0$:

$$N_{[]}(2\rho L_n, \mathcal{G}_n, \|\cdot\|_{\infty,\mathcal{Z}}) \le N(\rho, \Theta_n, \|\cdot\|_\infty)$$

The right side is the ρ-covering number of parameter box Θ_n in ℓ^∞. Partitioning the axes into $1 + \lfloor C_\theta/\rho \rfloor$ intervals of width 2ρ gives:

$$N(\rho, \Theta_n, \|\cdot\|_\infty) \le (1 + C_\theta/\rho)^{p_n}$$

For $\varepsilon = 2\rho L_n \in (0, 2M']$, taking logs:

$$\log N_{[]}(\varepsilon, \mathcal{G}_n, \|\cdot\|_{\infty,\mathcal{Z}}) \le p_n \log(1 + 2C_\theta L_n/\varepsilon) \tag{32}$$

Let $A = 2C_\theta C_L$ and $y = Ap_n^{D-1}/\varepsilon$, then:

$$\log(1 + 2C_\theta L_n/\varepsilon) \le \log(1 + Ap_n^{D-1}/\varepsilon) = \log(1 + y) \tag{33}$$

Define the constants:

$$p_{\min} = \max\left\{2eM', (2M')^{D-1}\right\}, \qquad C_{\text{ent}} = D + \log(2A)$$

For $p_n \ge p_{\min}, \varepsilon \in (0, 2M']$:

$$\begin{aligned} p_n/\varepsilon \ge 2eM'/(2M') = e &\implies \log(p_n/\varepsilon) \ge 1 \\ p_n \ge (2M')^{D-1} \ge \varepsilon^{D-1} &\implies (D-1)\log p_n - \log\varepsilon \le D\log(p_n/\varepsilon) \end{aligned}$$

Splitting on y, if $y < 1$ then $\log(1 + y) \le \log 2 \le \log(2A)$, and if $y \ge 1$:

$$\log(1 + y) \le \log(2y) = \log(2A) + (D-1)\log p_n - \log\varepsilon \le \log(2A) + D\log(p_n/\varepsilon)$$

In both cases and using $\log(p_n/\varepsilon) \ge 1, \log(2A) \ge 0$:

$$\log(1 + y) \le (\log(2A) + D)\log(p_n/\varepsilon) = C_{\text{ent}}\log(p_n/\varepsilon)$$

Then via Eqs. (32) and (33):

$$\log N_{[]}(\varepsilon, \mathcal{G}_n, \|\cdot\|_{\infty,\mathcal{Z}}) \le p_n \log(1 + 2C_\theta L_n/\varepsilon) \le p_n C_{\text{ent}} \log(p_n/\varepsilon) \tag{34}$$

Proof of Proposition 3 First observe that the zero set is shared between any g_1, g_2 as it depends only on the censoring mechanism:

$$\begin{aligned} \{(\mathbf{x}, y, \Delta) : p_g(\mathbf{x}, y, \Delta) = 0\} = &\{(\mathbf{x}, y, 1) : S_C(y \mid \mathbf{x}) = 0\} \\ &\cup \{(\mathbf{x}, y, 0) : y \in [0, 1),\ f_C(y \mid \mathbf{x}) = 0\} \\ &\cup \{(\mathbf{x}, 1, 0) : \alpha(\mathbf{x}) = 0\} \end{aligned}$$

Fix the convention that on this zero set, $\frac{p_{g_1}}{p_{g_2}} = 1 \implies \log\frac{p_{g_1}}{p_{g_2}} = 0$, and focus on the remaining space outside of it, where $S_C, f_C, \alpha > 0$. Let $\eta = \|g_1 - g_2\|_{\infty,\mathcal{Z}}$. By MVT on $u \mapsto e^u$ and integrating over y given $g_i(\mathbf{x}, y) \in [-B, B]$:

$$|\Lambda_{g_1} - \Lambda_{g_2}| \le e^B \eta$$

Across the three strata of μ:

(i) $\Delta = 1$:

$$\begin{aligned} \frac{p_{g_1}}{p_{g_2}} = \frac{S_C e^{g_1} e^{-\Lambda_{g_1}}}{S_C e^{g_2} e^{-\Lambda_{g_2}}} &= e^{g_1 - g_2} e^{-(\Lambda_{g_1} - \Lambda_{g_2})} \\ \implies \left|\log\frac{p_{g_1}}{p_{g_2}}\right| \le |g_1 - g_2| + |\Lambda_{g_1} - \Lambda_{g_2}| &\le \eta + e^B\eta \le (1 + e^B)\eta \end{aligned}$$

(ii) $\Delta = 0, y \in [0, 1)$:

$$\frac{p_{g_1}}{p_{g_2}} = \frac{f_C\, e^{-\Lambda_{g_1}}}{f_C\, e^{-\Lambda_{g_2}}} \implies \left|\log\frac{p_{g_1}}{p_{g_2}}\right| \le |\Lambda_{g_1} - \Lambda_{g_2}| \le e^B\eta \le (1 + e^B)\eta$$

(iii) $\Delta = 0, y = 1$: Same as the above, with $f_C \to \alpha, y \to 1$.

Thus the required bound holds across the μ-support.

Proof of Proposition 4 The bracketing integral is:

$$\mathcal{J}_{[]}(\delta, \mathcal{P}_n, H) = \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{P}_n, H)} d\varepsilon.$$

Recall the sup-norm bound $\mathcal{G}_n \subseteq \{g : \|g\|_{\infty,\mathcal{Z}} \leq M'\}$. Invoke Corollary 3 at $B = M'$ to transfer the Hellinger count to sup-norm count.

$$\log N_{[]}(\varepsilon, \mathcal{P}_n, H) \leq \log N_{[]}\left(\varepsilon/\sqrt{C_h(M')}, \mathcal{G}_n, \|\cdot\|_{\infty,\mathcal{Z}}\right), \qquad C_h(B) := \frac{5}{4}e^{3B}$$

Pass to Proposition 2 at scale $\varepsilon/\sqrt{C_h(M')}$:

$$\log N_{[]}(\varepsilon, \mathcal{P}_n, H) \leq C_{\text{ent}} p_n(\log(p_n/\varepsilon) + \kappa), \qquad \kappa := \tfrac{1}{2}\log C_h(M')$$

on $\varepsilon \in (0, 2M'\sqrt{C_h(M')}]$ (so the substituted scale lies in $(0, 2M']$) and $J_n \geq J^*_{\text{ent}} := \min\{J \in \mathbb{N}_{\geq 1} : E(J+m) \geq p_{\min}\}$ (so $p_n \geq p_{\min}$).

Set $c_0 := E(1+m)$. $J_n \geq 1$ gives $p_n = EJ_n + Em \leq c_0 J_n$, and a threshold ensuring $\log(J_n/\varepsilon) \geq 0$ on the operating domain is $J_n \geq \lceil 2M'\sqrt{C_h(M')}\rceil$. Define

$$J^* := \max\left(J^*_{\text{ent}}, \lceil 2M'\sqrt{C_h(M')}\rceil\right).$$

For $J_n \geq J^*$ and $\varepsilon \in (0, 2M'\sqrt{C_h(M')}]$, $\log(J_n/\varepsilon) \geq 0$. The inequality $A + B \leq \max(1, B)(A + 1)$ for $A \geq 0$, $B \in \mathbb{R}$ at $A = \log(J_n/\varepsilon)$, $B = \log c_0 + \kappa$ gives $\log(p_n/\varepsilon) + \kappa \leq \max(1, \log c_0 + \kappa)[\log(J_n/\varepsilon) + 1]$. Substituting $p_n \leq c_0 J_n$ and setting $C_0 := C_{\text{ent}} c_0 \max(1, \log c_0 + \kappa)$:

$$\log N_{[]}(\varepsilon, \mathcal{P}_n, H) \leq C_0 J_n(\log(J_n/\varepsilon) + 1) \tag{35}$$

$J_n[\log(J_n/\varepsilon) + 1] \geq 1$ on this domain, so the integrand in the bracketing integral satisfies:

$$\begin{aligned}\sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{P}_n, H)} &\leq \sqrt{1 + C_0 J_n(\log(J_n/\varepsilon) + 1)} \\ &\leq \sqrt{(C_0+1)J_n}\sqrt{\log(J_n/\varepsilon) + 1}\end{aligned}$$

$\sqrt{(C_0+1)J_n}$ is constant in ε and factors out. Substituting $u = \log(J_n/\varepsilon)$ and integrating by parts:

$$\int_0^\delta \sqrt{\log(J_n/\varepsilon) + 1} d\varepsilon \leq 2\delta\sqrt{\log(J_n/\delta) + 1}$$

Therefore, the bracketing integral becomes:

$$\mathcal{J}_{[]}(\delta, \mathcal{P}_n, H) \leq \sqrt{(C_0+1)J_n} \cdot 2\delta\sqrt{\log(J_n/\delta) + 1}$$

Substitute $K = 2\sqrt{C_0 + 1}$:

$$\mathcal{J}_{[]}(\delta, \mathcal{P}_n, H) \leq K\sqrt{J_n}\delta\sqrt{\log(J_n/\delta) + 1} =: \widetilde{J}_n(\delta) \tag{36}$$

for $\delta \in (0, 2M'\sqrt{C_h(M')}]$ and $J_n \geq J^*$. $M_0 \geq 1 \implies 2M'\sqrt{C_h(M')} > \sqrt{2}$, so $(0, \sqrt{2}] \subseteq (0, 2M'\sqrt{C_h(M')}]$ and thus Eq. (36) holds for $\delta \in (0, \sqrt{2}]$, giving result (iii).

It remains to verify the shape of $\widetilde{J}_n(\delta)$ on $(0, \sqrt{2}]$. Set $u(\delta) := \log(J_n/\delta)$. $J_n \geq J^* \geq 32 > \sqrt{2} \geq \delta$ gives $u > 0$. Differentiating $\delta\sqrt{u+1}$ in δ (with $u'(\delta) = -1/\delta$) gives $\frac{2u+1}{2\sqrt{u+1}} > 0$, so $\widetilde{J}_n = K\sqrt{J_n}\delta\sqrt{u+1}$ is strictly increasing, giving result (i).

Finally, $\widetilde{J}_n(\delta)/\delta^{3/2} = K\sqrt{J_n}\delta^{-1/2}\sqrt{u+1}$. $\delta^{-1/2}$ and $\sqrt{u+1}$ are strictly positive and strictly decreasing on $\delta \in (0, \sqrt{2}]$, so their product is strictly decreasing, giving result (ii).

A.6 Technical Lemmas

Lemma 1 (Telescoping). *Let pairs $(\Phi_\ell, \tilde{\Phi}_\ell)_{\ell=1}^D$ of continuous maps on nested domains $\mathcal{Z}_\ell \subseteq \tilde{\mathcal{Z}}_\ell$ ($\ell \in \{1, \ldots, D+1\}$) with $\Phi_\ell : \mathcal{Z}_\ell \to \mathcal{Z}_{\ell+1}$ and $\tilde{\Phi}_\ell : \tilde{\mathcal{Z}}_\ell \to \tilde{\mathcal{Z}}_{\ell+1}$, where:*

1. *Each $\tilde{\Phi}_\ell$ is L_0-Lipschitz on $\tilde{\mathcal{Z}}_\ell$ in the sup-norm, $L_0 \geq 1$, and*
2. *$\|\tilde{\Phi}_\ell(\mathbf{u}) - \Phi_\ell(\mathbf{u})\|_\infty \leq \eta_\ell$ for every $\mathbf{u} \in \mathcal{Z}_\ell$.*

Denote the compositions $F_D = \Phi_D \circ \cdots \circ \Phi_1$, $\tilde{F}_D = \tilde{\Phi}_D \circ \cdots \circ \tilde{\Phi}_1$. Then, the composition error:

$$\|F_D - \tilde{F}_D\|_{\infty,\mathcal{Z}_1} \leq D\, L_0^{D-1} \max_{1\leq \ell \leq D} \eta_\ell \tag{37}$$

Proof: This is shown by telescoping $F_D - \tilde{F}_D$ into D incremental one-layer swaps of Φ_ℓ for $\tilde{\Phi}_\ell$. Sum the post-swap maps' cumulative Lipschitz constants $L_0^{D-\ell}$ from assumption (1) and the per-layer perturbations η_ℓ from assumption (2). The final inequality follows from $L_0^{D-\ell} \leq L_0^{D-1}$.

Lemma 2 (Spline bounds)**.** *For an order-m B-spline expansion $s_{\boldsymbol{\theta}} = \sum_k \theta_k B_k^{I,J}$ on a closed bounded interval I with J uniform interior knots, m-fold boundary knots, and mesh $\bar{\Delta} = |I|/(J+1)$ it holds that:*

(i) $\|s_{\boldsymbol{\theta}}\|_{\infty,I} \leq \|\boldsymbol{\theta}\|_\infty$, and

(ii) $\|s'_{\boldsymbol{\theta}}\|_{L^\infty(I)} \leq c_{\text{slope}} \|\boldsymbol{\theta}\|_\infty / \bar{\Delta}$ for c_{slope} depending only on m.

Proof: Partition of unity on I (Schumaker, 2007, Thm. 4.20) and non-negativity of B_k give property (i). For property (ii), express the right derivative $D_+ s$ as an order $m-1$ B-spline expansion, whose coefficients are bounded by $c_{\text{slope}}\|\boldsymbol{\theta}\|_\infty/\bar{\Delta}$ (Schumaker, 2007, Thm. 5.9), and convert this to a bound on the derivative at every point on I except the knot points, obtaining the essential supremum $\|s'_{\boldsymbol{\theta}}\|_{L^\infty(I)}$.

Lemma 3 (Sup-norm Hellinger bound)**.** *Let $B \in [0,\infty)$ and define $C_h(B) = \frac{5}{4}e^{3B}$. For every pair of measurable log-hazards $g_1, g_2 : \mathcal{Z} \to \mathbb{R}$ with $\|g_i\|_{\infty,\mathcal{Z}} \leq B$ ($i = 1, 2$):*

$$H^2(p_{g_1}, p_{g_2}) \leq C_h(B) \|g_1 - g_2\|^2_{\infty,\mathcal{Z}} \tag{38}$$

Proof: Decompose $H^2(p_{g_1}, p_{g_2}) = H^2_{\text{ac}}(p_{g_1}, p_{g_2}) + H^2_{\text{at}}(p_{g_1}, p_{g_2})$ via $\mu = \mu_{\text{ac}} + \mu_{\text{at}}$

$$H^2_{\text{ac}}(p_{g_1}, p_{g_2}) = \sum_{\Delta\in\{0,1\}} \int_{\mathcal{X}} \int_0^1 \left(\sqrt{p_{g_1}(\mathbf{x}, y, \Delta)} - \sqrt{p_{g_2}(\mathbf{x}, y, \Delta)}\right)^2 dy\, dP_X(\mathbf{x})$$

$$H^2_{\text{at}}(p_{g_1}, p_{g_2}) = \int_{\mathcal{X}} \left(\sqrt{p_{g_1}(\mathbf{x}, 1, 0)} - \sqrt{p_{g_2}(\mathbf{x}, 1, 0)}\right)^2 dP_X(\mathbf{x})$$

Define $a_g(\mathbf{x}, y) = e^{g(\mathbf{x},y)/2}$ and $b_g(y \mid \mathbf{x}) = e^{-\Lambda_g(y|\mathbf{x})/2}$ where Λ_g is the cumulative hazard induced by g. Factorise $\sqrt{p_g}$ over the three possibilities:

$$\begin{cases} \sqrt{p_g(\mathbf{x}, y, 1)} = \sqrt{S_C(y \mid \mathbf{x})} a_g(\mathbf{x}, y) b_g(y \mid \mathbf{x}) & \Delta = 1, y \in [0, 1) \\ \sqrt{p_g(\mathbf{x}, y, 0)} = \sqrt{f_C(y \mid \mathbf{x})}\, b_g(y \mid \mathbf{x}) & \Delta = 0, y \in [0, 1) \\ \sqrt{p_g(\mathbf{x}, 1, 0)} = \sqrt{\alpha(\mathbf{x})}\, b_g(1 \mid \mathbf{x}) & \Delta = 0, y = 1 \end{cases}$$

Let $\eta = \|g_1 - g_2\|_{\infty,\mathcal{Z}}$ and derive Lipschitz bounds:

(i)
$$|a_{g_1}(\mathbf{x}, y) - a_{g_2}(\mathbf{x}, y)| \leq \frac{1}{2} e^{B/2} \eta \tag{39}$$
by MVT on $u \mapsto e^{u/2}$ given $g_i(\mathbf{x}, y) \in [-B, B]$.

(ii)
$$|\Lambda_{g_1}(y \mid \mathbf{x}) - \Lambda_{g_2}(y \mid \mathbf{x})| \leq e^B \eta \tag{40}$$
by MVT on $u \mapsto e^u$ and integration over y.

(iii)
$$|b_{g_1}(y \mid \mathbf{x}) - b_{g_2}(y \mid \mathbf{x})| \leq \frac{1}{2} e^B \eta \tag{41}$$
by MVT on $u \mapsto e^{-u/2}$ and (ii).

Per-factor upper bounds: $a_{g_i} \le e^{B/2}$ from $g_i \le B$; $b_{g_i} \le 1$ from $\Lambda_{g_i} \ge 0$; $S_C \le 1$ from (C3). Product Lipschitz from $a_1 b_1 - a_2 b_2 = (a_1 - a_2)\, b_1 + a_2\, (b_1 - b_2)$ with (i), (iii):

$$|a_{g_1} b_{g_1} - a_{g_2} b_{g_2}| \le \frac{1}{2} e^{B/2} (1 + e^B)\, \eta \le e^{3B/2}\, \eta \tag{42}$$

Substitute the squared difference and integrate against μ:

(i) $\Delta = 1, y \in [0, 1)$:

$$\left(\sqrt{p_{g_1}(\mathbf{x}, y, 1)} - \sqrt{p_{g_2}(\mathbf{x}, y, 1)}\right)^2 = S_C(y \mid \mathbf{x})\, (a_{g_1} b_{g_1} - a_{g_2} b_{g_2})^2 \le S_C(y \mid \mathbf{x}) e^{3B} \eta^2$$

$$\implies \int_{\mathcal{X}} \int_0^1 \left(\sqrt{p_{g_1}(\mathbf{x}, y, 1)} - \sqrt{p_{g_2}(\mathbf{x}, y, 1)}\right)^2 dy\, dP_X(\mathbf{x}) \le e^{3B} \eta^2,$$

using $\int_0^1 S_C(y \mid \mathbf{x})\, dy \le 1$ from $S_C \le 1$ and $P_X(\mathcal{X}) = 1$.

(ii) $\Delta = 0, y \in [0, 1)$:

$$\left(\sqrt{p_{g_1}(\mathbf{x}, y, 0)} - \sqrt{p_{g_2}(\mathbf{x}, y, 0)}\right)^2 = f_C(y \mid \mathbf{x})\, (b_{g_1} - b_{g_2})^2 \le f_C(y \mid \mathbf{x})\, \frac{1}{4} e^{2B} \eta^2$$

$$\implies \int_{\mathcal{X}} \int_0^1 \left(\sqrt{p_{g_1}(\mathbf{x}, y, 0)} - \sqrt{p_{g_2}(\mathbf{x}, y, 0)}\right)^2 dy\, dP_X(\mathbf{x}) \le \frac{1}{4} e^{2B} \eta^2 \int_{\mathcal{X}} \int_0^1 f_C(y \mid \mathbf{x})\, dy\, dP_X(\mathbf{x})$$

(iii) $\Delta = 0, y = 1$:

$$\left(\sqrt{p_{g_1}(\mathbf{x}, 1, 0)} - \sqrt{p_{g_2}(\mathbf{x}, 1, 0)}\right)^2 = \alpha(\mathbf{x})\, (b_{g_1}(1 \mid \mathbf{x}) - b_{g_2}(1 \mid \mathbf{x}))^2 \le \alpha(\mathbf{x})\, \frac{1}{4} e^{2B}\, \eta^2$$

$$\implies \int_{\mathcal{X}} \left(\sqrt{p_{g_1}(\mathbf{x}, 1, 0)} - \sqrt{p_{g_2}(\mathbf{x}, 1, 0)}\right)^2 dP_X(\mathbf{x}) \le \frac{1}{4} e^{2B}\, \eta^2 \int_{\mathcal{X}} \alpha(\mathbf{x})\, dP_X(\mathbf{x})$$

Combine (ii) and (iii) via (C3) $\int_0^1 f_C(y \mid \mathbf{x})\, dy + \alpha(\mathbf{x}) = 1$ and $P_X(\mathcal{X}) = 1$:

$$\frac{1}{4} e^{2B}\, \eta^2 \int_{\mathcal{X}} \left(\int_0^1 f_C(y \mid \mathbf{x})\, dy + \alpha(\mathbf{x})\right) dP_X(\mathbf{x}) = \frac{1}{4} e^{2B}\, \eta^2$$

Summing the three strata into $H^2 = H^2_{\mathrm{ac}} + H^2_{\mathrm{at}}$ and since $B \ge 0$:

$$H^2(p_{g_1}, p_{g_2}) \le e^{3B}\, \eta^2 + \frac{1}{4} e^{2B}\, \eta^2 \le \frac{5}{4} e^{3B}\, \eta^2 = C_h(B)\, \|g_1 - g_2\|^2_{\infty, \mathcal{Z}}$$

Corollary 3. *Under Lemma 3, for density class $\mathcal{P} = \{p_g : g \in \mathcal{G}\}$ with $\mathcal{G} \subseteq \{g : \|g\|_{\infty, \mathcal{Z}} \le B\}$ and every $\varepsilon > 0$:*

$$\log N_{[]}(\varepsilon, \mathcal{P}, H) \le \log N_{[]}(\varepsilon / \sqrt{C_h(B)}, \mathcal{G}, \|\cdot\|_{\infty, \mathcal{Z}}) \tag{43}$$

Proof: Apply (van der Vaart & Wellner, 2023, Thm. 2.7.17) to class $\mathcal{F} = \{\sqrt{p_g} : g \in \mathcal{G}\}$ indexed by $(\mathcal{G}, \|\cdot\|_{\infty, \mathcal{Z}})$ with norm $L^2(\mu)$. The 3-case Lipschitz bounds on $|\sqrt{p_{g_1}} - \sqrt{p_{g_2}}|$ from Lemma 3's factorisation Eqs. (39) to (41) give envelope:

$$F(\mathbf{x}, y, \Delta) = \begin{cases} e^{3B/2} & \Delta = 1, y \in [0, 1) \\ \frac{1}{2}\, e^B \sqrt{f_C(y \mid \mathbf{x})} & \Delta = 0, y \in [0, 1) \\ \frac{1}{2}\, e^B \sqrt{\alpha(\mathbf{x})} & \Delta = 0, y = 1 \end{cases}$$

Integrating F^2 over μ and combining the two $\Delta = 0$ strata $\int_0^1 f_C(y \mid \mathbf{x})\, dy + \alpha(\mathbf{x}) = 1$:

$$\|F\|^2_{L^2(\mu)} \le e^{3B} + \frac{1}{4} e^{2B} \le \frac{5}{4} e^{3B} = C_h(B)$$

(van der Vaart & Wellner, 2023, Thm. 2.7.17) then gives, for every $\eta > 0$ and under $N(\eta, \mathcal{G}, \|\cdot\|_{\infty,\mathcal{Z}}) < \infty$:

$$N_{[]}\left(2\sqrt{C_h(B)}\eta, \{\sqrt{p_g} : g \in \mathcal{G}\}, L^2(\mu)\right) \leq N\left(\eta, \mathcal{G}, \|\cdot\|_{\infty,\mathcal{Z}}\right) \tag{44}$$

An $L^2(\mu)$-bracket (ψ_L, ψ_U) for $\sqrt{p_g}$ produces a Hellinger bracket (ℓ, u) for p_g of the same width via $\ell = (\psi_L \vee 0)^2$, $u = (\psi_U \vee 0)^2$. Then $\ell \leq p_g \leq u$ μ almost everywhere (since $\sqrt{p_g} \geq 0$), and $H^2(u, \ell) \leq \int (\psi_U - \psi_L)^2 \, d\mu$ since $\sqrt{u} = \psi_U \vee 0$, $\sqrt{\ell} = \psi_L \vee 0$, and $x \mapsto x \vee 0$ is 1-Lipschitz. Thus:

$$N_{[]}(\varepsilon, \mathcal{P}, H) \leq N_{[]}(\varepsilon, \{\sqrt{p_g} : g \in \mathcal{G}\}, L^2(\mu)) \tag{45}$$

Midpoints of a sup-norm r-bracket cover of $\mathcal{G}$ form an $r/2$-cover:

$$N(r/2, \mathcal{G}, \|\cdot\|_{\infty,\mathcal{Z}}) \leq N_{[]}(r, \mathcal{G}, \|\cdot\|_{\infty,\mathcal{Z}}) \tag{46}$$

Substitute $\eta = \varepsilon/(2\sqrt{C_h(B)})$ in Eq. (44) and $r = \varepsilon/\sqrt{C_h(B)}$ in Eq. (46). Then through Eq. (45):

$$\begin{aligned}
\log N_{[]}(\varepsilon, \mathcal{P}, H) &\leq \log N_{[]}(\varepsilon, \{\sqrt{p_g} : g \in \mathcal{G}\}, L^2(\mu)) \\
&\leq \log N(\varepsilon/(2\sqrt{C_h(B)}), \mathcal{G}, \|\cdot\|_{\infty,\mathcal{Z}}) \\
&\leq \log N_{[]}(\varepsilon/\sqrt{C_h(B)}, \mathcal{G}, \|\cdot\|_{\infty,\mathcal{Z}})
\end{aligned}$$

Lemma 4 (Measurable selector)**.** *For every $\eta_n \geq 0$ and $J_n \geq J_0$ there exists a Borel-measurable $\hat{g}_n : \Omega \to \mathcal{G}_n$ with $\mathcal{L}_n(\hat{g}_n) \geq \mathcal{L}_n(\bar{g}_n) - \eta_n$, where $\bar{g}_n$ is the sieve approximant of Proposition 1.*

$\mathcal{G}_n$ is parameterised by its B-spline coefficient vector $\boldsymbol{\theta} \in \Theta_n^{\text{feas}} \subset \mathbb{R}^{p_n}$ where $\Theta_n^{\text{feas}} = \{\boldsymbol{\theta} \in \Theta_n : \|g_{\boldsymbol{\theta}}\|_{\infty,\mathcal{Z}} \leq M'\}$. The map $(\boldsymbol{\theta}, \omega) \mapsto \mathcal{L}_n(g_{\boldsymbol{\theta}}; \omega)$ is Carathéodory because it is (a) continuous in $\boldsymbol{\theta}$ for each ω via B-spline linearity in $\boldsymbol{\theta}$ and dominated convergence with envelope $e^{M'}$ on the cumulative-hazard integral $\int_0^{Y_i} e^{g_{\boldsymbol{\theta}}(X_i,s)} \, ds$; and (b) Borel in ω for each $\boldsymbol{\theta}$. This implies joint Borel measurability, and $\omega \mapsto \mathcal{L}_n(\bar{g}_n; \omega)$ is Borel by (b) since $\bar{g}_n$ is deterministic. Define the set:

$$\mathcal{S}_n(\omega) = \left\{\boldsymbol{\theta} \in \Theta_n^{\text{feas}} : \mathcal{L}_n(g_{\boldsymbol{\theta}}; \omega) \geq \mathcal{L}_n(\bar{g}_n; \omega) - \eta_n\right\}$$

This set is non-empty, containing the coefficient vector of $\bar{g}_n$, closed-valued by continuity in $\boldsymbol{\theta}$, and has Borel graph by joint measurability. Kuratowski-Ryll-Nardzewski (Aliprantis & Border, 1999, Thm. 17.13) then supplies a Borel selector $\hat{\boldsymbol{\theta}}_n : \Omega \to \Theta_n^{\text{feas}}$ with $\hat{\boldsymbol{\theta}}_n(\omega) \in \mathcal{S}_n(\omega)$. The composition $\hat{g}_n := g_{\hat{\boldsymbol{\theta}}_n}$ is thus Borel via the continuous parameterisation $\boldsymbol{\theta} \mapsto g_{\boldsymbol{\theta}}$.

B Implementation Details

KAPLAN-HR is implemented with the `pykan` library v0.2.8 by Liu et al. (2025b) for the KAN layers. We list the initialisation and training parameters in Table 2.

During training, the knot vectors are periodically re-fitted to the observed activation distribution via `pykan`'s grid-extension at a tunable schedule of epochs $\mathcal{U}$. Parameters are optimised with Adam using mini-batches of tunable size B. The learning rate is constant on the clinical datasets and follows a cosine schedule in the simulations (Table 2). Training stops when the validation NLL has not decreased below its running minimum for P consecutive epochs. The complete training procedure is given in Algorithm 1.

We use trapezoidal integration for both prediction and training. The prediction grid $\boldsymbol{\tau}$ used for evaluation (and validation during training) is $\tau_0 = 0$ together with the quantiles of the training observed times $\{Y_i\}_{i=1}^n$ at $K + 1$ equally spaced probability levels, duplicates removed, with $K = 256$. During training, computation of the NLL requires $\hat{\Lambda}_{\text{HR}}(Y_i \mid \mathbf{x}_i)$ for each subject. For efficiency, we evaluate this on a sub-grid $\boldsymbol{\tau}_{\mathcal{B}}$ of bin indices $\{b_i := \max\{k : \tau_k \leq Y_i\} : i \in \mathcal{B}\}$ in the mini-batch $\mathcal{B}$.

Factored first-layer evaluation Evaluating $\text{KAN}_{\boldsymbol{\theta}}(\mathbf{x}_i, t)$ at $n \times K$ (sample, time) pairs is a major cost for KAPLAN-HR. This cost can be reduced by exploiting the additive structure of Eq. (4), decomposing the first layer output as:

$$z_j^{(1)} = \sum_{i=1}^{d} \varphi_{1,i,j}(x^{(i)}) + \varphi_{1,i,d+1}(t) =: F_j(\mathbf{x}) + G_j(t). \tag{47}$$

Since F_j depends only on $\mathbf{x}$ and G_j only on t, they may be evaluated independently and then combined by broadcast addition. Subsequent layers operate on the full $z^{(1)}$, which mixes $\mathbf{x}$ and t, so the factorisation does not extend beyond the first layer.

Table 2: KAN configuration. Monospaced names are the `pykan` v0.2.8 API; symbols in parentheses are the paper's. 'tuned' entries are swept per (dataset, variant).

Parameter	Simulations	Real Data	Description
Spline order (m)	4	tuned $\in \{3, 4, 5\}$	`pykan` $k = m - 1$
`grid`	adaptive	tuned $\in \{2, 3, 4, 5\}$	Grid intervals ($J_n + 1$)
`grid_range`	$[-1, 1]$	$[-1, 1]$	Initial knot range
`grid_eps`	0.02	0.02	0: quantile; 1: uniform
`noise_scale`	0.3	0.3	Coefficient init noise
`scale_sp`	1	1	w^{s} init scale
`scale_base_mu`	0	0	w^{b} init mean
`scale_base_sigma`	1	1	w^{b} init sigma
`sp_trainable`	`True`	`True`	w^{s} trained
`sb_trainable`	`True`	`True`	w^{b} trained
`affine_trainable`	`False`	`False`	Affine frozen
`base_fun`	`silu`	`silu`	Residual function
Learning rate	10^{-2} (10^{-3} for $n > 32768$)	tuned	Adam, initial value
Learning rate schedule	cosine, $T_{\max} = 1000$	constant	Decay over training
Batch size (B)	512	128	Mini-batch size
Grid re-fits	every 5 epochs	tuned $\in \{5, 10, 15, 20\}$	Grid update frequency
ℓ_1 penalty (λ_{ℓ_1})	10^{-4}	tuned	Weight in Eq. (10)
Patience (P)	150	50	early stopping patience
Time grid (K)	256	256	Λ_{HR} mesh resolution

Algorithm 1 Training KAPLAN-HR

Require: Data $\mathcal{D} = \{(\mathbf{x}_i, Y_i, \Delta_i)\}_{i=1}^n$, time grid $\boldsymbol{\tau} = (\tau_0 \equiv 0, \tau_1, \ldots, \tau_K)$, batch size B, grid update epochs $\mathcal{U}$, patience P

Initialise $\mathrm{KAN}_{\boldsymbol{\theta}}$ with input dimension $d + 1$
Compute bin indices $b_i \leftarrow \max\{k : \tau_k \le Y_i\}$ for $i = 1, \ldots, n$
for epoch $e = 1, 2, \ldots$ **do**
  if $e \in \mathcal{U}$ **then** re-fit B-spline knot vectors
  for each batch $\mathcal{B} \subset \{1, \ldots, n\}$, $|\mathcal{B}| = B$ **do**
    $(a_1, \ldots, a_R) \leftarrow \mathrm{sorted}(\mathrm{unique}(\{b_i : i \in \mathcal{B}\}))$ ▷ R active bin indices
    Sub-grid $\boldsymbol{\tau}_{\mathcal{B}} \leftarrow (\tau_{\mathcal{B},r} := \tau_{a_r})_{r=1}^R$ with widths $\tilde{\delta}_r \leftarrow \tau_{\mathcal{B},r} - \tau_{\mathcal{B},r-1}$ ($\tau_{\mathcal{B},0} \equiv 0$)
    Evaluate $g_{i,r} \leftarrow \mathrm{KAN}_{\boldsymbol{\theta}}(\mathbf{x}_i, \tau_{\mathcal{B},r})$ for $i \in \mathcal{B}$, $r = 0, 1, \ldots, R$ ▷ Factored, cf. Eq. (47)
    $h_{i,r} \leftarrow \exp(g_{i,r})$
    $H_{i,r} \leftarrow \sum_{m=1}^{r} \frac{1}{2} (h_{i,m-1} + h_{i,m})\, \tilde{\delta}_m$ ▷ Cumulative sum (trapezoid)
    $r_i \leftarrow$ position of b_i in $(a_1, \ldots, a_R)$
    $\mathcal{L}_{\mathrm{HR}} \leftarrow \frac{1}{B} \sum_{i \in \mathcal{B}} [H_{i,r_i} - \Delta_i\, g_{i,r_i}] + \mathcal{L}_{\ell_1}$
    Adam step on $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{HR}}$
  if validation NLL has not improved for P epochs **then break**
return $\mathrm{KAN}_{\boldsymbol{\theta}}$

C Experimental Details

C.1 Computing Platform

All experiments were run on a high-performance computing cluster, using Intel Ice Lake processors. We set the memory limit to 64GB.

C.2 Simulation Experiments

We define four synthetic DGPs:

$$\zeta_r(z) := \sum_{j=0}^{7} q_j^{-r}\, \psi_r\big(q_j \pi(z-\phi_j)\big), \quad \psi_1 = |\sin|, \quad \psi_2 = \sin\cdot|\sin|$$

with $q_j = 2 \cdot 1.5^j$, $\quad \phi_j = (0.137 + 0.618034\, j) \bmod 1$

DGP-1: $g^*(\mathbf{x}, t) = 4\,\zeta_2(x_1) + 2\,\zeta_2(x_2) - 0.3\sin(3\pi t) + 0.2 + \log 3$

DGP-2: $g^*(\mathbf{x}, t) = \zeta_1(x_1) + 0.5\,\zeta_1(x_2) - 0.3\sin(3\pi t) - 1.2 + \log 3$

DGP-3: $g^*(\mathbf{x}, t) = \zeta_1\big(\frac{x_1 + 0.5x_2}{1.5}\big) - 0.9\,t - 0.75 + \log 3$

DGP-4: $g^*(\mathbf{x}, t) = 1.75\,\zeta_1(x_1)\,\big|\sin\big(4\pi(t-0.137)\big)\big| + 0.3\sin(\pi x_2) - 0.4 + \log 3$

Covariates are transformed via the multiscale function ζ_r. This function's q_j^{-r} scaling gives it a bounded r-th weak derivative, placing it in $\mathcal{F}_r^{\mathrm{KAN}}$ with smoothness r, while its kinks at every scale serve to stress the sieve's approximation error. DGP-1 uses ζ_2 ($r = 2$); DGPs 2 to 4 use ζ_1 ($r = 1$). DGP-1 and DGP-2 are additive; DGP-3 composes ζ_1 with a linear combination of x_1 and x_2; DGP-4 multiplies $\zeta_1(x_1)$ by a function of time, so its hazards are not proportional.

For each DGP we generate $R = 20$ independent training sets, each with a testing set of 2000 observations. We further split each training set $80/20$ and use the validation sub-set for early stopping. We draw i.i.d. covariates $x_1, \ldots, x_5 \sim \mathrm{Uniform}(0, 1)$ (x_3, x_4, x_5 are intentionally left as noise) and generate continuous event times by inversion from the conditional hazard $\lambda(t \mid x) = \exp(g(x, t))$. Survival quantities are evaluated on a uniform grid of $K = 256$ points on $(0, 1]$. Ground-truth survival $S(t \mid x) = \exp\left(-\int_0^t \exp(g(x, u))du\right)$ is computed by composite Simpson's rule on 2^{14} subintervals of $[0, 1]$. Censoring times $C \sim \mathrm{Exp}(\lambda_c)$ are drawn independently of T, and we apply administrative censoring at the horizon $t = 1$ and calibrate λ_c to yield an overall censoring rate of $\sim 30\%$.

For DGPs 1 and 2 we fit KAPLAN-HR models with structure $[d_0, 1]$, and for DGPs 3 and 4 we use deeper structure $[d_0, 3, 1]$. Each spline's grid carries $\lceil g_r\,(n/n_0)^{1/(2r+1)} \rceil$ uniform intervals with $n_0 = 512$, $g_1 = 2.5$ and $g_2 = 5$, so J_n grows at the required rate $n^{1/(2r+1)}$. The other model parameters are listed in Table 2. The GAM baselines are fitted using the `mgcv` R package (Wood et al., 2016), setting $k = 10$ for the covariates and $k = 15$ for the time and the remaining parameters.

C.3 Real-World Clinical Dataset Experiments

Datasets and Preprocessing We evaluate on six standard survival analysis datasets: Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012), Rotterdam Tumor Bank and German Breast Cancer Study Group (RotGBSG) (Foekens et al., 2000; Schumacher et al., 1994), National Wilms' Tumor Study (NWTCO) (Breslow & Chatterjee, 1999), Assay Of Serum Free Light Chain (FLCHAIN) (Dispenzieri et al., 2012), and Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) (Knaus et al., 1995), with covariates selected to match Kvamme & Borgan (2021); and Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson et al., 2016), following the preprocessing of Purushotham et al. (2018). Each dataset is split 10 times stratified on the event indicator into 64% train, 16% validation, and 20% test sets with random seeds 0-9. Covariates are standardised using training set statistics at each split. Times are min-max normalised to $(0, 1]$ using the full-sample range before splitting. Rows with any missing values are dropped. Table 3 summarises the six datasets. They span sample sizes from 1,937 to 35,993, covariate counts from 7 to 26, and censoring rates from 32% to 90%.

Baseline Models We compare performance against: Cox proportional hazards, implemented by `lifelines` (Davidson-Pilon, 2019); a generalised additive model (GAM), implemented by the `mgcv` R package (Wood et al., 2016); HARE (Kooperberg et al., 1995), implemented by the `polspline` R package; DeepSurv (Katzman et al., 2018), CoxTime (Kvamme et al., 2019), and DeepHit (Lee et al., 2018), implemented by `pycox` (Kvamme et al., 2019); Deep Survival Machines (DSM) (Nagpal et al., 2021), implemented by `auton-survival` (Nagpal et al., 2022); CoxKAN (Knottenbelt et al., 2025); and SuMo-Net (Rindt et al., 2022).

Table 3: Benchmark dataset characteristics: complete-case cohort size n, covariate count d, and the censoring rate.

Dataset	n	d	Censoring (%)
METABRIC	1,937	9	41.9
MIMIC-III	35,993	26	90.1
RotGBSG	2,232	7	43.2
FLCHAIN	6,524	17	69.9
NWTCO	4,028	13	85.8
SUPPORT	8,872	24	32.0

Hyperparameter Tuning All tunable models undergo hyperparameter optimisation using Weights & Biases (Weights and Biases, 2026) to minimise the IBS (Graf et al., 1999) on the validation set in split 0 (out of 10), separately for each dataset, over 100 trials per (dataset, model) pair. The search is Bayesian for every model except DSM, whose hyperparameter space is entirely categorical and is swept at random. Table 4 gives the sweep ranges for each model hyperparameter.

Scoring Survival predictions are evaluated on: the time-dependent concordance index (C-TD) by Antolini et al. (2005), implemented by `pycox` (Kvamme et al., 2019); the integrated Brier score (IBS) (Graf et al., 1999), implemented by `scikit-survival` (Pölsterl, 2020); the integrated calibration index (ICI) (Austin et al., 2020), implemented by `lifelines` (Davidson-Pilon, 2019), measured at the median training set event time (per split); and the D-calibration p-value (D-cal) (Haider et al., 2020), implemented by `SurvivalEVAL` (Qi et al., 2024). All metrics except D-cal are averaged over the 10 data splits and reported as the mean $\pm$ standard deviation. The D-cal is reported as the number of splits (out of 10) with $p > 0.05$. Statistical comparisons between KAPLAN-HR and the best-scoring baseline model on each dataset use the two-sided Wilcoxon signed-rank test on the 10 per-split metric values, with Holm-Bonferroni correction applied across the six datasets per metric.

D Verification of the Sieve Conditions

We directly verify that the fitted estimators of the simulation experiments fulfil conditions (C5) and (C6).

Method We construct $\bar{g}_n$ explicitly for each DGP by least-squares fitting order-m B-splines on uniform knots to each analytic edge of g^*, with the number of interior knots following the schedule of Appendix C.2 at each n. We calculate $M_0 = \|g^*\|_{\infty,\mathcal{Z}}$ on a $(128+1)^3$ tensor grid in (x_1, x_2, t) together with a 2^{18}-point Sobol sweep of the input domain, and we set C_θ just above the largest coefficient of any $\bar{g}_n$ on the schedule. For each of the $R = 20$ fits at every DGP and n, we compare the trained model's spline coefficients against C_θ, and we calculate $\|\hat{g}_n\|_{\infty,\mathcal{Z}}$ by evaluating over a 2^{16}-point sweep of the input domain, to compare against $M' = 2M_0$. For (C6) we measure the shortfall $\hat{\eta}_n = (\mathcal{L}_n(\bar{g}_n) - \mathcal{L}_n(\hat{g}_n))_+$ on the same fits and check that the ratio $\hat{\eta}_n / (n^{-2r/(2r+1)} \log n)$ is bounded by a constant C_η.

Results Table 5 reports the (C5) quantities against their thresholds. We find that $40/160$ fits on DGP-3 and $6/160$ fits on DGP-4 exceed M', concentrated in the small sample sizes and diminishing at larger n, with one fit above M' at $n = 2^{15}$ on DGP-4. No fit on DGP-1 or DGP-2 exceeds M'. The coefficient bound is exceeded in $8/32$ cells, at a maximum of 32.48 against $C_\theta = 12$, with no growth in n. Condition (C5) admits any n-independent coefficient bound, so this rescales C_θ and leaves the rate unchanged. Table 6 reports the (C6) shortfall ratio. We find that $621/640$ fits attain at least the log-likelihood of $\bar{g}_n$, so their shortfall is zero, with the misses confined to $n \geq 2^{14}$. The largest ratio anywhere on the grid is 0.595, so (C6) holds across the grid with $C_\eta = 0.6$.

Table 4: Hyperparameter search spaces for all models.

Hyperparameter	Range / values	Distribution
KAPLAN-HR		
Learning rate	$[10^{-4}, 10^{-2}]$	log-uniform
Spline knots J_n	$\{2, 3, 4, 5\}$	categorical
Spline order m	$\{3, 4, 5\}$	categorical
KAN layers D	$\{2, 3\}$	categorical
Hidden dimension Q	$\{5, 8, 16, 32\}$	categorical
Grid update frequency	$\{5, 10, 15, 20\}$	categorical
ℓ_1 penalty λ_{ℓ_1}	$[0, 10^{-3}]$	uniform
Batch size	128	constant
Max training epochs	200	constant
Shared time grid size K	256	constant
DeepSurv, CoxTime, DeepHit		
Hidden layers	$\{1, 2, 4\}$	categorical
Nodes per layer	$\{32, 64, 128, 256\}$	categorical
Learning rate	$[10^{-4.5}, 10^{-1.5}]$	log-uniform
Weight decay	$\{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.4\}$	categorical
Dropout	$[0, 0.5]$	uniform
Batch normalisation	{true, false}	categorical
Batch size	$\{32, 64, 128, 256, 512\}$	categorical
CoxTime only		
Shrinkage λ	$\{0, 0.001, 0.01, 0.1\}$	categorical
Log-duration transform	{true, false}	categorical
DeepHit only		
Ranking weight α	$[0, 1]$	uniform
Ranking scale σ	$\{0.1, 0.25, 1, 5, 10\}$	categorical
Num. durations	$\{10, 25, 50, 100\}$	categorical
Deep Survival Machines (DSM)		
Num. experts k	$\{4, 6, 8\}$	categorical
Discount factor	$\{0.5, 0.75, 1.0\}$	categorical
Hidden layers	$\{1, 2\}$	categorical
Nodes per layer	$\{50, 100\}$	categorical
Output distribution	{LogNormal, Weibull}	categorical
Learning rate	$\{10^{-3}, 10^{-4}\}$	categorical
Batch size	100	constant
SuMo-Net		
Covariate sub-net depth	$\{1, 2, 4\}$	categorical
Time sub-net depth	$\{1, 2, 4\}$	categorical
Nodes per layer	$\{8, 16, 32\}$	categorical
Dropout	$\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$	categorical
Weight decay	$\{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.4\}$	categorical
Batch size	$\{25, 50, 100, 250\}$	categorical
Learning rate	10^{-3}	constant
CoxKAN		
Same ranges used in the authors' reproducibility code Knottenbelt et al. (2025).		

Table 5: Sieve membership (C5) of the simulation estimators: maximum over $R = 20$ fits of $\|\hat{\theta}_i\|_\infty$ against C_θ, and of $\sup|\hat{g}_i|$ against M'. Additionally, the number of replicates with $\sup|\hat{g}_i| > M'$. Bold marks the presence of fits that exceed the sieve.

DGP	quantity	threshold	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}
DGP-1	$\max_i \|\hat{\theta}_i\|_\infty$	12.00	4.30	3.04	3.67	2.82	2.70	2.65	1.45	0.92
	$\max_i \sup\|\hat{g}_i\|$	6.45	4.95	4.37	4.14	3.84	3.61	3.50	3.34	3.33
	$\#\{i : \sup\|\hat{g}_i\| > M'\} / R$		0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/20
DGP-2	$\max_i \|\hat{\theta}_i\|_\infty$	12.00	3.43	3.17	3.46	2.90	3.85	2.44	1.48	0.96
	$\max_i \sup\|\hat{g}_i\|$	4.49	3.88	3.29	2.85	2.73	2.45	2.32	2.31	2.30
	$\#\{i : \sup\|\hat{g}_i\| > M'\} / R$		0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/20
DGP-3	$\max_i \|\hat{\theta}_i\|_\infty$	12.00	**17.55**	**12.42**	8.07	7.52	4.22	3.98	**17.38**	4.45
	$\max_i \sup\|\hat{g}_i\|$	3.43	**10.95**	**8.62**	**8.89**	**5.00**	2.78	2.76	2.27	2.26
	$\#\{i : \sup\|\hat{g}_i\| > M'\} / R$		15/20	12/20	7/20	6/20	0/20	0/20	0/20	0/20
DGP-4	$\max_i \|\hat{\theta}_i\|_\infty$	12.00	11.87	**18.34**	**32.48**	6.66	7.25	**19.46**	**19.67**	**14.61**
	$\max_i \sup\|\hat{g}_i\|$	6.77	**8.11**	**7.91**	**14.83**	5.95	4.41	4.26	**7.82**	4.22
	$\#\{i : \sup\|\hat{g}_i\| > M'\} / R$		2/20	2/20	1/20	0/20	0/20	0/20	1/20	0/20

Table 6: Near-maximisation (C6) of the delivered fits: maximum over $R = 20$ replicates of $\hat{\eta}_n/(n^{-2r/(2r+1)} \log n)$ per DGP and n, where $\hat{\eta}_n$ is the log-likelihood shortfall of the fit against the certified sieve approximant. Every cell lies below $C_\eta = 0.60$.

DGP	r	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}	$\max_n$
DGP-1	2	0	0	0	0	0	0	0.595	0	0.595
DGP-2	1	0	0	0	0	0	0.087	0.085	0	0.087
DGP-3	1	0	0	0	0	0	0	0	0.153	0.153
DGP-4	1	0	0	0	0	0	0	0	0	0.000

E Additional Experiments

E.1 Covariate Dimension

The simulation study sets the ambient dimension to $p = 5$, of which two covariates carry signal, per Appendix C.2. To test the effect of varying the ambient dimension, we repeat the experiment at $p \in \{2, 4, 8, 16, 32, 64, 128\}$, holding the DGP fixed and adding noise covariates, each tied to an active one through a one-factor Gaussian copula at latent $\rho = 0.5$. Per the main experiments we use 30% censoring and $R = 20$ replicates.

Figure 4 shows the resultant convergence curves. The curves stack vertically in order of p at every n, showing that the ambient dimension increases the absolute error in terms of ISE_S. The curves continue to converge over the n range, and the fitted slopes steepen with p. This supports that the added noise covariates impose a finite-sample cost that the estimator absorbs as n grows, and not a change of exponent in the rate.

E.2 Censoring Rate Ablation

The simulation study calibrates the censoring distribution to an overall rate of 30%. To test the effect of heavier censoring we repeat the experiment at 60% and 90%.

Figure 5 shows the resultant convergence curves. At 60% the curves sit above the 30% curves and descend with them. At 90% the error is several times larger and the descent is slower on DGP-1, DGP-2 and DGP-4, with DGP-3 the exception.

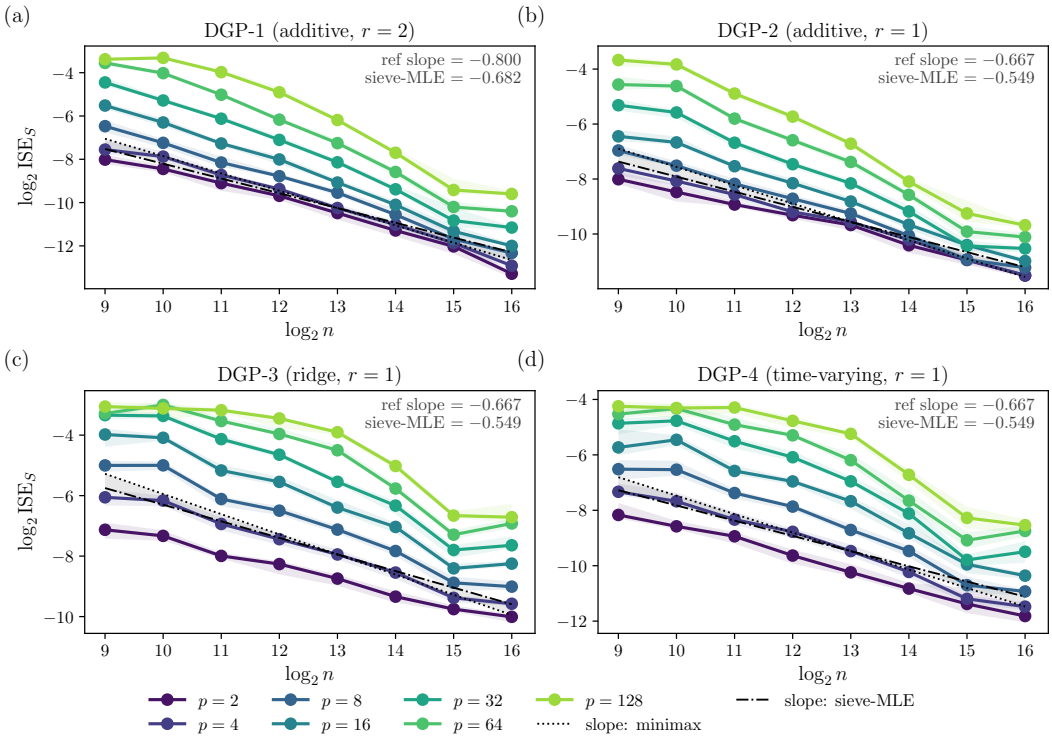


Figure 4: Log-log convergence of ISE_S vs. n for KAPLAN-HR at ambient dimension p, with two active covariates and $p - 2$ noise covariates tied to them through a one-factor Gaussian copula at latent $\rho = 0.5$. Median and IQR over 20 repetitions, compared to the minimax slope $-2r/(2r + 1)$ and the Corollary 2 slope (shown only for $p = 4$).

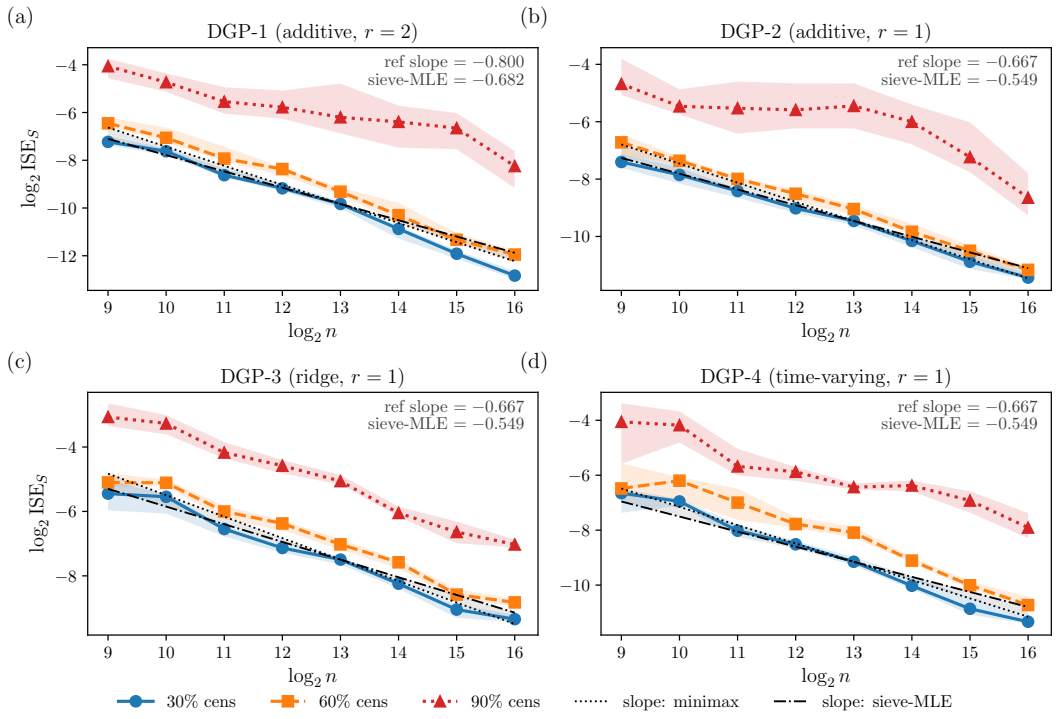


Figure 5: Log-log convergence of ISE_S vs. n for KAPLAN-HR at overall censoring rates of 30%, 60% and 90%. Median and IQR over $R = 20$ repetitions. The 30% curve is the configuration of Fig. 2, and the reference slopes are anchored to it.

E.3 SiLU Ablation

The implemented KAN edges are B-splines plus a SiLU residual, per Eq. (3). To measure the residual's contribution, we repeat the simulation and benchmark experiments with it removed.

Figure 6 repeats Fig. 2 with the ablated estimator overlaid. The two curves track each other on all four DGPs, with no systematic gap at any sample size. Tables 7 to 9 report the benchmark performance after refitting without the residual at the selected KAPLAN-HR hyperparameters. We find that the residual's contribution is concentrated in calibration, where KAPLAN-HR attains the lower ICI on five of six datasets. On C-TD and IBS it leads on four of six, with the two significant C-TD gaps on NWTCO and SUPPORT.

Table 7: C-TD (↑, mean over 10 splits) with and without the SiLU residual. The residual-free model is refit at the KAPLAN-HR hyperparameters, so that only `base_fun` differs. $^*p < 0.05$ vs. KAPLAN-HR (paired Wilcoxon over splits, Holm-corrected across datasets).

Arm	METABRIC	MIMIC-III	RotGBSG	FLCHAIN	NWTCO	SUPPORT
KAPLAN-HR	**0.6690**	**0.8136**	0.6769	0.7935	**0.7223**	**0.6665**
−SiLU	0.6592	0.8091	**0.6792**	**0.7946**	0.7140*	0.6553*

F Real-World Clinical Dataset Results

Table 10 reports the tuned hyperparameters for KAPLAN-HR. Tables 11 to 14 present the complete benchmark results.

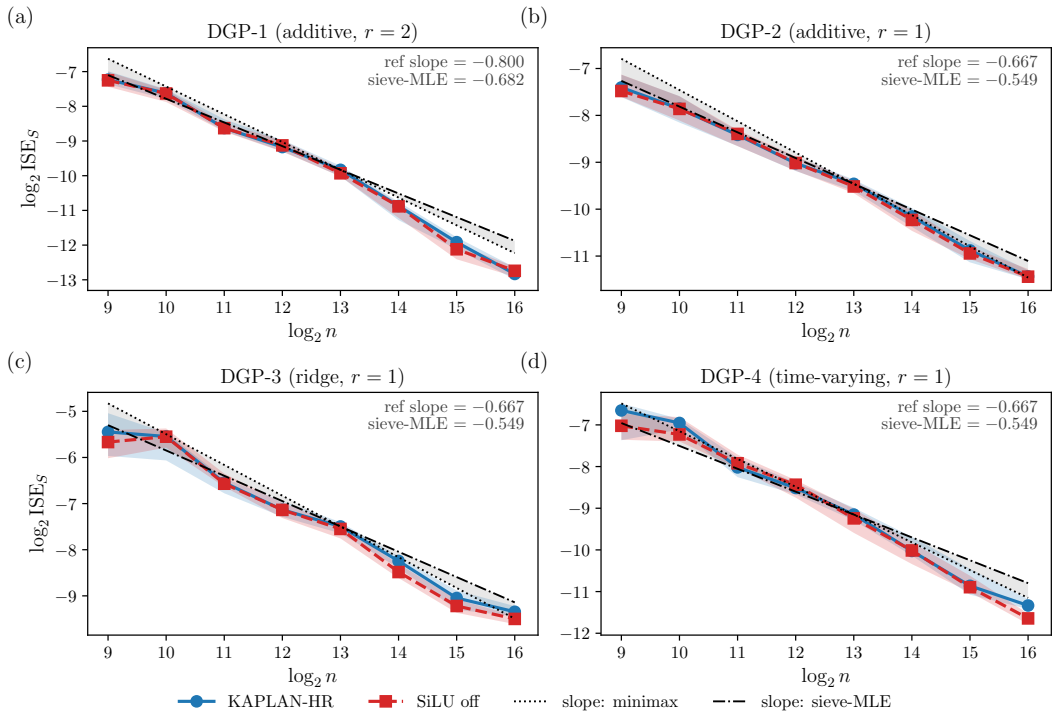


Figure 6: Log-log convergence of ISE_S vs. n for KAPLAN-HR with and without the SiLU residual. Median and IQR over $R = 20$ repetitions. Reference slopes as in Fig. 2.

Table 8: IBS (↓, mean over 10 splits) with and without the SiLU residual. The residual-free model is refit at the KAPLAN-HR hyperparameters, so that only `base_fun` differs. $^*p < 0.05$ vs. KAPLAN-HR (paired Wilcoxon over splits, Holm-corrected across datasets).

Arm	METABRIC	MIMIC-III	RotGBSG	FLCHAIN	NWTCO	SUPPORT
KAPLAN-HR	**0.1603**	**0.1742**	0.1800	0.0983	**0.1024**	**0.1837**
−SiLU	0.1716*	0.1760	**0.1797**	**0.0979**	0.1039	0.1857*

Table 9: ICI (↓, mean over 10 splits) with and without the SiLU residual. The residual-free model is refit at the KAPLAN-HR hyperparameters, so that only `base_fun` differs. $^*p < 0.05$ vs. KAPLAN-HR (paired Wilcoxon over splits, Holm-corrected across datasets).

Arm	METABRIC	MIMIC-III	RotGBSG	FLCHAIN	NWTCO	SUPPORT
KAPLAN-HR	**0.0231**	**0.0086**	**0.0280**	0.0186	**0.0334**	**0.0251**
−SiLU	0.0759*	0.0104	0.0288	**0.0182**	0.0371	0.0344

Table 10: Tuned hyperparameters for KAPLAN-HR per dataset, selected on validation integrated Brier score. LR = Adam learning rate; J_n = spline knot count; m = spline order; D = number of KAN layers; Q = hidden layer dimension; Grid upd. = spline grid update frequency; λ_{ℓ_1} = ℓ_1 penalty on spline coefficients and their first differences.

Dataset	LR	J_n	m	D	Q	Grid upd.	λ_{ℓ_1}
METABRIC	2.38×10^{-4}	2	5	3	16	20	4.01×10^{-4}
MIMIC-III	3.42×10^{-3}	4	3	3	5	15	6.56×10^{-4}
RotGBSG	1.95×10^{-4}	3	3	2	32	10	6.07×10^{-4}
FLCHAIN	2.24×10^{-4}	4	4	2	5	15	9.24×10^{-4}
NWTCO	7.71×10^{-4}	3	3	2	16	20	9.69×10^{-4}
SUPPORT	2.00×10^{-4}	5	3	3	5	15	8.32×10^{-4}

Table 11: Time-dependent concordance (C-TD, ↑, mean ± std) benchmark results.

Model	METABRIC	MIMIC-III	RotGBSG	FLCHAIN	NWTCO	SUPPORT
CoxPH	.64604 ± .01828	.73747 ± .01064	.65363 ± .01658	.70935 ± .16190	.71277 ± .02291	.66087 ± .01111
DeepSurv	.65652 ± .02116	.75626 ± .01128	.67253 ± .01732	.79137 ± .01215	.64814 ± .10781	.66409 ± .01081
CoxTime	.67022 ± .01656	.78593 ± .01023	.66925 ± .01332	.78604 ± .01337	.70753 ± .02164	.65493 ± .00932
DeepHit	**.67357 ± .01948**	.80840 ± .01192	.66855 ± .01777	.77604 ± .00888	.71320 ± .02338	.65966 ± .00990
DSM	.66914 ± .02440	.80324 ± .00906	.67499 ± .01442	.78954 ± .01176	.71332 ± .02050	.66544 ± .01212
GAM	.63576 ± .01016	.63567 ± .07206	.66797 ± .01382	.66891 ± .03373	.60283 ± .02323	.64581 ± .00959
HARE	.66606 ± .01524	.78628 ± .01214	.66891 ± .01489	.77820 ± .05352	.71461 ± .02140	.66411 ± .01265
CoxKAN	.65710 ± .01776	.76325 ± .00809	.66377 ± .01777	**.79664 ± .01228**	.70802 ± .02516	.66303 ± .01024
SuMo-Net	.66228 ± .01682	.79652 ± .01296	.67089 ± .01467	.79194 ± .00994	.71855 ± .01535	.64312 ± .00941
KAPLAN-HR	.66895 ± .01618	**.81362 ± .01005**	**.67692 ± .01532**	.79349 ± .01292	**.72228 ± .02198**	**.66646 ± .00937**

Table 12: Integrated Brier Score (IBS, ↓, mean ± std) benchmark results.

Model	METABRIC	MIMIC-III	RotGBSG	FLCHAIN	NWTCO	SUPPORT
CoxPH	.16057 ± .00739	.20115 ± .02953	.18576 ± .00755	.11678 ± .03318	.10547 ± .00331	.18442 ± .00608
DeepSurv	.16069 ± .00662	.20359 ± .02834	**.17975 ± .00758**	.10101 ± .00336	.10936 ± .00680	**.18321 ± .00616**
CoxTime	.16038 ± .00699	.18813 ± .02913	.18131 ± .00682	.10979 ± .01681	.10603 ± .00560	.18687 ± .00634
DeepHit	**.15950 ± .00659**	.18174 ± .03097	.18404 ± .00745	.10038 ± .00350	**.09592 ± .01012**	.18437 ± .00693
DSM	.16385 ± .00757	.18740 ± .03165	.18165 ± .00621	**.09807 ± .00278**	.09806 ± .00429	.18459 ± .00614
GAM	.17900 ± .00774	.20227 ± .03103	.18278 ± .00463	.11990 ± .00351	.10973 ± .00259	.20065 ± .00620
HARE	.16435 ± .00691	.18969 ± .03202	.18142 ± .00725	.12945 ± .09561	.10163 ± .00661	.18567 ± .00713
CoxKAN	.16030 ± .00688	.18946 ± .02955	.18312 ± .00768	.09830 ± .00272	.10459 ± .00376	.18400 ± .00663
SuMo-Net	.16303 ± .00519	.18101 ± .03618	.18085 ± .00794	.09997 ± .00291	.10167 ± .00378	.18364 ± .00697
KAPLAN-HR	.16032 ± .00558	**.17415 ± .02414**	.17999 ± .00766	.09834 ± .00251	.10240 ± .00392	.18370 ± .00639

Table 13: ICI at median event time (ICI Q2, ↓, mean ± std) benchmark results.

Model	METABRIC	MIMIC-III	RotGBSG	FLCHAIN	NWTCO	SUPPORT
CoxPH	.02470 ± .01066	.00829 ± .00359	.03555 ± .01333	.04795 ± .05869	.02888 ± .00659	.02929 ± .00828
DeepSurv	.02602 ± .01260	.01576 ± .00446	.03261 ± .00814	.01799 ± .00595	.01885 ± .00705	.03269 ± .00937
CoxTime	.02330 ± .00805	.01898 ± .00303	**.02793 ± .00882**	.03816 ± .04466	.04206 ± .00760	.02874 ± .01023
DeepHit	.03026 ± .01267	.00832 ± .00171	.03234 ± .00993	.02532 ± .01004	.03667 ± .01233	.02378 ± .00519
DSM	.03481 ± .01251	.00969 ± .00341	.04192 ± .01410	**.01445 ± .00610**	.02949 ± .00893	.03938 ± .00466
GAM	.04463 ± .01367	.03695 ± .02480	.04993 ± .01954	.02275 ± .01180	**.01544 ± .01540**	.04225 ± .00757
HARE	.03199 ± .00868	**.00790 ± .00197**	.03032 ± .00831	.05605 ± .12191	.04403 ± .02169	.03051 ± .00857
CoxKAN	.02784 ± .01247	.00897 ± .00194	.02799 ± .00621	.02109 ± .00673	.03353 ± .01097	.02945 ± .00840
SuMo-Net	.05589 ± .01949	.00890 ± .00180	.03813 ± .01796	.03141 ± .00623	.03047 ± .00991	**.02148 ± .00402**
KAPLAN-HR	**.02310 ± .00585**	.00857 ± .00222	.02803 ± .00678	.01862 ± .00562	.03338 ± .00836	.02512 ± .00587

Table 14: D-calibration pass count (k/n) benchmark results. Cells with $k = n$ (all splits pass at $p > 0.05$) in bold.

Model	METABRIC	MIMIC-III	RotGBSG	FLCHAIN	NWTCO	SUPPORT
CoxPH	**10/10**	**10/10**	**10/10**	8/10	**10/10**	5/10
DeepSurv	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**	4/10
CoxTime	**10/10**	**10/10**	**10/10**	8/10	**10/10**	4/10
DeepHit	8/10	**10/10**	1/10	8/10	**10/10**	0/10
DSM	**10/10**	**10/10**	7/10	**10/10**	**10/10**	0/10
GAM	**10/10**	4/10	9/10	6/10	**10/10**	0/10
HARE	**10/10**	**10/10**	**10/10**	9/10	8/10	9/10
CoxKAN	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**	5/10
SuMo-Net	8/10	**10/10**	8/10	**10/10**	**10/10**	6/10
KAPLAN-HR	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**	5/10