WHEN THE LOSS STOPS TRACKING THE METRIC: VALUE-MONOTONICITY IN DEEP SURVIVAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A validation loss is routinely interpreted as if a lower value implied a better model. Deep survival prediction makes the mismatch stark, since models there are commonly trained with likelihood objectives, evaluated by the concordance index (C-index), and selected on the numerical value of the validation loss. Standard surrogate theory does not justify that interpretation. Fisher consistency constrains the minimizer, not the surrogate value at the intermediate points where early stopping and checkpoint selection occur. We formalize the missing property as *value-monotonicity*, requiring the loss to fall as the metric rises, to within a stated tolerance, along a training trajectory rather than only at the optimum. We prove that the Cox partial likelihood admits no finite such tolerance on a score class closed under rescaling, because rescaling leaves the C-index unchanged while the likelihood varies without bound, and the discrete-time likelihoods decouple by the same mechanism. The sigmoid concordance loss (SCL), in contrast, approximates one minus the C-index with a two-sided, computable tolerance, so a decrease beyond twice that tolerance certifies a concordance gain. Across eighteen datasets and four modalities the absolute rank correlation between validation SCL and validation C-index averages 0.97 and never falls below 0.91, against 0.37 to 0.46 for the likelihood losses. Replacing C-index-based checkpointing by loss-based checkpointing moves the SCL's test C-index by at most 0.006, against up to 0.064 for the baselines. The SCL also retains competitive discrimination, with the best mean within-dataset rank, 2.00, among the six losses run on every dataset, although the leading methods are not statistically separated. The contribution is therefore not improved discrimination, but a loss whose numerical value remains interpretable as a proxy for the evaluation metric during training.

1 INTRODUCTION

A falling validation loss is read as a better model, and that reading drives early stopping, checkpoint selection and the comparison of runs. Nothing in surrogate theory licenses it. Fisher consistency constrains the minimizer, not the value at the intermediate points where those decisions are taken. Deep survival prediction makes the gap concrete, because the objective and the metric there are not the same kind of quantity.

Survival models estimate the time from a baseline observation to a clinical event, from data in which many subjects are censored and their event is not yet observed. A central evaluation target is risk ordering: whether subjects who experience the event sooner receive higher predicted risk. The concordance index (C-index) (Harrell et al., 1982; Uno et al., 2011) measures exactly that, and it depends on the model only through the ordering it induces. Deep survival models differ in architecture and in how they represent time, from Cox-parameterized networks (Katzman et al., 2018) through discrete-time likelihoods (Yu et al., 2011; Lee et al., 2018; Gensheimer & Narasimhan, 2019) to the multiple-instance aggregators used in pathology (Chen et al., 2022; Jaume et al., 2024), and they share one choice at the level of the objective. They are trained by maximizing a likelihood and evaluated by the C-index.

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.

A likelihood is optimized at a single conditional distribution, while the C-index is a function of the ordering alone and is unchanged by any increasing rescaling of the risk score. The usual justification is that the two agree at the optimum and that the C-index can be recomputed on a validation set whenever needed. Neither addresses how the loss is actually used. For the survival likelihoods the reading is not even well posed, since their values shift with the censoring rate, the batch size and the risk-set size.

We show that this failure is structural for likelihood losses but avoidable. We then study the sigmoid concordance loss (SCL), the mean over comparable pairs of a smoothed misordering indicator, as an objective that avoids it.

Contributions.

- We name and formalize *value-monotonicity*, and separate it from Fisher consistency and from one-sided surrogate bounds.
- We prove that no finite tolerance exists for the Cox partial likelihood on a class closed under rescaling, whereas the SCL admits a tolerance of order τ.
- We give the SCL a two-sided, computable tolerance: a decrease beyond twice it certifies a concordance gain.
- We measure the property on eighteen datasets and four modalities, and characterize its empirical limits.

The scope is narrow by design. The SCL matches the field on discrimination rather than dominating it, and of the twenty paired imaging comparisons against a published loss four resolve, two of them the other way. Monitoring a Cox-trained model with the SCL value recovers most of the selection agreement, so the benefit lies in which value is read rather than in which loss was trained. The question this paper answers is when the value of a training objective can be read as the metric it is meant to optimize.

2 Related work

Optimum alignment against value alignment. The gap between training loss and evaluation metric is recognized across machine learning (Huang et al., 2019), and the statistical response has been *consistency*. Bartlett et al. (2006) and Steinwart (2007) characterize when minimizing a surrogate recovers the metric optimum, Gao & Zhou (2015) identify which pairwise surrogates are consistent for AUC, and Elgui et al. (2023) give consistency results specific to the concordance index. Non-asymptotic $\mathcal{H}$-consistency bounds (Awasthi et al., 2022) relate the excess metric risk to the excess surrogate risk uniformly over the adopted hypothesis class, but they remain *one-sided*. None of this says the loss *value* orders models the way the metric does before convergence, which is the question we ask.

Concordance surrogates. Prior concordance surrogates were designed for optimization, their guarantees showing that optimizing the surrogate improves or bounds concordance. We ask instead that the value of the surrogate approximate the value of the metric, because early stopping and checkpoint selection act on values. Smoothing the concordance index has a long history. Yan et al. (2003) and Herschtal & Raskutti (2004) smooth the Wilcoxon-Mann-Whitney statistic for binary classification, where there is no censoring. Steck et al. (2007) give the survival version, defining the metric over the order graph whose edges are exactly the comparable pairs we use and optimizing two concave *lower bounds* on it; later deep survival work adds smooth pairwise ranking terms to likelihood objectives, among them the exponential terms of DeepHit (Lee et al., 2018) and TripleSurv (Zhang et al., 2024) that we reimplement here. Our claim differs in the direction of the guarantee. Steck et al. (2007) consider the plain sigmoid and set it aside, observing that it approximates the indicator but is not a bound, and take the log-sigmoid because it is one. A lower bound certifies that the metric is at least the surrogate and says nothing about whether the surrogate's value moves with the metric's, which is what a practitioner relies on when reading a loss curve. Theorem 1 is two-sided, and two-sidedness is what value-monotonicity requires. Learning-to-rank (Burges et al., 2005; Qin et al., 2010) and deep AUC maximization (Yuan et al., 2022) optimize objectives of the same family without censoring. The closest work in spirit is relational surrogate loss learning (Huang et al.,

2022), which argues that a surrogate need only rank models the way the metric does and, to that end, *learns* a surrogate network offline by maximizing its rank correlation with the metric. We share that intuition and differ in three ways. We analyze an analytic loss rather than a learned black box, we prove the correlation rather than fitting it, and we read the loss as a monitoring and selection signal. We compare against it in Section 5, where a faithful adaptation to the concordance index does not attain reliable value-monotonicity because its learned term collapses alone and must be combined with a likelihood whose value our theory shows decouples.

Separating ranking from calibration. The closest recent method, GRAFT (Ashhad et al., 2026), shares our starting point that survival likelihoods conflate ranking with calibration, and separates the two by optimizing what its authors call a C-index-aligned ranking loss, calibrating afterwards. Its Equation 10 defines the negative normalized covariance of the soft ranks of Blondel et al. (2020) and the ranks of event times imputed from local Kaplan-Meier estimators, which is a differentiable Spearman correlation rather than a pairwise concordance surrogate. We share the goal and differ in the mechanism. The distinction suggests a tracking mismatch, since soft Spearman targets a different rank statistic from the concordance index, and a loss aligned with the former need not be aligned with the latter. We run it under our protocol and report the comparison in Section 5 and Appendix B.18, where its discrimination is close to the SCL's while its value tracks the C-index far less closely, consistent with the distinction between the two rank statistics.

Objective-metric alignment in survival. A line of work argues that the standard survival likelihood does not optimize the criterion one cares about and proposes objectives targeting other desiderata, without necessarily improving discrimination, among them explicit calibration (Goldstein et al., 2020) and game-theoretic proper losses (Han et al., 2021). Deep survival models otherwise differ mainly in architecture and time representation (Gensheimer & Narasimhan, 2019; Kvamme et al., 2019; Tang et al., 2022; Nagpal et al., 2021; Zhang et al., 2024) or in the pathology aggregator (Chen et al., 2022; 2021; Jaume et al., 2024), but share the likelihood-trained, C-index-evaluated design that we analyze. We add the training-trajectory property and the accompanying negative result for likelihood losses.

3 VALUE-MONOTONICITY

Setup. We observe right-censored data $\{(x_n, t_n, \delta_n)\}$ where $\delta_n = 1$ marks an observed event. A model emits a scalar risk $f(x) \in \mathbb{R}$, higher meaning shorter survival. A pair (i, j) is *comparable* if $t_i < t_j$ and $\delta_i = 1$. The concordance index is the probability the model orders a comparable pair correctly, $C(f) = \Pr\big(f(x_i) > f(x_j) \mid t_i < t_j,\ \delta_i = 1\big)$, with ties counted as one half; Harrell's and Uno's variants differ only in how pairs are weighted (Harrell et al., 1982; Uno et al., 2011) and both depend on f through its ordering alone.

Fisher consistency is not enough. Fisher consistency asks that $\arg\min_f L(f) = \arg\max_f C(f)$, which is a statement about the optimum. When the loss is used to monitor training, stop early, and select a checkpoint, the relevant question is whether the value of $L(f)$ moves with $C(f)$ at every point of the trajectory. Informally, a loss is value-monotone if models with meaningfully lower loss also have higher concordance, along the trajectory and not only at the optimum. The property is approximate: it allows a bounded deviation rather than requiring an exact monotone map between loss and metric.

Definition 1 (Value-monotonicity)**.** *A loss L is ε-*value-monotone *in the C-index on a score class $\mathcal{F}$ if $L(f) = h(C(f)) + r(f)$ for a strictly decreasing h and a residual r with $\sup_{f\in\mathcal{F}} |r(f)| \le \varepsilon$, so that a loss decrease beyond 2ε implies a higher C-index throughout training rather than only in the limit. Every loss admits such a decomposition for some ε, so the definition is informative only when ε is small relative to the loss changes observed during training. Then 2ε is the resolution at which the loss curve can be read, a wider decrease certifying a concordance gain (Proposition 3) and a narrower one certifying nothing. On a class closed under upward rescaling the SCL keeps ε of order the temperature, while the Cox partial likelihood admits no finite ε at all (Proposition 2).*

All results rest on one structural fact.

Lemma 1 (Rank-invariance)**.** *For every strictly increasing φ, $C(\varphi \circ f) = C(f)$. Hence every class of scores sharing an ordering lies inside a level set of C, and any direction that reshapes a score*

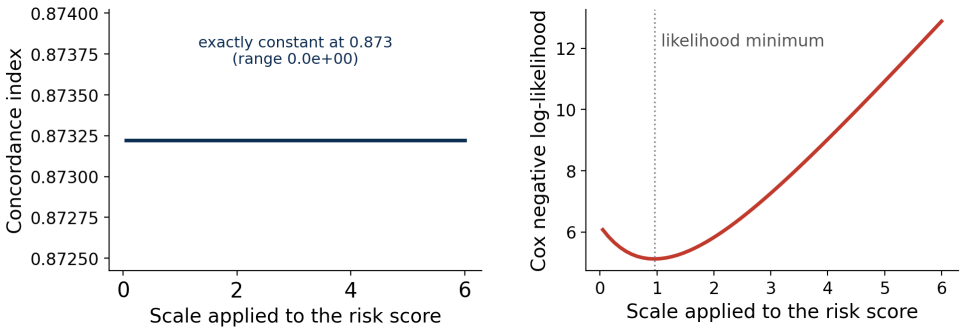


Figure 1: Proposition 1 on synthetic Cox data ($n = 800$, five covariates, 30 per cent censoring). Rescaling preserves rank, so the concordance index is constant at 0.8732 to four decimal places (left) while the Cox negative log partial likelihood runs from 5.12 to 12.88, a factor of two and a half (right). The left panel is plotted over a span of 0.0016 of concordance index, so the flatness is not an artefact of the range.

without changing its order leaves C stationary to first order. The inclusion is strict: scores with different orderings can still share a concordance value, so a level set is a union of such classes.

What the argument needs of a loss is only that it is *not* invariant under strictly increasing reparameterizations of the score. Any objective minimized at a single conditional law rather than at an order-equivalence class has this property, and so does the Cox partial likelihood, which profiles the baseline hazard out and is therefore not a proper scoring rule for the full conditional distribution but is still scale-sensitive: $L_{\mathrm{cox}}(\alpha f) \neq L_{\mathrm{cox}}(f)$. The always-available instance is rescaling.

Proposition 1 (Rescaling decoupling)**.** *Let L_{cox} be the negative Cox partial log-likelihood and f non-constant on some risk set. Then $C(\alpha f) = C(f)$ for all $\alpha > 0$, so $h = f$ is exactly rank-preserving with $dC(f)[f] = 0$; and $\phi(\alpha) = L_{\mathrm{cox}}(\alpha f)$ is strictly convex,*

$$\phi''(\alpha) = \frac{1}{|D|} \sum_{i \in D} \mathrm{Var}_{p_\alpha^{(i)}}[f] > 0, \qquad p_\alpha^{(i)} = \mathrm{softmax}(\alpha f) \text{ over the risk set } R(i).$$

Thus $\phi'(1) \neq 0$ unless $\alpha = 1$ is the unique stationary point of ϕ: rescaling is an explicit direction along which the likelihood strictly changes while the C-index is exactly constant, which Figure 1 shows on synthetic Cox data. That exception is not a technicality, since it is what happens at a training optimum on the same data, which is why Observation 1 is stated for the monitored held-out loss.

Proof sketch. $\alpha \mapsto \alpha f$ preserves ordering, so $C(\alpha f)$ is constant by Lemma 1. The likelihood is a sum of log-sum-exp terms in αf, each convex in α with second derivative the softmax variance of f over its risk set, which is positive whenever f is non-constant there; so ϕ is strictly convex and ϕ' vanishes at one α only. Full argument in Appendix A. □

Proposition 1 exhibits the direction; the following says why the same mechanism does not level the two families, since it applies to the SCL at fixed τ as well.

Proposition 2 (The separation is in the size of the tolerance)**.** *Fix f_0 and let $\mathcal{F}_A = \{\alpha f_0 : 1 \leq \alpha \leq A\}$, on which C is constant by Lemma 1. If the gap law of f_0 satisfies the non-atomic-gap condition A1 of Appendix A, with constants (B, η) bounding the density of comparable-pair score gaps near zero, then so does every αf_0 with $\alpha \geq 1$, so $\varepsilon_\tau(\mathcal{F}_A) \leq 2B\tau + O(e^{-\eta/\tau})$* uniformly in A. *If instead f_0 misorders a pair inside some risk set, then $\sup_{\alpha \leq A} L_{\mathrm{cox}}(\alpha f_0) \to \infty$ as $A \to \infty$, so no finite tolerance exists for the Cox partial likelihood on $\mathcal{F}_A$.*

Rescaling upward makes comparable pairs *less* nearly tied, which is why the SCL's tolerance does not degrade while each Cox term grows linearly in the scale; Appendix A confirms both numerically.

4 THE SIGMOID CONCORDANCE LOSS

The loss. For a minibatch with comparable-pair set $\mathcal{P}$, the sigmoid concordance loss (SCL) is the mean smoothed misordering rate,

$$L_\tau(f) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \sigma\Big(-\frac{f(x_i) - f(x_j)}{\tau}\Big), \qquad \sigma(u) = \frac{1}{1+e^{-u}}, \tag{1}$$

with temperature $\tau > 0$ fixed at 0.1 in all experiments. As $\tau \to 0$ the sigmoid approaches the misordering indicator and $L_\tau(f) \to 1 - C(f)$, so its value *estimates* one minus the metric rather than merely sharing its optimum (Theorem 1). Because the loss depends on score differences only through their smoothed sign, it is asymptotically invariant to the magnitude of the scores; this is why we use the sigmoid rather than a margin surrogate such as the squared hinge, whose value changes under order-preserving rescaling and is therefore not value-monotone even though it is AUC-consistent.

Minibatch estimator and cost. The loss in equation 1 is a within-minibatch average over comparable pairs, so no global risk set is needed and it is computed with a single pairwise comparison of the B scores in a batch. Forming the comparable-pair mask $\{(i,j) : t_i < t_j,\ \delta_i = 1\}$ and the smoothed differences costs $O(B^2)$ time and memory, negligible next to the backbone forward pass for the batch sizes we use ($B \le 256$). The minibatch value is therefore an estimate of the full-data value rather than a function of the batch size, which the Cox partial likelihood is not, since its risk sets grow with B; Appendix B.17 measures both. Because the summand depends only on score differences, the loss is invariant to any additive shift of f and, as $\tau \to 0$, to any positive rescaling, the two order-preserving moves that Section 3 shows a likelihood cannot ignore. Algorithm 1 in Appendix B gives the full step.

A convex special case. For a linear model the squared-margin relaxation of equation 1 is convex and defines a survival ranking SVM solvable to global optimality, which Appendix B compares against the deep loss.

The hybrid variant. To test whether value-tracking survives any admixture of a likelihood term, we also study a *hybrid* loss $L_\tau(f) + \lambda\, L_{\mathrm{haz}}(g)$, $\lambda > 0$, in which the network emits a scalar risk f for the concordance term and K hazard logits g for a logistic-hazard negative log-likelihood, so the two terms act on separate output heads and the backbone, not only the loss, differs from the other methods. This is a diagnostic, not a proposed method, and the theory does not predict its outcome, since Theorem 2 gives no magnitude and covers the SCL too.

Theorem 1 (Value-monotonicity of the SCL)**.** *Under a non-atomic-gap assumption on the score differences, $L_\tau(f) = (1 - C(f)) + O(\tau)$ as $\tau \to 0$, uniformly over the score class, with constant at most $2B$ where B bounds the density of the score gap near zero. Write $\varepsilon_\tau = \sup_f |L_\tau(f) - (1 - C(f))| \le 2B\tau + o(\tau)$.*

The restriction to a score class is not cosmetic. The downward orbit $\{\alpha f : \alpha \to 0\}$ leaves every such class, because $L_\tau(\alpha f) = L_{\tau/\alpha}(f) \to \frac{1}{2}$ while C is unchanged, so on a perfectly concordant score the approximation error approaches $\frac{1}{2}$. No uniform bound holds over a full orbit at fixed τ, and the alignment we report is alignment along the trajectories our models actually follow.

Proof sketch. For a comparable pair the summand is $\sigma(-Z)$ with $Z = (f(x_i) - f(x_j))/\tau$, and $\sigma(-Z) \to \mathbb{1}[f(x_i) < f(x_j)]$ pointwise as $\tau \to 0$, whose expectation is $1 - C(f)$. The sigmoid differs appreciably from the step only in a window of width $O(\tau)$ around a tie, and the bounded gap density caps the mass there, uniformly over the score class. Full proof in Appendix A. □

What the theorem gives. The approximation is two-sided, so the loss value can be read as one minus the metric at a known resolution.

Proposition 3 (Certified improvement)**.** *If $|L_\tau(f) - (1 - C(f))| \le \varepsilon_\tau$ uniformly, then $L_\tau(f') < L_\tau(f) - 2\varepsilon_\tau$ implies $C(f') > C(f)$.*

A decrease larger than $2\varepsilon_\tau$ certifies a gain and a smaller one certifies nothing. The tolerance is computable without knowing the gap law: Theorem 1 provides a bound that can be evaluated directly

from the validation predictions. A practitioner can therefore apply the certificate without further assumptions.

Where the guarantee weakens. At a fixed τ the approximation is not an order isomorphism. It weakens when many comparable-pair score gaps are small relative to τ, since the sigmoid then never saturates and the loss drifts toward $1/2$ whatever the ordering. A three-point counterexample in Appendix A puts a perfectly concordant score at $L_\tau = 0.467$ beside a discordant one at $L_\tau = 0.333$. Since $L_\tau(\alpha f) = L_{\tau/\alpha}(f)$, that is the downward rescaling orbit, and the SCL is rescaling-invariant only as $\tau \to 0$. The failure mode is therefore near ties, which a practitioner can check in one pass over the validation predictions with the comparable-pair gap diagnostic of Appendix B.10. What we measure is alignment in the regime our models occupy rather than a guarantee over every score, and it is structural there, since to leading order the loss is an affine image of the metric and Section 3 shows a likelihood cannot be.

5 Experiments

Protocol. Four modalities and eighteen datasets: ten tabular clinical and genomic cohorts, two CT/PET cohorts (Saeed et al., 2025; 2026; Simpson et al., 2024), two MRI glioma cohorts (Calabrese et al., 2022; Bakas et al., 2022), and four TCGA pathology cohorts. Every dataset uses event-stratified five-fold cross-validation with a held-out validation fold, pooled out-of-fold predictions, and ten fold partitions. Within a modality all neural methods share the same backbone and head architecture, differing only in the loss and in the output dimensionality it requires, K logits for the discrete-time methods and one for the scalar-risk ones. No hyperparameter search is performed for any method. Baselines are the Cox partial likelihood, MTLR, DeepHit, TripleSurv, a squared-hinge margin loss, and a random survival forest (Ishwaran et al., 2008), with GRAFT, ReLoss, X-CAL and inverse-weighted survival games added in Appendix B.6. TripleSurv is reimplemented, no public code being available, with its ranking term entering the objective in the direction the paper's Equation 6 gives; Appendix B.7 records that we first read the sign the other way and what the error cost the baseline. Appendix B gives the full protocol and the per-dataset tables, and Appendix B.14 measures what the stored matrices' cohort-level standardization costs (Table A16).

Loss-metric decoupling on a training trajectory (Figure 2). We use SUPPORT because it is the largest tabular cohort, 9105 patients, which gives the smoothest training trajectories. It also gives a conservative illustration, since its average baseline discrepancy between the two selection rules is 0.002, the smallest of the ten tabular cohorts, against a mean of 0.008. Nine additional cohorts are in Appendix B.20. All six losses reach the same validation C-index plateau here, near 0.88, and their loss curves then behave in six different ways. MTLR and the SCL illustrate the two extremes. MTLR's validation loss keeps falling for another sixty epochs after its C-index peaks, while that C-index decays from 0.885 to 0.868, a rank correlation of -0.72, and the checkpoint its loss selects tests 0.010 below the one the C-index selects; for the SCL the two curves move together at $+1.00$ and both rules stop at epoch 49. The likelihood losses in between say little about the metric either way. This is the behaviour Observation 1 describes. Accordingly the selection cost is small on SUPPORT; Figure 3 quantifies it across all eighteen datasets.

What value alignment buys (Figure 3). Switching the selection rule from validation C-index to validation loss changes the test C-index of the SCL by at most 0.006, on no dataset by more than 0.02. For the likelihood and margin losses the change reaches 0.064 and exceeds 0.02 on up to six datasets, and which datasets those are cannot be known in advance. The signed difference is negative on most cohorts (Table A7), so the rules disagree unpredictably rather than one being uniformly better, and Appendix B.8 offers the validation fold's event count as the most plausible account of the direction. Per dataset the coupling averages 0.97 for the SCL and never falls below 0.91, against 0.37 to 0.46 for the likelihood losses and 0.53 for the squared hinge, which fall below 0.5 on ten to sixteen of the eighteen datasets (Figure A1, Table A2). The SCL never reverses sign between folds, while the others reverse on ten to thirty-one per cent of theirs.

Discrimination is preserved, not improved (Table 1). Across the eighteen datasets the SCL has the best mean within-dataset rank of the six losses run on every one, 2.00 against 3.06 for the next,

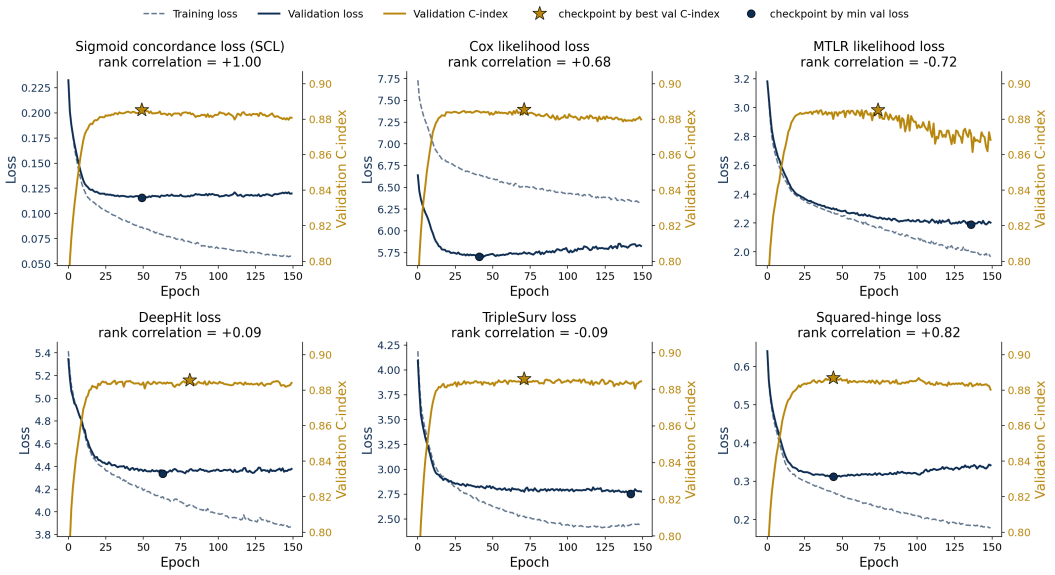


Figure 2: Training and validation trajectories on SUPPORT: training loss (dashed), validation loss (solid), validation C-index (gold, right axis). Above each panel is the rank correlation between negative validation loss and validation C-index over the second half of training. Gold star and navy circle mark the checkpoints chosen by validation C-index and by validation loss. They coincide for the SCL but separate for the other losses, whose loss keeps moving after the C-index has plateaued. MTLR is the sharpest case, its validation loss falling throughout while its C-index decays.

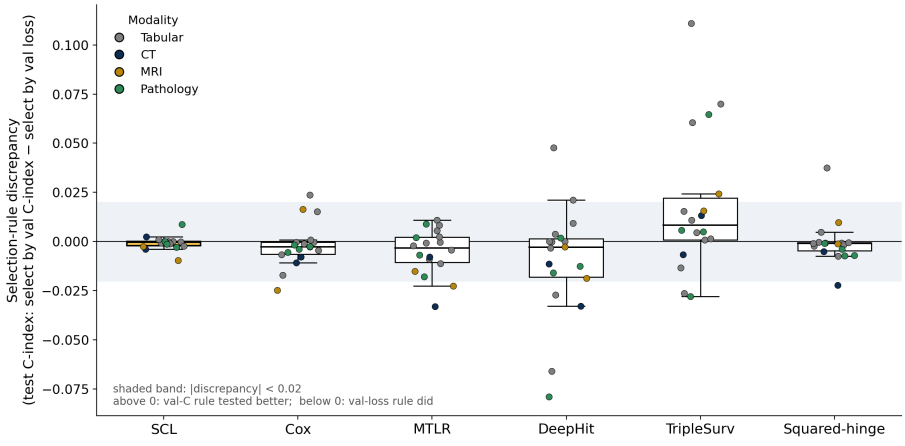


Figure 3: Discrepancy between the two checkpoint-selection rules, one point per dataset, coloured by modality: test C-index of the checkpoint chosen by validation C-index minus that of the checkpoint chosen by validation loss. A negative value means the loss-selected checkpoint tested better, so this measures divergence, not accuracy lost. Shaded band: absolute discrepancy below 0.02. The SCL stays within 0.006 on every dataset and the squared hinge within 0.016; the four likelihood losses leave the band on one to six datasets each and reach 0.064.

and it holds that place in 0.998 of 5000 resamples of the datasets. The margin still does not separate it, since 1.06 is below the critical difference of 1.78 at eighteen datasets (Table A1). It is not uniformly ahead either. A random survival forest matches it on the tabular pool, and among the imaging cohorts it places second twice and fourth twice. Under the by-loss rule the leading four losses fall within 0.23 of one another. Neither ranking carries the argument, since which way the by-loss rule falls on a cohort tracks that cohort's validation event count more closely than the loss (Appendix B.8); what separates the losses is the size of the disagreement between the two rules

Table 1: Discrimination by modality, blocks ordered from the pool with the most datasets to the fewest. Every cell comes from the benchmark runs, so a row's integrated Brier score (IBS, lower better) is computed from the same checkpoint as its C-index. The best value within each modality block is bold for each metric. Per-cohort values are in Tables A3 to A5, ranks with bootstrap intervals in Table A1. A random survival forest reaches 0.735 on the tabular pool, level with the SCL to within 0.0001; it has no training trajectory and so no coupling or selection rule, and is reported per dataset in Table A3.

Modality	Loss	Harrell C-index	Uno C-index	td-AUC	IBS ↓
Tabular 10 datasets	**SCL**	**0.735**	**0.736**	**0.778**	0.146
	Squared-hinge	0.730	0.731	0.772	0.144
	Cox	0.730	0.731	0.771	0.143
	MTLR	0.728	0.729	0.771	**0.142**
	TripleSurv	0.713	0.714	0.755	0.144
	DeepHit	0.698	0.699	0.737	0.192
Pathology 4 datasets	**SCL**	**0.679**	**0.666**	**0.688**	0.143
	Cox	0.670	0.662	0.680	**0.140**
	Squared-hinge	0.669	0.661	0.679	0.142
	TripleSurv	0.661	0.661	0.667	0.143
	MTLR	0.661	0.655	0.670	0.145
	DeepHit	0.651	0.642	0.652	0.158
CT/PET 2 datasets	MTLR	**0.610**	**0.615**	**0.628**	0.161
	Squared-hinge	0.599	0.604	0.611	**0.151**
	SCL	0.595	0.599	0.610	0.153
	TripleSurv	0.594	0.598	0.616	0.156
	Cox	0.591	0.597	0.601	0.152
	DeepHit	0.565	0.568	0.579	0.182
MRI 2 datasets	DeepHit	**0.670**	**0.666**	**0.726**	0.173
	Squared-hinge	0.664	0.661	0.719	0.165
	SCL	0.659	0.656	0.713	0.165
	MTLR	0.656	0.652	0.704	0.171
	Cox	0.654	0.653	0.706	**0.164**
	TripleSurv	0.650	0.647	0.698	0.168

(Table A2). On imaging we bootstrapped the *difference* with patients and fold partitions resampled together (Table A15 in Appendix B.13). Four of the twenty comparisons against a published loss resolve, and two of them go against the SCL: DeepHit and TripleSurv are both ahead of it on UPENN-GBM, by 0.015 and 0.017. It is ahead of DeepHit on HECKTOR by 0.070 and of TripleSurv on UCSF-PDGM by 0.026. Imaging discrimination is unresolved at these cohort sizes, in both directions.

Evidence for the proposed mechanism. Observation 1 names a candidate cause, namely that the likelihood gradient occupies rank-preserving directions, of which rescaling is the canonical one. Measuring the absolute cosine between $\nabla_f L$ and the centred score vector on METABRIC, SUPPORT and GBSG, the Cox gradient carries a rescaling alignment an order of magnitude above the SCL's, 0.04 to 0.24 against 0.003 to 0.008, and it grows from the first half of training to the second on two of the three cohorts and falls on the third. The component is persistent but not dominant and its growth is not universal, so we read it as evidence consistent with the mechanism rather than proof of it.

Training with the loss against reading it (Appendix B.12). Our experiments so far confound training with the SCL and having a concordance-based value to read. We separate the two by evaluating the SCL on the validation predictions of a model trained on another loss and selecting by that value. For the likelihood losses it recovers most of the selection quality: Cox's worst deviation from the reference rule falls from 0.0094 to 0.0054 and MTLR's from 0.0163 to 0.0021. For DeepHit it does not, and the deviation rises from 0.0643 to 0.0724. The reason is a readability condition rather than a property of the training loss. A monitored SCL is informative only when few comparable pairs are nearly tied at the scale of τ, and DeepHit's derived risk has a standard deviation of 0.10, so more

than half of its pairs are. Standardizing that risk before reading the surrogate brings the deviation to 0.0265. Across 72 loss-dataset cells the near-tie fraction predicts the approximation error at rank correlation 0.97 (Appendix B.10).

Ablations and related loss families. Replacing the sigmoid by the squared hinge preserves the C-index but drops the coupling to 0.73 from 0.98 on the ten tabular trajectory captures behind Figure 2, and adding a likelihood anchor collapses it at every weight tested, to 0.39 from 0.97 on the eighteen-dataset measurement of Table A2, while the anchored model's discrimination is indistinguishable from the SCL's, 0.734 against 0.735 on the tabular pool (Figure A3). The anchor reduces the coupling substantially without improving discrimination. Calibration is the obvious reason to add one, and it does not deliver that either. On the same runs and the same checkpoints the anchored model's integrated Brier score is worse than the plain SCL's on seventeen of eighteen cohorts, 0.275 against 0.148 on average. Its survival curve comes from a hazard head that carries weight one half beside a dominant ranking term, rather than from a baseline hazard fitted on the training fold. The statement is about adding an anchor to the SCL, not about likelihoods, which calibrate well alone. Appendix B.6 places the SCL among the alternative loss families, and none of them reaches value alignment by learning or reweighting a loss.

6 Discussion

The results support judging a loss not only by the metric it can reach but by whether its value tracks that metric during training. Two measurements in Section 5 are consistent with the gap being structural rather than incidental, the rescaling alignment of the likelihood gradient and the collapse of the coupling when a likelihood term is added to the SCL. The same invariance mismatch can arise beyond survival, and Theorem 2 is stated in the general form. Whenever a metric is invariant under a transformation of the model outputs but the loss is not, the loss can vary along directions that leave the metric stationary. Its value may therefore move even when the metric does not, however well the two agree at the optimum. The familiar instance is binary classification, where the area under the ROC curve is invariant to any strictly increasing reparameterization of the score and the cross-entropy is not, so a falling validation cross-entropy does not entail a rising area under the curve. Normalized discounted cumulative gain, reciprocal rank, recall at k and mean average precision each factor through an ordering or a thresholded decision that the training loss does not see. This suggests a diagnostic rather than a theorem: before reading a loss curve as a metric curve, ask which group leaves the metric fixed and whether the loss respects it. The remaining question is quantitative, how far the loss can move along such directions, and the theorem gives no magnitude. The SCL satisfies the same hypothesis and stays readable because its tolerance is of order τ, while the Cox likelihood's is unbounded. What made the value readable here was that the loss is, to leading order, an affine image of the metric's own estimator, a strictly stronger property than consistency; we show it suffices and do not claim it is necessary. The alignment comes with a trade-off. A loss built to track an estimator inherits that estimator's variance as well as its location, so the same property that makes the value readable makes it no smoother than the metric it reports. We prove and measure all of this only for survival and the C-index.

The practical consequence is narrower than a recommendation to prefer one selection rule. On a likelihood loss the two rules disagree by up to 0.064 of test C-index, and its direction tracks how many events the validation fold happens to carry rather than anything visible at the time of the choice (Appendix B.8), so neither rule can be recommended over the other. What the SCL offers is that the choice does not arise, its two rules agreeing to within 0.006 on every dataset. Four limitations bound the result. Theorem 1 fails over a rescaling orbit at fixed τ. The finite-sample bound carries $1/p$. The temperature is untuned. And the property belongs to the single-horizon form. Averaging the surrogate over horizons to obtain a survival curve directly erodes the coupling and worsens the calibration it was meant to buy (Table A10 in Appendix B.9).

What the paper leaves behind is a distinction rather than a method. Consistency says where a loss ends up; value-monotonicity says whether its value can be read on the way there, and the two come apart for the objectives survival analysis actually trains on. The SCL is one loss that closes the gap for the concordance index, and the question it raises is which other metric and objective pairs fail the same test.

AI USE STATEMENT

In this work we used generative AI tools for tasks with required disclosure: developing conceptual framing, formulating mathematical claims, supplying and writing proof ingredients, proposing hypotheses, designing experiments, implementing methods, generating the synthetic data used only for Figure 1, and assisting data analysis and the interpretation of results. We did not use them for translation, and the remaining required-disclosure tasks do not apply. Additionally we used them for tasks with recommended disclosure: writing and editing code and figures, drafting and editing parts of this paper, and literature search and reference formatting. We have reviewed all AI-assisted work: every claim was checked against its proof and against constructed counterexamples, every reported number was regenerated from committed result files rather than transcribed, and the authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Full proofs and explicit assumptions for every theorem statement are in Appendix A, and the protocol, backbones, hyperparameters and per-dataset results in Appendix B; all eighteen datasets are publicly available for research. The anonymized supplement carries the code, the result exports and the out-of-fold predictions every table and figure is built from, so each number can be rechecked without retraining.

REFERENCES

Mohammad Ashhad, Robert Hoehndorf, and Ricardo Henao. GRAFT: Decoupling ranking and calibration for survival analysis. *arXiv preprint arXiv:2602.07884*, 2026.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning (ICML)*, pp. 1117–1174, 2022.

Spyridon Bakas, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Ujjwal Baid, et al. The university of pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. *Scientific Data*, 9(1):453, 2022.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning (ICML)*, 2020.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 89–96, 2005.

Evan Calabrese, Javier E. Villanueva-Meyer, Jeffrey D. Rudie, Andreas M. Rauschecker, Ujjwal Baid, Spyridon Bakas, Soonmee Cha, John T. Mongan, and Christopher P. Hess. The university of california san francisco preoperative diffuse glioma MRI dataset. *Radiology: Artificial Intelligence*, 4(6):e220058, 2022.

Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew F. K. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4025, 2021.

Richard J. Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, and Faisal Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8): 865–878, 2022.

Kevin Elgui, Alex Nowak, and Geneviève Robin. A statistical learning take on the concordance index for survival analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *PMLR*, pp. 4712–4731, 2023.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 939–945, 2015.

Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, and Rajesh Ranganath. X-CAL: Explicit calibration for survival analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xintian Han, Mark Goldstein, Aahlad Puli, Thomas Wies, Adler Perotte, and Rajesh Ranganath. Inverse-weighted survival games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Alan Herschtal and Bhavani Raskutti. Optimising area under the ROC curve using gradient descent. In *International Conference on Machine Learning (ICML)*, 2004.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Angel Bautista, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. Addressing the loss-metric mismatch with adaptive loss alignment. In *International Conference on Machine Learning (ICML)*, 2019.

Tao Huang, Zekang Li, Hua Lu, Nian Ye, Jiyang Han, Han Hu, Han Wang, and Xiaojun Chang. Relational surrogate loss learning. In *International Conference on Learning Representations (ICLR)*, 2022.

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

Guillaume Jaume, Anurag Vaidya, Richard J. Chen, Drew F. K. Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11579–11590, 2024.

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, 2018.

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.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, 1991.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2314–2321, 2018.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989.

Chirag Nagpal, Xinyu Rachel 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, 2021.

Tao Qin, Tie-Yan Liu, and Hang Li. A general approximation framework for direct optimization of information retrieval measures. *Information Retrieval*, 13(4):375–397, 2010.

Numan Saeed, Salma Hassan, Shahad Hardan, Ahmed Aly, Darya Taratynova, Umair Nawaz, Ufaq Khan, Muhammad Ridzuan, Vincent Andrearczyk, Adrien Depeursinge, et al. A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. *arXiv preprint arXiv:2509.00367*, 2025.

Numan Saeed, Salma Hassan, Shahad Hardan, Lishan Cai, Xinglong Liang, Moona Mazher, Abdul Qayyum, Vincent Andrearczyk, Adrien Depeursinge, Mathieu Hatt, et al. HEad and neCK tumOR (HECKTOR) 2025: Benchmark of segmentation, diagnosis, and prognosis in multimodal PET/CT. *arXiv preprint arXiv:2606.20143*, 2026.

Amber L. Simpson, Jacob Peoples, John M. Creasy, Gabor Fichtinger, Natalie Gangai, Krishna N. Keshavamurthy, Andras Lasso, Jinru Shia, Michael I. D'Angelica, and Richard K. G. Do. Preoperative CT and survival data for patients undergoing resection of colorectal liver metastases. *Scientific Data*, 11(1):172, 2024.

Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C. Raykar. On ranking in survival analysis: Bounds on the concordance index. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2007.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

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.

Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

Lian Yan, Robert Dodier, Michael C Mozer, and Richard Wolniewicz. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *International Conference on Machine Learning (ICML)*, pp. 848–855, 2003.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pp. 1845–1853, 2011.

Zhuoning Yuan, Zhishuai Guo, Nitesh Chawla, and Tianbao Yang. Compositional training for end-to-end deep AUC maximization. In *International Conference on Learning Representations (ICLR)*, 2022.

Liwen Zhang, Lianzhen Zhong, Fan Yang, Di Dong, Hui Hui, and Jie Tian. TripleSurv: Triplet time-adaptive coordinate learning approach for survival analysis. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):9464–9475, 2024.

A Theoretical results and proofs

A.1 The general form of the decoupling

Proposition 1 exhibits one order-preserving direction. The same mechanism holds for order-preserving directions in general, at the cost of a statement close to a restatement of its hypothesis, and we record it here rather than in the body.

Theorem 2 (General decoupling)**.** *Let G be the strictly increasing reparameterizations acting on scores by composition and L Hadamard-differentiable and not G-invariant (the Cox partial likelihood, the discrete-time likelihoods, and the likelihood term of DeepHit all qualify, and so does the SCL at any fixed $\tau > 0$). Then the orbit-tangent space is contained in the rank-preserving cone, $T_f(G \cdot f) \subseteq \ker dC(f)$, and at any f that is not orbit-critical the gradient of L has a nonzero component in $T_f(G \cdot f)$: a descent direction of L along which the C-index is first-order stationary.*

Observation 1 (Uninformative late in training)**.** *Once the ranking has approximately converged, the value of the* monitored *loss keeps moving while the C-index moves only through finite-sample noise, so the loss value becomes uninformative about the metric. The statement is about the held-out value, not about residual descent on the training objective: for a linear last layer the rescaling direction is orbit-critical exactly at a training stationary point. What drives the held-out value is that the training and validation losses have different orbit minimizers while the C-index is exactly constant along the orbit for both (Appendix A).*

This appendix states and proves the two claims behind the method. The first is that on a score class satisfying Assumption A1 the SCL is ε_τ-value-monotone in the concordance index with $\varepsilon_\tau \leq 2B\tau + o(\tau)$, so its value tracks the metric across training and not only at its minimizer. The second is that no finite tolerance is available for the standard survival likelihood losses, whose value moves without bound while the concordance index is exactly fixed, most severely late in training. The mechanism behind the second claim is not exclusive to likelihoods and covers the SCL at a fixed τ as well; what separates them is the first claim's bound. Both follow from a single structural fact, namely that the concordance index is invariant to strictly increasing reparameterizations of the risk score.

A.2 Setup

The data are triples (X, T, δ) with covariates $X \in \mathcal{X}$, event or censoring time $T > 0$, and event indicator $\delta \in \{0, 1\}$. A model is a risk score $f : \mathcal{X} \to \mathbb{R}$, where a higher score denotes higher risk and shorter survival, and we write $f_i = f(X_i)$. A pair (i, j) is comparable if $T_i < T_j$ and $\delta_i = 1$, so that the earlier subject is known to have failed first, and we write $\mathcal{P}$ for the distribution of comparable pairs. The population concordance index is

$$C(f) = \Pr\big(f_i > f_j \mid (i,j) \text{ comparable}\big) = \mathbb{E}_{(i,j)\sim\mathcal{P}}\big[\mathbf{1}\{f_i > f_j\}\big],$$

with ties split as one half. The Harrell and Uno versions differ only in how the inverse-probability-of-censoring weights reweight $\mathcal{P}$, which does not affect Lemma 1, Theorem 1 or Propositions 1, 1a and 2. It does affect Proposition A1: the weights $1/\hat{G}(T)^2$ are not uniformly bounded, so the kernel leaves $[0, 1/2]$ and the bounded-differences step below does not apply to the Uno form. Proposition A1 is stated for the unweighted concordance index.

Assumption A1 (non-atomic gap). There are constants $\eta > 0$ and $B < \infty$ such that for every $f \in \mathcal{F}$ the score gap $Z = f_i - f_j$ under $\mathcal{P}$ has a density bounded by B on $(-\eta, \eta)$. Both constants are common to the class, which is what the uniformity in Theorem 1 and Proposition A1 requires; the $o(\tau)$ remainder there is $O(e^{-\eta/\tau})$ and so depends on η.

This is a genuine restriction on $\mathcal{F}$ and not a regularity formality. It is not implied by X having a density and f being non-degenerate: for $X \sim \mathrm{Unif}[-1, 1]$ and $f(x) = x^3$, which is strictly increasing, $f(X)$ has density proportional to $|u|^{-2/3}$ and the gap density at zero is infinite. Numerically the estimated gap density on $(-h, h)$ grows as $1.3, 6.4, 28$ for $h = 10^{-1}, 10^{-3}, 10^{-5}$, against a constant 0.5 for the identity score. A1 therefore excludes score classes that concentrate comparable pairs near ties, which is the same condition the near-tie diagnostic of Appendix B estimates in practice.

A.3 The structural fact

Lemma 1 (rank-invariance of the concordance index). For every strictly increasing function $\varphi : \mathbb{R} \to \mathbb{R}$, $C(\varphi \circ f) = C(f)$.

Proof. Since φ is strictly increasing, $\varphi(f_i) > \varphi(f_j)$ if and only if $f_i > f_j$, so the event inside the expectation is unchanged. □

Lemma 1 says that the concordance index factors through the ordering of f alone, so every class of scores sharing an ordering lies inside a level set of C. The inclusion is strict: scores with different

orderings can share a concordance value, so a level set is in general a union of order-equivalence classes. A loss is faithful to the concordance index precisely when, informally, its value does not move within such a class. Define the rank-preserving tangent cone at a score f as the set of directions that change f without changing C to first order, $R_f = \{h : \frac{d}{dt} C(f + th)\big|_{t=0} = 0\}$.

A.4 The SCL is value-monotone

For temperature $\tau > 0$ the population SCL over comparable pairs is

$$L_\tau(f) = \mathbb{E}_{(i,j)\sim\mathcal{P}}\big[\sigma(-(f_i - f_j)/\tau)\big], \qquad \sigma(u) = \frac{1}{1 + e^{-u}}.$$

Theorem 1 (value-monotonicity). For every f and every $\tau > 0$, with no assumption on the law of the score gap,

$$\big|L_\tau(f) - (1 - C(f))\big| \;\le\; \mathbb{E}\big[\sigma(-|Z_f|/\tau)\big] \;\le\; \inf_{k>0}\Big\{\Pr(|Z_f| < k\tau) + e^{-k}\Big\},$$

where $Z_f = f_i - f_j$ on comparable pairs. Under Assumption A1 the right-hand side is at most $2B\tau + O(e^{-\eta/\tau})$ uniformly over $\mathcal{F}$, so $L_\tau(f) = (1 - C(f)) + O(\tau)$ as $\tau \to 0$. We write $\varepsilon_\tau = \sup_{f\in\mathcal{F}} |L_\tau(f) - (1 - C(f))|$ for the resulting uniform tolerance.

Three consequences of stating it this way are worth drawing. First, the middle expression is *exactly computable* from a set of predictions, so every checkpoint carries its own certified tolerance and Proposition 3 can be applied without knowing B. Second, exact ties cost nothing: $\sigma(0) = 1/2$ and the concordance index splits ties as one half, so an atom of Z_f at zero contributes zero to the deviation. This matters because A1 forbids such atoms while several of our cohorts have duplicated covariate rows and therefore exhibit them, and the conclusion holds there regardless. Third, $\Pr(|Z_f| < \tau)$ is the near-tie fraction that Appendix B measures, which is therefore not a proxy for the tolerance but, with the e^{-k} term, an upper bound on it.

The statement is an approximation, not an order isomorphism. At a fixed positive τ the value of L_τ is *not* a monotone function of C: Proposition 3 below gives the guarantee that does hold, and Remark A1 exhibits a three-point counterexample showing that the guarantee cannot be strengthened.

Proof. Let $Z = f_i - f_j$ have density p. Then $L_\tau(f) = \mathbb{E}[\sigma(-Z/\tau)]$ and $1 - C(f) = \mathbb{E}[\mathbf{1}\{Z < 0\}]$ up to the measure-zero tie set under Assumption A1. Their difference is $\mathbb{E}[g_\tau(Z)]$ with $g_\tau(z) = \sigma(-z/\tau) - \mathbf{1}\{z < 0\}$, which satisfies $|g_\tau(z)| \le e^{-|z|/\tau}$, because for $z > 0$ we have $\sigma(-z/\tau) \le e^{-z/\tau}$ and for $z < 0$ we have $1 - \sigma(-z/\tau) = \sigma(z/\tau) \le e^{z/\tau}$. Sharper and without reference to a density, $|g_\tau(z)| = \sigma(-|z|/\tau)$ exactly, which gives the first inequality; splitting the expectation at $|z| = k\tau$ and using $\sigma(-u) \le e^{-u}$ gives the second. Under A1, bounding the density near zero and the tails separately gives $\int e^{-|z|/\tau} p(z)\, dz \le 2B\tau + O(e^{-\eta/\tau})$. The constant $2B$ is not sharp: since $\int_0^\infty \sigma(-z/\tau)\, dz = \tau \log 2$, the same argument gives $2(\log 2)B\tau$. □

Proposition 3 (certified improvement). Suppose $|L_\tau(f) - (1 - C(f))| \le \varepsilon_\tau$ for every $f \in \mathcal{F}$. Then for any $f, f' \in \mathcal{F}$,

$$L_\tau(f') < L_\tau(f) - 2\varepsilon_\tau \implies C(f') > C(f).$$

Proof. $1 - C(f') \le L_\tau(f') + \varepsilon_\tau < L_\tau(f) - \varepsilon_\tau \le 1 - C(f)$. □

So a decrease in the loss larger than $2\varepsilon_\tau$ certifies an increase in the concordance index, and a decrease smaller than $2\varepsilon_\tau$ certifies nothing. This is the precise sense in which the value of the SCL can be read as the metric, and it is the sense we use throughout: the loss is a usable proxy at the resolution $2\varepsilon_\tau$, which Theorem 1 bounds by $4B\tau + o(\tau)$. The tolerance degrades when B is large, that is when the score gaps of comparable pairs concentrate near zero on the scale of τ.

Remark A1 (the tolerance is necessary). Take three uncensored subjects with $t_1 < t_2 < t_3$, so all three pairs are comparable, and $\tau = 0.1$. For $f = (0.02, 0.01, 0)$ the ordering is perfect, $C = 1$, yet every score gap is small relative to τ and $L_\tau = 0.467$. For $f' = (1, 2, 0)$ the first pair is inverted, $C = 2/3$, but the gaps are large relative to τ and $L_\tau = 0.333$. The second model has the lower loss and the worse concordance index. Consistently with Proposition 3, the smallest uniform

tolerance covering this pair is $\varepsilon_\tau = 0.467$, the loss improvement is 0.133, and $0.133 < 2\varepsilon_\tau$, so the guarantee correctly declines to certify it. The example also shows where the difficulty lies: it violates Assumption A1 in the quantitative sense that the gap density near zero is arbitrarily large at this scale. Shrinking all scores by a factor α is the general form of the construction, since $L_\tau(\alpha f) = L_{\tau/\alpha}(f)$; the SCL is invariant to positive rescaling only in the limit $\tau \to 0$, and at a fixed τ a sufficiently shrunk score drives the loss to $1/2$ regardless of its ordering. In our experiments this regime is not reached and the measured coupling is between 0.91 and 1.00 on every dataset, but nothing in the theory rules it out. The practical check is not the overall score scale, which a single outlier can inflate while leaving most pairs nearly tied, but the fraction of comparable pairs whose gap is smaller than τ, which estimates the density B that the bound is stated in; Appendix B measures it.

Corollary A1 (limiting objective). By dominated convergence $\lim_{\tau\to 0} L_\tau(f) = 1 - C(f)$ pointwise for every f, so $\arg\min_f \lim_{\tau\to 0} L_\tau(f) = \arg\max_f C(f)$. We record this only to fix the limiting objective; as stated it is the identity $\arg\min(1 - C) = \arg\max C$ and carries no content beyond it. In particular it is *not* Fisher consistency in the usual sense, which would concern $\arg\min_f L_\tau$ at a fixed τ or the limit of those minimizers, neither of which we establish; Remark A1 shows the fixed-τ minimizer can differ.

Why the sigmoid and not the squared hinge. The squared-hinge, or margin-based, surrogate $L_{\mathrm{hinge}}(f) = \mathbb{E}[(m - (f_i - f_j))_+^2]$ is consistent for the area under the curve (Gao & Zhou, 2015) and therefore shares Corollary A1, but it fails the test of Proposition 2. On the rescaling class $\mathcal{F}_A$ the contribution of a misordered pair grows like α^2 while the concordance index is exactly fixed, so $\varepsilon_\tau(\mathcal{F}_A) \geq cA^2$ and no finite tolerance survives $A \to \infty$, exactly as for the Cox likelihood. The SCL escapes this because its summand is bounded by one and, by Proposition 2(i), its tolerance on that class does not grow with A at all. We should be clear about what this does and does not predict: the property at work is boundedness together with approximating the misordering indicator, which a clipped or ramp surrogate would share, so the theory distinguishes the sigmoid from the squared hinge but does not single it out among bounded step approximations. We report the hinge comparison as the ablation it is and do not claim the sigmoid is the unique choice.

A.5 Likelihood losses provably decouple

The discrete-time and logistic-hazard negative log-likelihoods, and the likelihood term of DeepHit, are categorical log scores over the time bins and therefore strictly proper scoring rules in the sense of Gneiting & Raftery (2007), so their population minimizer is the true conditional law of the event given the covariates, a single point in distribution space rather than a class of orderings. The Cox partial likelihood is not a proper scoring rule for that law, since the baseline hazard is profiled out and the partial likelihood identifies only the regression coefficients; what the argument below needs is weaker and the Cox loss has it, namely that the loss is not invariant under strictly increasing reparameterizations of the score. Let G be the group of strictly increasing continuously differentiable reparameterizations acting on scores by $\varphi \cdot f = \varphi \circ f$. By Lemma 1 the concordance index is invariant under G, so $C(\varphi \cdot f) = C(f)$ for all $\varphi \in G$. Write $G \cdot f$ for the orbit of f, that is its class of order-equivalent scores, and $T_f(G \cdot f)$ for the tangent space of that orbit at f, which consists of the directions that reshape values while preserving order.

Theorem 2 (general decoupling). Let L be Hadamard-differentiable, which the chain rule along the curved path $\varphi_\alpha \circ f$ requires and Gateaux differentiability alone would not give, and not invariant under G, which the Cox partial likelihood, the discrete-time likelihoods and the likelihood term of DeepHit all satisfy, for example because $L(2f) \neq L(f)$ by Proposition 1 below, and which the SCL also satisfies at any fixed $\tau > 0$. Then the orbit tangent space is contained in the rank-preserving cone, $T_f(G \cdot f) \subseteq R_f = \ker dC(f)$ wherever the latter is defined, the inclusion being the exact statement that C is constant along the orbit rather than merely first-order stationary, and at any f that is not a critical point of L restricted to its orbit, the gradient $\nabla L(f)$ has a nonzero component in $T_f(G \cdot f)$. Equivalently, there is a descent direction of L along which the concordance index is first-order stationary, so part of the available descent on L is invisible to the concordance index. Gradient flow itself does not remain in the order-equivalence class, since ∇L generically has a component normal to the orbit; what the theorem isolates is the tangential part, which moves the loss value without moving the metric.

Two caveats on how much this delivers. The conclusion is close to a restatement of the hypothesis: "not a critical point of L restricted to its orbit" and "∇L has a nonzero orbit-tangential component" say the same thing, and the role of the non-invariance hypothesis is only to guarantee that non-orbit-critical points exist at all, not that a training trajectory stays among them, which Observation 1 addresses separately and negatively for the training objective. The substantive content of this section is therefore Proposition 1, which exhibits the direction explicitly and computes the sign of the derivative, rather than the geometric restatement here. Finally, G as defined is closed under composition but not under inverses, so $G \cdot f$ is a semigroup orbit; the argument uses only that the tangent set $T_f(G \cdot f) = \{\psi \circ f\}$ is a linear space, which holds regardless.

Proof. For any differentiable curve $\alpha \mapsto \varphi_\alpha$ in G with φ_0 the identity, the map $\alpha \mapsto C(\varphi_\alpha \circ f)$ is constant by invariance, so differentiating at $\alpha = 0$ gives $dC(f)[\xi] = 0$ for $\xi = \frac{d}{d\alpha}(\varphi_\alpha \circ f)\big|_0 \in T_f(G \cdot f)$, which shows $T_f(G \cdot f) \subseteq \ker dC(f)$. Since L is not invariant under G, the map $\alpha \mapsto L(\varphi_\alpha \circ f)$ is non-constant for some such curve, so $\frac{d}{d\alpha} L(\varphi_\alpha \circ f)\big|_0 = dL(f)[\xi] \neq 0$ unless f is a critical point of L restricted to its orbit. That direction, or its sign flip, is a descent direction of L lying in R_f. □

Before the rescaling proposition we record the statement that actually separates the two families, since Theorem 2 by itself does not: it applies to the SCL as well.

Proposition 2 (the separation is in the size of the tolerance). Fix a score f_0 and let $\mathcal{F}_A = \{\alpha f_0 : 1 \leq \alpha \leq A\}$ be the upward rescaling class, on which the concordance index is constant by Lemma 1. Then

(i) if the gap law of f_0 satisfies A1 with constants (B, η), then every αf_0 with $\alpha \geq 1$ satisfies it with (B, η) as well, so $\varepsilon_\tau(\mathcal{F}_A) \leq 2B\tau + O(e^{-\eta/\tau})$ ***uniformly in*** A;

(ii) if f_0 misorders at least one pair inside some risk set, then $\sup_{\alpha \leq A} L_{\mathrm{cox}}(\alpha f_0) \to \infty$ as $A \to \infty$, so no finite tolerance is available for the Cox partial likelihood on $\mathcal{F}_A$.

Proof. (i) The gap of αf_0 is αZ_0, whose density is $p_0(z/\alpha)/\alpha \leq B/\alpha \leq B$ on $(-\eta, \eta) \subseteq (-\alpha\eta, \alpha\eta)$ for $\alpha \geq 1$. Theorem 1 then applies with the same constants. Rescaling upward makes comparable pairs *less* nearly tied, which is why the tolerance does not degrade. (ii) Let $j \in R(i)$ with $f_{0j} > f_{0i}$ for some event i. The corresponding term of the negative partial log-likelihood is $\log \sum_{k \in R(i)} e^{\alpha f_{0k}} - \alpha f_{0i} \geq \alpha(f_{0j} - f_{0i}) \to \infty$. All other terms are nonnegative. □

Note that (ii) fails exactly when f_0 orders every risk set perfectly, which is the case $C(f_0) = 1$; away from it the divergence is linear in the scale. Numerically, on synthetic Cox data with $n = 200$ and an imperfect score, scaling α from 1 to 64 leaves the concordance index at 0.7425 to four decimals and keeps $|L_\tau - (1 - C)|$ below 0.0003, while the near-tie fraction falls from 0.040 to 0.0005 and the Cox value rises from 4.06 to 122.3.

This is the precise content of the informal claim that a likelihood value "can move without bound while the concordance index is fixed". It is a statement about a class closed under upward rescaling, and it is worth being explicit that on a class of *bounded* scores, $\sup_x |f(x)| \leq M$, the Cox value is bounded by $2M + \log n$ and a finite tolerance does exist for it, though a useless one: it is of the order of the whole range of the loss rather than of τ.

The canonical instance of this mechanism is rescaling, which we make fully explicit and rigorous.

Proposition 1 (rescaling-direction decoupling). Let L_{cox} be the negative Cox partial log-likelihood on a sample with event set D and risk sets $R(i)$, and let f be any risk score that is non-constant on some risk set. Define $\phi(\alpha) = L_{\mathrm{cox}}(\alpha f)$ for $\alpha > 0$. Then $C(\alpha f) = C(f)$ for all $\alpha > 0$ by Lemma 1, so the direction $h = f$ is exactly rank-preserving with $dC(f)[f] = 0$; and ϕ is smooth and strictly convex, with

$$\phi''(\alpha) = \frac{1}{|D|} \sum_{i \in D} \mathrm{Var}_{p_\alpha^{(i)}}[f] > 0, \qquad p_\alpha^{(i)} = \mathrm{softmax}(\alpha f) \text{ over } R(i),$$

so $\phi'(1) \neq 0$ for all f outside a measure-zero set, and at any such f one of the two rescaling directions strictly decreases ϕ while the concordance index is exactly constant.

Strict convexity does not imply an interior minimizer, and we do not claim one. When f perfectly orders every risk set, ϕ is strictly decreasing and the infimum is at $\alpha \to \infty$. When the ordering is

perfectly wrong the infimum is at $\alpha \to 0$: for two subjects with times $1 < 2$, the earlier one having the event, and scores $(0, 1)$, the Cox loss is $0.698, 0.744, 1.313, 10.000$ at $\alpha = 0.01, 0.1, 1, 10$, strictly increasing over all positive scales. An interior minimizer exists only when ϕ' changes sign, which holds for intermediate orderings but not in general. What the proposition needs, and all it asserts, is that $\phi'(1) \neq 0$, which is what makes rescaling a descent direction the concordance index cannot see.

Proof. The concordance identity is Lemma 1. Writing $\phi(\alpha) = -|D|^{-1} \sum_{i \in D} [\alpha f_i - \log \sum_{j \in R(i)} e^{\alpha f_j}]$ gives $\phi'(\alpha) = -|D|^{-1} \sum_i [f_i - \mathbb{E}_{p_\alpha^{(i)}} f]$ and $\phi''(\alpha) = |D|^{-1} \sum_i \mathrm{Var}_{p_\alpha^{(i)}} f \geq 0$, strict whenever f is non-constant on some risk set. Strict convexity yields at most one interior minimizer, and $\phi'(1) = 0$ only on a measure-zero set; if f perfectly orders every risk set then ϕ is strictly decreasing and $\phi'(1) < 0$. □

The discrete-time case. Rescaling a scalar score has no direct analogue when the network emits K logits, so we give the construction that replaces it. Write the predicted distribution over bins as $p(x) \in \Delta^{K-1}$. All the discrete-time methods we run derive their scalar risk as a fixed linear functional of that distribution, $R(p) = w^\top p$: MTLR integrates the cumulative hazard, DeepHit accumulates the incidence up to the horizon, and the TripleSurv risk is $1 - \sum_k p_k t_k^{\mathrm{bin}}$. The concordance index sees the model only through the ordering of $R(p(x))$.

Proposition A2 (discrete-time decoupling). Let $K \geq 3$, let w be non-constant, and let the model class be rich enough to realize any measurable $x \mapsto p(x)$ in the interior of Δ^{K-1}. For each x let $\mathcal{S}(x) = \{q \in \Delta^{K-1} : w^\top q = R(p(x))\}$ be the fibre through $p(x)$. Every p' with $p'(x) \in \mathcal{S}(x)$ for all x leaves each subject's risk, and therefore C, exactly unchanged. The discrete-time negative log-likelihood is convex in p, each term being $-\log$ of a linear function of p, and it is unbounded above on $\prod_x \mathcal{S}(x)$, since a fibre of dimension $K - 2 \geq 1$ contains points that send the observed bin's mass, or a censored subject's surviving mass, to zero. Hence no finite ε makes a discrete-time likelihood ε-value-monotone on a class closed under these moves.

Proof. The first claim is immediate: $R(p'(x)) = R(p(x))$ for every x, so the ordering of the risks, hence C, is unchanged. For the second, fix x and a bin k with $p_k(x) > 0$ whose mass enters that subject's likelihood term. Because w is non-constant there are bins $i \neq j$ with $w_i \neq w_j$, so the segment obtained by moving mass out of k and compensating within $\{i, j\}$ to hold $w^\top q$ fixed stays in $\mathcal{S}(x)$ and reaches $q_k \to 0$, along which $-\log q_k \to \infty$. The C-index is constant along that segment and the likelihood is not bounded on it. □

The mechanism is the one Proposition 1 exhibits for Cox: a direction the metric cannot see, along which the likelihood moves without bound. What differs is only the group. For a scalar risk it is rescaling; for a distribution over bins it is movement inside the level set of the derived risk.

Numerically, on the eight-hundred-sample synthetic Cox dataset of Figure 1, sweeping the scale over $\alpha \in [0.05, 6]$ leaves the concordance index constant at 0.8732 to four decimal places while the Cox negative log-likelihood ranges from 5.12 to 12.88, with its minimum at an intermediate scale, which is the phenomenon shown in the validation trajectories of the main text. Any optimizer that adjusts the effective scale of the score, through the weight norm, a temperature, or a normalization-layer scale, moves the likelihood without moving the concordance index.

Observation 1 (uninformativeness late in training). This is a mechanism consistent with what we measure, not a consequence of the results above; we label it an observation for that reason. Along a training trajectory, once the ranking has approximately converged, the value of the *monitored* likelihood keeps moving while the concordance index moves only through finite-sample noise, so the value of the loss becomes uninformative about the metric.

Two points of care. First, the statement is about the loss evaluated on the held-out split, not about descent on the training objective. For a model whose last layer is linear, scaling that layer scales f, so at any stationary point of the training loss the derivative along the rescaling direction, $\varphi'(1) = \langle \nabla_f L, f \rangle$, is exactly zero: the rescaling direction is orbit-critical precisely at convergence, and Proposition 1 provides no training-set descent there. We confirm this numerically on synthetic Cox data, where $\varphi'(1) = -2 \times 10^{-7}$ at the fitted maximum partial likelihood against 2.6 at a random coefficient vector.

Second, we offer the following as an explanatory mechanism consistent with the observations, not as a derivation: the training and validation losses have *different* orbit minimizers while the concordance index is exactly constant along the orbit for both. On the same synthetic data the orbit-optimal scale is 1.00 on the training split and 1.09 on the validation split, and the validation concordance index is 0.860470 at every scale from 0.4 to 3.0, to six decimal places. So as optimization drives the score scale toward the training-orbit optimum, the validation loss moves along its own orbit away from its own optimum, changing in value while the validation concordance index does not change at all. This accounts for a displacement; it does not by itself establish that the displacement continues to grow, which would need assumptions about the optimization dynamics that we do not make. What we claim is the existence of the rank-preserving direction, which is proved, and the observed behaviour of the monitored value, which is measured in Section 5. We state this as uninformativeness rather than as systematic anti-correlation, because a strongly negative rank correlation measured over such a converged window is a noise artefact rather than a genuine inversion. The defensible statement supported by the data is that the sigmoid surrogate stays informatively coupled, with rank correlations concentrated near one and low variance, while every likelihood loss becomes uninformative, with a low and high-variance rank correlation and nonzero selection regret.

A.6 A FINITE-SAMPLE VALUE-MONOTONICITY BOUND

Theorem 1 is a population statement. In practice we minimize, and monitor, the empirical loss over the observed comparable pairs, so what matters for the training-time argument is whether the empirical loss value tracks the concordance index on a finite sample and uniformly over the model class explored during optimization. We give the corresponding finite-sample bound. Let $\mathcal{F}$ be any class of risk scores, no boundedness being needed since the sigmoid kernel is bounded whatever the scores are, let $\sigma_\tau(u) = \sigma(-u/\tau)$, and let the empirical loss and empirical concordance over the sample of n subjects be

$$\widehat{L}_\tau(f) = \frac{1}{|\widehat{\mathcal{P}}|} \sum_{(i,j)\in\widehat{\mathcal{P}}} \sigma_\tau\big(f_i - f_j\big), \qquad \widehat{C}(f) = \frac{1}{|\widehat{\mathcal{P}}|} \sum_{(i,j)\in\widehat{\mathcal{P}}} \mathbf{1}\{f_i > f_j\},$$

where $\widehat{\mathcal{P}}$ is the set of observed comparable pairs.

Pair conventions. Two conventions have to be fixed, and they are not the same convention.

The first is what the loss and the metric each count as a comparable pair. The SCL sums over pairs with $t_i < t_j$ strictly and $\delta_i = 1$. The reported concordance index is the `lifelines` implementation, which additionally admits a pair in which subject i has an event and subject j is censored at the *same* observed time, since the event is then known to have occurred no later than the censored subject's last observation. Both exclude two events at an identical time. The difference is therefore one-sided and small: across the ten tabular cohorts the metric admits a median of 0.02 per cent more pairs than the loss, and 0.10 per cent on average, rising to 0.82 per cent on SEER, whose follow-up is the most heavily discretised. Evaluated on the same out-of-fold scores the two conventions give concordance indices differing by at most 0.00025, on LUNG, and by 0.00012 on SEER. So the limit $\lim_{\tau\to 0} L_\tau = 1 - C$ is exact for the loss's own convention and holds for the reported metric up to that difference, which is an order of magnitude below ε_τ at the temperature we use and two orders below the between-run variability of Appendix B. We state it rather than absorb it.

The second convention is how pairs are counted in the finite-sample argument, and it governs the rest of this section. For an ordered pair (i, j) with $i \neq j$ write $c_{ij} = \mathbf{1}\{T_i < T_j,\ \delta_i = 1\}$, so that i is the subject with the earlier observed event. At most one of c_{ij} and c_{ji} is one. Let

$$p = \Pr\big[T_1 < T_2,\ \delta_1 = 1\big]$$

for two independent subjects, which is the *ordered*-pair comparability probability. The two disjoint events $\{T_1 < T_2, \delta_1 = 1\}$ and $\{T_2 < T_1, \delta_2 = 1\}$ are exchangeable, so $p \leq 1/2$ always, with equality only in the absence of censoring and ties. On the ten tabular cohorts of this study p ranges from 0.12 on SEER, the most heavily censored, to 0.43 on SUPPORT. The fraction of *unordered* pairs that are comparable is $2p$ and is the quantity one obtains by counting directed comparable pairs against $\binom{n}{2}$; the two differ by a factor of two and must not be mixed.

Proposition A1 (finite-sample value-monotonicity). Let Assumption A1 hold uniformly over $\mathcal{F}$ with gap-density bound B, and let $\delta \in (0, 1)$. There is a universal constant C_0, independent of p, τ,

n and $\mathcal{F}$, such that if $\lfloor n/2 \rfloor \geq 2p^{-2}\log(4/\delta)$ then with probability at least $1-\delta$ over the sample, uniformly over $f \in \mathcal{F}$,

$$\left|\widehat{L}_\tau(f) - (1 - C(f))\right| \;\leq\; \underbrace{2B\tau + o(\tau)}_{\text{approximation}} \;+\; \frac{2}{p}\underbrace{\left[\frac{C_0}{\tau}\,\mathcal{R}_n(\mathcal{F}) + 2\sqrt{\frac{\log(4/\delta)}{2\lfloor n/2 \rfloor}}\right]}_{\text{estimation}},$$

where $\mathcal{R}_n(\mathcal{F})$ is the Rademacher complexity (Bartlett & Mendelson, 2002) of $\mathcal{F}$. The empirical loss therefore estimates one minus the concordance index uniformly over the class, with the error controlled by the temperature through the approximation term, by the capacity of $\mathcal{F}$ and the sample size through the estimation term, and by the censoring through $1/p$. The bound degrades as $p \to 0$ and is vacuous in the limit: the effective sample size for this estimate is the number of comparable pairs, not the number of patients.

Proof. By the triangle inequality the error splits into an approximation term $|L_\tau(f) - (1 - C(f))|$, which Theorem 1 bounds by $2B\tau + o(\tau)$ uniformly over $\mathcal{F}$, and an estimation term $|\widehat{L}_\tau(f) - L_\tau(f)|$, which is what remains.

The empirical loss is a *ratio* of U-statistics, not a U-statistic: both the comparable set and its cardinality are random, and conditioning on the comparable set does not make the summands independent. We therefore treat the ratio directly. With $m = n(n-1)$ ordered pairs put

$$\widehat{N}(f) = \tfrac{1}{m}\sum_{i \neq j} c_{ij}\,\sigma\big((f_j - f_i)/\tau\big), \qquad \widehat{D} = \tfrac{1}{m}\sum_{i \neq j} c_{ij},$$

so $\widehat{L}_\tau(f) = \widehat{N}(f)/\widehat{D}$, $\mathbb{E}\widehat{D} = p$ and $L_\tau(f) = \mathbb{E}\widehat{N}(f)/p$. Both are U-statistics of order two once written with symmetric kernels,

$$h_f(z_i, z_j) = \tfrac{1}{2}\big[c_{ij}\sigma((f_j - f_i)/\tau) + c_{ji}\sigma((f_i - f_j)/\tau)\big] \in [0, \tfrac{1}{2}], \qquad g(z_i, z_j) = \tfrac{1}{2}(c_{ij} + c_{ji}) \in [0, \tfrac{1}{2}],$$

since $\widehat{N}(f) = \binom{n}{2}^{-1}\sum_{i<j} h_f(z_i, z_j)$ and $\widehat{D} = \binom{n}{2}^{-1}\sum_{i<j} g(z_i, z_j)$.

On the event $\widehat{D} > 0$, adding and subtracting $p\,\mathbb{E}\widehat{N}$,

$$\left|\widehat{L}_\tau(f) - L_\tau(f)\right| = \frac{\left|\widehat{N}p - \mathbb{E}\widehat{N}\,\widehat{D}\right|}{\widehat{D}\,p} \leq \frac{1}{\widehat{D}}\Big(\left|\widehat{N} - \mathbb{E}\widehat{N}\right| + L_\tau(f)\left|\widehat{D} - p\right|\Big) \leq \frac{1}{\widehat{D}}\Big(\left|\widehat{N} - \mathbb{E}\widehat{N}\right| + \left|\widehat{D} - p\right|\Big),$$

using $L_\tau(f) \in [0, 1]$. Note that the bound is now on $\widehat{N}$ and $\widehat{D}$ and we do not return to $\widehat{L}_\tau$ at any later step.

For a U-statistic of order two with a symmetric kernel bounded in $[0, \frac{1}{2}]$, Hoeffding's reduction (Hoeffding, 1963) writes it as an average over permutations of the data of averages of $\lfloor n/2 \rfloor$ independent and identically distributed terms. The deviation bound then follows from Hoeffding's convexity step $\exp(\lambda U) \leq \operatorname{avg}_\pi \exp(\lambda V_\pi)$ applied to those averages, together with the bounded-differences inequality (McDiarmid, 1989) for each of them, rather than from a union bound over the permutations, which would be vacuous. This gives, for the f-free statistic $\widehat{D}$,

$$\left|\widehat{D} - p\right| \leq \sqrt{\frac{\log(4/\delta)}{2\lfloor n/2 \rfloor}} \qquad \text{with probability } 1 - \delta/2.$$

On that event $\widehat{D} \geq p - \sqrt{\log(4/\delta)/(2\lfloor n/2 \rfloor)} \geq p/2$ whenever $\lfloor n/2 \rfloor \geq 2p^{-2}\log(4/\delta)$, which is the stated sample-size condition and supplies the factor $1/\widehat{D} \leq 2/p$. Squaring the requirement $\sqrt{\log(4/\delta)/(2\lfloor n/2 \rfloor)} \leq p/2$ gives exactly this, so the condition is on $\lfloor n/2 \rfloor$ and not on n.

For $\widehat{N}$ the same reduction, followed by symmetrization over the supremum within $\mathcal{F}$, gives

$$\sup_{f \in \mathcal{F}}\left|\widehat{N}(f) - \mathbb{E}\widehat{N}(f)\right| \leq 2\,\mathcal{R}_n\big(\sigma_\tau \circ \Delta\mathcal{F}\big) + \sqrt{\frac{\log(4/\delta)}{2\lfloor n/2 \rfloor}} \qquad \text{with probability } 1 - \delta/2,$$

where $\Delta\mathcal{F} = \{(x, x') \mapsto f(x) - f(x') : f \in \mathcal{F}\}$ and the comparability indicator, being bounded by one and independent of f, only rescales the kernel. Since $u \mapsto \sigma(u/\tau)$ is $\frac{1}{4\tau}$-Lipschitz, Talagrand's contraction inequality (Ledoux & Talagrand, 1991) gives $\mathcal{R}_n(\sigma_\tau \circ \Delta\mathcal{F}) \leq \frac{1}{4\tau}\mathcal{R}_n(\Delta\mathcal{F})$,

and $\mathcal{R}_n(\Delta\mathcal{F}) \le 2\mathcal{R}_n(\mathcal{F})$ by subadditivity. Three bookkeeping points. The complexity the reduction delivers is that of the pair class at sample size $\lfloor n/2 \rfloor$, which is at most $8\,\mathcal{R}_n(\mathcal{F})$ since $n\mathcal{R}_n$ is nondecreasing, so writing $\mathcal{R}_n(\mathcal{F})$ in the statement costs only a numerical factor. The contraction inequality in its absolute-value form wants $\psi(0) = 0$, whereas our kernel has $\psi(0) = c/4$; recentering costs a further $1/(4\sqrt{\lfloor n/2 \rfloor})$, which is of the same order as the concentration term and is absorbed into it. Collecting these numerical factors into C_0, which involves no p, τ, n or $\mathcal{F}$, and combining the two events by a union bound, yields the estimation term of the statement multiplied by $2/p$. Adding the approximation term completes the proof. □

The concentration half of the same argument applied to the indicator kernel shows that $\widehat{C}(f)$, which is the same ratio of U-statistics and so carries the same $2/p$ factor, concentrates around $C(f)$ at the rate $p^{-1}\sqrt{\log(1/\delta)/\lfloor n/2 \rfloor}$ for each fixed f. Uniformity over $\mathcal{F}$ does *not* transfer: the indicator is not Lipschitz, so the contraction step is unavailable and a uniform statement would need the Vapnik-Chervonenkis dimension of $\{\mathbf{1}\{f(x) > f(x')\} : f \in \mathcal{F}\}$. For the use we make of it, monitoring selects among finitely many checkpoints on a held-out split, so a union bound over checkpoints suffices. With that caveat the empirical concordance and the empirical loss track the same population quantity, and their trajectories move together up to the temperature and complexity terms above. The bound is stated with a universal constant rather than optimized constants; sharper pairwise concentration is available through the moment method for U-processes but is not needed for the qualitative training-time conclusion.

The two bounds pull against each other in τ. Theorem 1 improves as $\tau \to 0$, its tolerance being of order $B\tau$, while the estimation term of Proposition A1 carries a $1/\tau$ from the Lipschitz constant of σ_τ and so degrades in the same limit. Balancing them gives an optimal temperature of order $\sqrt{\mathfrak{R}_n(\mathcal{F})/(pB)}$ rather than the smallest τ available. We fix $\tau = 0.1$ throughout and do not tune it, so this is a statement about what the theory recommends and not about how our experiments were run; the practical consequence is that driving τ toward zero is not free, which the near-tie diagnostic does not by itself reveal.

A.7 Remarks on scope

Proposition 1 makes the rescaling case rigorous, which already shows that decoupling is available at every score that is not critical along the rescaling direction. The general Theorem 2 extends this to arbitrary rank-preserving directions through the orbit-tangent argument. A fully formal function-space statement that a nonzero rank-preserving gradient component persists for arbitrary directions, rather than only for rescaling, is a strengthening in the same spirit, which we do not pursue here because the rescaling case already establishes the core claim.

B Experimental details and additional results

B.1 Protocol and implementation

Algorithm 1 Sigmoid concordance loss, one optimization step

1. **input** minibatch scores $f \in \mathbb{R}^B$ from the model, times $t \in \mathbb{R}^B$, events $\delta \in \{0,1\}^B$, temperature τ
2. $M_{ij} \leftarrow \mathbb{1}[t_i < t_j] \cdot \mathbb{1}[\delta_i = 1]$ *// comparable-pair mask, $O(B^2)$*
3. $D_{ij} \leftarrow f_i - f_j$ *// all pairwise score gaps*
4. $L \leftarrow \big(\sum_{ij} M_{ij}\,\sigma(-D_{ij}/\tau)\big)\big/\big(\sum_{ij} M_{ij}\big)$ *// mean smoothed misordering rate $= \widehat{1 - C}$*
5. update model parameters by a gradient step on L

All methods share the training protocol below and differ only in the loss. The study evaluates a loss, so each modality uses a standard and deliberately unremarkable backbone.

Cross-validation and the validation split. Folds are formed by shuffling the event and censored indices separately and splitting each into five parts, so every fold carries the same event rate. For

outer fold k we test on fold k, validate on fold $k + 1 \bmod 5$, and train on the remaining three folds. The validation set is therefore a held-out fold rather than a split of the training fold, no patient is ever in both the validation and the test set of the same run, and each fold serves as validation exactly once. The five test-fold risk vectors are concatenated into one out-of-fold vector and each metric is computed once on it; Appendix B.15 measures how much this pooling differs from averaging the C-index within folds. The whole procedure is repeated over ten fold partitions, which reshuffle the folds, and reported as mean $\pm$ standard deviation across them.

Hyperparameters and tuning. No hyperparameter search is performed, for any method. Within a modality every neural method shares the backbone, optimizer, learning rate, weight decay, batch size, and epoch budget given below, and differs only in the loss; each loss uses the setting its own authors report, listed in the baseline configurations below. The comparison is therefore a controlled one in which the loss is the only free variable, and it is not a claim that each baseline is at its own optimum. A search favouring any one loss would confound the comparison, and we did not have the budget to tune eleven methods fairly on eighteen datasets.

Optimization. Adam, learning rate 10^{-3}, weight decay 10^{-4}, one hundred epochs. The validation set is evaluated every two epochs and the checkpoint with the best validation Harrell C-index is used for testing, except in Table A7 where the selection rule is the variable of interest. The trajectory figures are drawn from separate 150-epoch runs on the tabular and pathology cohorts, so that the late-training behaviour the theory concerns is visible well past the point where the C-index has plateaued; all reported metrics come from the 100-epoch runs.

Which run produces which number. The canonical run set is one execution of the protocol over ten fold partitions on every one of the eighteen datasets, fixed before the numbers were read and not selected among repetitions. Two scripts train that configuration, and they answer different questions, so we name which one backs each table. The per-cohort benchmark runs (the `*_cv` scripts) produce every discrimination, calibration and coupling number, including the integrated Brier scores, which come from the same training run and the same checkpoint as the C-index beside them. The checkpoint-selection runs (the `selection_analysis_*` scripts) produce Table A7 and Figure 3, because the quantity there is a paired within-run comparison of two selection rules applied to the same trained models, which requires both checkpoints of a single run. The two sets train the same configuration independently and therefore differ by the retraining spread quantified in Appendix B.16. A separate four-repetition experiment, used only in that appendix, measures how much of the imaging spread is retraining nondeterminism rather than the fold partition; it backs no number in the main text. We do not set deterministic algorithm flags, so runs are not bitwise reproducible.

Loss. The SCL uses temperature $\tau = 0.1$, fixed and not tuned. The discrete-time methods (MTLR and DeepHit) bin the time axis at the quantiles of the event times, giving roughly equal numbers of events per bin.

Backbones and batch sizes. Tabular: a multilayer perceptron with two hidden layers of 64 units, ReLU, batch normalization, dropout 0.3; batch size $\min(256, \max(32, n/4))$ for training-fold size n. CT, PET-CT, and MRI: a compact 3D convolutional encoder producing a 64-dimensional feature, concatenated with the clinical covariates, followed by a small multilayer perceptron head with dropout 0.2; batch size 16, limited by 3D volume memory. Pathology: a multilayer perceptron on the 768-dimensional TITAN slide feature, layer normalization, hidden widths 256 and 128, GELU, dropout 0.3; batch size 64. The pairwise loss forms comparable pairs within a batch, so larger batches are preferable where memory allows.

Imaging preprocessing. Volumes are cropped to a tumor-centered cube of 128 mm guided by the segmentation and resampled to $64 \times 64 \times 64$ voxels at 2 mm isotropic spacing, which preserves physical scale across patients. MRI sequences are z-scored within the brain region per sequence, CT intensities are windowed to a fixed soft-tissue Hounsfield range, and the HECKTOR PET and CT inputs use the challenge-provided normalization.

Software. PyTorch, with lifelines and scikit-survival for the survival metrics and torchmtlr for multi-task logistic regression.

B.2 Evaluation metrics

Harrell C-index. The fraction of comparable pairs whose risks are correctly ordered, ties in the risk contributing one half.

Truncated Uno C-index. The inverse-probability-of-censoring-weighted C-index, restricted to comparable pairs whose case event time is at most the eightieth percentile of the event times. The truncation is used because the weights become unstable in the sparse tail of the follow-up, which otherwise inflates the variance of the estimate.

Time-dependent AUC. The cumulative/dynamic area under the curve averaged over twenty evaluation times placed at the tenth to ninetieth percentiles of the test event times.

Integrated Brier score. For the scalar-risk methods (SCL, Cox, TripleSurv, hybrid) a Breslow baseline estimated on the training fold turns the risk into a survival function $S(t \mid x) = \exp(-H_0(t)e^{f(x)})$; the discrete-time methods use their native predicted survival curves. The score is integrated over a common grid of twenty time points spanning the observed event times. Lower is better.

B.3 Baseline configurations

All neural baselines share the backbone and optimization above; the loss and the width of the output head are the only differences. **Cox**: negative Cox partial log-likelihood with Breslow handling of ties. **MTLR**: discrete-time negative log-likelihood (torchmtlr) with weight regularization strength one; for every coupling result the monitored validation loss is the pure negative log-likelihood with the weight penalty removed, so that the quantity compared against the C-index reflects data fit rather than the weight norm, consistently with the other losses. The training objective is unchanged. **DeepHit**: a discrete-time likelihood term plus an exponential ranking term, ranking weight 0.2, smoothing 0.1. **TripleSurv**: reimplemented from the paper, since no public code is available, as a weighted sum of a likelihood term, an exponential time-adaptive pairwise ranking term, and a calibration term, with weights 1.0, 1.0, 0.5, ranking temperature 1.0, time scale 1.0; the ranking term enters with the sign its Equation 6 gives, which maximizes it and is the concordant direction, as Appendix B.7 sets out; the survival term inside the likelihood is clamped to a small positive floor before the logarithm, because one minus the cumulative softmax can become a small negative number at the last time bin and otherwise produces non-finite values for censored subjects. **Random survival forest**: scikit-survival, one hundred trees, minimum fifteen samples per leaf. **Hybrid**: the SCL plus one half of a logistic-hazard negative log-likelihood anchor, reported to test the theory and not as a competing method.

B.4 Per-dataset results

Tables A3, A4 and A5 give the per-dataset numbers behind the pooled tabular and pathology rows of Table 1. Table A6 compares the convex linear ranking SVM of Section 4 against the deep SCL. Table A7 reports, for every cohort, the pooled test C-index under each of the two checkpoint-selection rules, which is the quantity summarized by the selection regret of Figure 3.

B.5 Study design and the anchor-weight sweep

Figure A2 shows the shared design: a modality-specific backbone feeds an output head, and the loss is the only component that varies apart from the width of that head. Figure A3 sweeps the likelihood-anchor weight of the hybrid variant, which is the evidence behind the claim in Section 5 that every anchor weight we tested degrades coupling without buying discrimination.

Table A1: Mean within-dataset rank over the eighteen datasets under both checkpoint-selection rules, for the six losses run on every dataset. Ranks are formed per dataset on the unrounded pooled test C-index with ties averaged, then averaged over datasets; intervals are percentile intervals from 5000 bootstrap resamples of the eighteen datasets, and P(best) is the fraction of resamples in which a loss attains the lowest mean rank. The Nemenyi critical difference at $k = 6$, $N = 18$ and $\alpha = 0.05$ is 1.78, against an observed gap from first to second of 1.06 under the first rule and 0.06 under the second, so neither ranking separates the leading losses. The last two columns do: they give the worst absolute difference in test C-index between the two rules and the number of datasets on which that difference exceeds 0.02.

Loss	By validation C-index			By validation loss			Rule disagreement	
	rank	95% CI	P(best)	rank	95% CI	P(best)	worst	> 0.02
SCL	2.00	[1.44, 2.61]	0.998	2.83	[2.11, 3.61]	0.364	0.006	0/18
Squared-hinge	3.06	[2.61, 3.56]	0.002	2.89	[2.33, 3.44]	0.212	0.016	0/18
MTLR	3.50	[2.78, 4.28]	0.000	3.00	[2.17, 3.83]	0.259	0.033	4/18
Cox	3.56	[2.89, 4.22]	0.000	3.06	[2.39, 3.72]	0.165	0.024	1/18
DeepHit	4.39	[3.50, 5.17]	0.000	4.67	[3.94, 5.33]	0.000	0.064	6/18
TripleSurv	4.50	[3.67, 5.22]	0.000	4.56	[3.89, 5.17]	0.000	0.032	6/18

Table A2: Coupling and selection-rule discrepancy, per dataset over the eighteen datasets. *Coupling* is the mean absolute Spearman correlation between the negative validation loss and the validation C-index over the second half of training, averaged over that dataset's fifty folds and then summarized across datasets; Appendix B.11 shows the conclusion does not depend on the window. We take magnitudes before averaging, since a signed average sends a loss whose direction reverses between folds toward zero and so reads as no relationship where the truth is a relationship of unstable sign. *Sign flips* is the fraction of folds whose correlation opposes their own dataset's mean, which is what a signed average hides. *Discrepancy* is the signed difference in test C-index between the two selection rules (Figure 3). Each loss contributes 900 fold correlations, except DeepHit, seven of whose folds give an undefined correlation because its validation C-index is constant over the window and are excluded. The separation is not in the averages alone: the SCL's weakest dataset, 0.914 on LUNG, sits above the strongest single cell of any baseline, the squared hinge's 0.874 on FLCHAIN.

Loss	Coupling: mean $\lvert r\rvert$ [range]	Below 0.5	Sign flips	Discrepancy: mean / worst $\lvert\cdot\rvert$
SCL	**0.97** [0.91, 1.00]	**0/18**	**0.00**	-0.001 **/ 0.006**
Squared-hinge	0.53 [0.23, 0.87]	10/18	0.10	-0.004 / 0.016
Cox	0.41 [0.17, 0.77]	11/18	0.22	-0.006 / 0.024
MTLR	0.46 [0.19, 0.72]	11/18	0.18	-0.008 / 0.033
DeepHit	0.37 [0.25, 0.70]	16/18	0.31	-0.014 / 0.064
TripleSurv	0.39 [0.27, 0.70]	14/18	0.25	-0.006 / 0.032

B.6 Related loss families

Table A8 groups the alternatives by what they optimize. The linear survival SVM of scikit-survival reaches 0.725 against the SCL's 0.735 but is convex-solved with no trajectory to monitor, and it is a different estimator from the squared-margin relaxation of our own loss, which reaches 0.734 on the same pool (Table A6). A faithful adaptation of relational surrogate loss learning (Huang et al., 2022) collapses when its learned surrogate is used alone (0.618) and, combined with the Cox likelihood as its authors recommend, recovers discrimination (0.729) but inherits the likelihood's decoupling, coupling only 0.50 and ranging from 0.16 to 0.93 across datasets. The calibration-oriented X-CAL (Goldstein et al., 2020) and inverse-weighted survival games (Han et al., 2021) target proper scoring rather than concordance and are likewise unstably coupled (0.27, 0.19). Learning or reweighting a loss does not deliver value alignment: the learned or calibration term cannot stand alone, and the likelihood it leans on is what decouples.

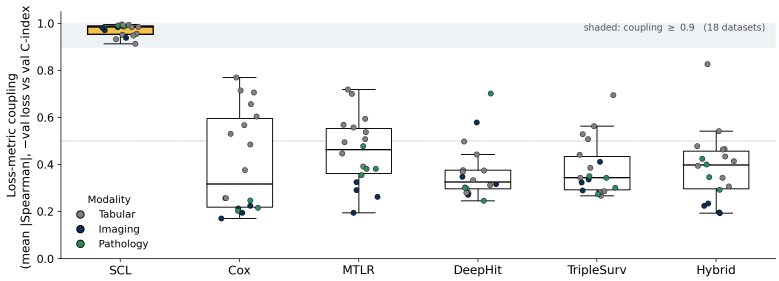


Figure A1: Loss-metric coupling per dataset, one point per (dataset, method), coloured by modality. The SCL occupies a narrow band on all eighteen datasets. The baselines have medians close to one another and each scatters across most of the range; the value plotted is the mean absolute per-fold correlation, the same statistic as Table A2.

Table A3: Per-dataset results on the first five tabular cohorts, one of ten. Pooled five-fold cross-validation, mean ± standard deviation over ten fold partitions. Coupling is the mean absolute loss-metric rank correlation over that cohort's fifty folds; the random survival forest has no training trajectory and no integrated Brier score under this protocol.

Dataset	Loss	Harrell C-index	Trunc. Uno	Time-dep. AUC	Coupling	Int. Brier
METABRIC	**SCL**	0.689 ± 0.007	0.679 ± 0.006	0.735 ± 0.007	0.987	0.164
	Squared-hinge	0.685 ± 0.007	0.675 ± 0.006	0.729 ± 0.007	0.328	0.169
	Cox	0.683 ± 0.009	0.674 ± 0.009	0.726 ± 0.010	0.258	0.165
	MTLR	0.671 ± 0.011	0.664 ± 0.009	0.712 ± 0.013	0.595	0.174
	DeepHit	0.672 ± 0.013	0.661 ± 0.013	0.709 ± 0.018	0.376	0.172
	TripleSurv	0.671 ± 0.013	0.661 ± 0.012	0.711 ± 0.014	0.386	0.167
	RSF	0.706 ± 0.004	0.696 ± 0.004	0.753 ± 0.005	n/a	n/a
SUPPORT	**SCL**	0.878 ± 0.001	0.899 ± 0.001	0.963 ± 0.000	0.994	0.132
	Squared-hinge	0.880 ± 0.001	0.900 ± 0.001	0.963 ± 0.001	0.818	0.125
	Cox	0.877 ± 0.002	0.898 ± 0.002	0.962 ± 0.001	0.531	0.122
	MTLR	0.877 ± 0.001	0.897 ± 0.001	0.961 ± 0.001	0.701	0.094
	DeepHit	0.877 ± 0.001	0.898 ± 0.001	0.962 ± 0.001	0.313	0.106
	TripleSurv	0.863 ± 0.007	0.884 ± 0.007	0.950 ± 0.005	0.281	0.119
	RSF	0.826 ± 0.001	0.843 ± 0.001	0.916 ± 0.001	n/a	n/a
GBSG	**SCL**	0.670 ± 0.011	0.675 ± 0.012	0.725 ± 0.013	0.952	0.178
	Squared-hinge	0.660 ± 0.010	0.664 ± 0.011	0.711 ± 0.011	0.688	0.175
	Cox	0.663 ± 0.011	0.667 ± 0.010	0.713 ± 0.011	0.658	0.174
	MTLR	0.666 ± 0.023	0.670 ± 0.025	0.718 ± 0.031	0.538	0.180
	DeepHit	0.625 ± 0.015	0.628 ± 0.015	0.666 ± 0.017	0.281	0.249
	TripleSurv	0.647 ± 0.021	0.649 ± 0.023	0.697 ± 0.028	0.442	0.173
	RSF	0.683 ± 0.008	0.685 ± 0.008	0.744 ± 0.010	n/a	n/a
ROTTERDAM	**SCL**	0.707 ± 0.002	0.710 ± 0.003	0.751 ± 0.004	0.985	0.161
	Squared-hinge	0.707 ± 0.003	0.711 ± 0.003	0.751 ± 0.003	0.844	0.156
	Cox	0.705 ± 0.003	0.708 ± 0.003	0.746 ± 0.005	0.716	0.156
	MTLR	0.708 ± 0.002	0.711 ± 0.002	0.755 ± 0.002	0.719	0.153
	DeepHit	0.655 ± 0.024	0.658 ± 0.024	0.694 ± 0.026	0.280	0.224
	TripleSurv	0.695 ± 0.003	0.698 ± 0.003	0.738 ± 0.002	0.344	0.158
	RSF	0.713 ± 0.002	0.716 ± 0.002	0.759 ± 0.002	n/a	n/a
FLCHAIN	**SCL**	0.796 ± 0.001	0.796 ± 0.001	0.819 ± 0.001	0.995	0.114
	Squared-hinge	0.795 ± 0.001	0.795 ± 0.001	0.818 ± 0.001	0.874	0.089
	Cox	0.794 ± 0.004	0.794 ± 0.004	0.817 ± 0.004	0.706	0.088
	MTLR	0.792 ± 0.001	0.792 ± 0.001	0.818 ± 0.001	0.392	0.088
	DeepHit	0.702 ± 0.021	0.706 ± 0.022	0.732 ± 0.022	0.303	0.178
	TripleSurv	0.745 ± 0.017	0.749 ± 0.016	0.773 ± 0.016	0.267	0.100
	RSF	0.788 ± 0.001	0.790 ± 0.002	0.816 ± 0.002	n/a	n/a

B.7 The TripleSurv ranking term, and an error of ours

TripleSurv has no public implementation, so the baseline is reimplemented from the paper, and one detail of that reimplementation was wrong for most of this work. We record it because it changed a baseline's numbers and because the shape of the mistake is easy to repeat. Equation 5 of that paper

Table A4: Per-dataset results on the last five tabular cohorts, two of ten. Pooled five-fold cross-validation, mean ± standard deviation over ten fold partitions. Coupling is the mean absolute loss-metric rank correlation over that cohort's fifty folds; the random survival forest has no training trajectory and no integrated Brier score under this protocol.

Dataset	Loss	Harrell C-index	Trunc. Uno	Time-dep. AUC	Coupling	Int. Brier
WHAS500	**SCL**	0.750 ± 0.010	0.753 ± 0.010	0.775 ± 0.010	0.950	0.160
	Squared-hinge	0.741 ± 0.009	0.744 ± 0.010	0.767 ± 0.010	0.645	0.165
	Cox	0.744 ± 0.013	0.749 ± 0.012	0.768 ± 0.014	0.568	0.162
	MTLR	0.742 ± 0.013	0.746 ± 0.014	0.769 ± 0.017	0.570	0.176
	DeepHit	0.727 ± 0.020	0.730 ± 0.020	0.748 ± 0.025	0.444	0.247
	TripleSurv	0.726 ± 0.013	0.731 ± 0.013	0.751 ± 0.016	0.695	0.161
	RSF	0.765 ± 0.004	0.769 ± 0.004	0.789 ± 0.004	n/a	n/a
SEER	**SCL**	0.720 ± 0.009	0.721 ± 0.009	0.739 ± 0.011	0.996	0.072
	Squared-hinge	0.716 ± 0.010	0.717 ± 0.010	0.735 ± 0.011	0.867	0.071
	Cox	0.720 ± 0.007	0.722 ± 0.007	0.740 ± 0.007	0.770	0.071
	MTLR	0.710 ± 0.013	0.709 ± 0.013	0.734 ± 0.011	0.495	0.073
	DeepHit	0.628 ± 0.025	0.627 ± 0.025	0.643 ± 0.026	0.334	0.168
	TripleSurv	0.692 ± 0.018	0.692 ± 0.019	0.717 ± 0.017	0.287	0.074
	RSF	0.725 ± 0.003	0.725 ± 0.003	0.745 ± 0.003	n/a	n/a
PBC	**SCL**	0.799 ± 0.010	0.790 ± 0.008	0.838 ± 0.009	0.934	0.126
	Squared-hinge	0.792 ± 0.010	0.781 ± 0.008	0.831 ± 0.012	0.657	0.129
	Cox	0.795 ± 0.012	0.787 ± 0.008	0.833 ± 0.013	0.605	0.126
	MTLR	0.793 ± 0.008	0.782 ± 0.010	0.832 ± 0.010	0.508	0.132
	DeepHit	0.783 ± 0.019	0.770 ± 0.019	0.820 ± 0.020	0.498	0.212
	TripleSurv	0.787 ± 0.016	0.776 ± 0.019	0.823 ± 0.019	0.563	0.129
	RSF	0.817 ± 0.003	0.804 ± 0.004	0.853 ± 0.004	n/a	n/a
LUNG	**SCL**	0.611 ± 0.017	0.606 ± 0.015	0.626 ± 0.023	0.914	0.220
	Squared-hinge	0.598 ± 0.010	0.594 ± 0.010	0.609 ± 0.017	0.420	0.226
	Cox	0.603 ± 0.015	0.600 ± 0.013	0.619 ± 0.023	0.376	0.224
	MTLR	0.605 ± 0.013	0.601 ± 0.013	0.621 ± 0.019	0.448	0.210
	DeepHit	0.597 ± 0.012	0.590 ± 0.014	0.610 ± 0.016	0.377	0.211
	TripleSurv	0.593 ± 0.018	0.588 ± 0.019	0.606 ± 0.025	0.530	0.219
	RSF	0.624 ± 0.009	0.618 ± 0.009	0.646 ± 0.013	n/a	n/a
TCGA-GBM	**SCL**	0.733 ± 0.007	0.729 ± 0.007	0.814 ± 0.008	0.956	0.137
	Squared-hinge	0.728 ± 0.006	0.725 ± 0.006	0.808 ± 0.008	0.607	0.139
	Cox	0.716 ± 0.009	0.714 ± 0.009	0.791 ± 0.010	0.486	0.141
	MTLR	0.715 ± 0.009	0.714 ± 0.009	0.787 ± 0.012	0.559	0.137
	DeepHit	0.718 ± 0.006	0.715 ± 0.006	0.792 ± 0.008	0.369	0.151
	TripleSurv	0.715 ± 0.013	0.713 ± 0.013	0.788 ± 0.015	0.508	0.142
	RSF	0.707 ± 0.005	0.707 ± 0.005	0.780 ± 0.006	n/a	n/a

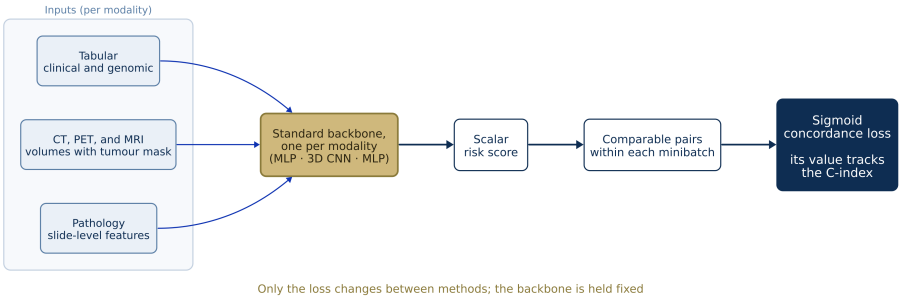


Figure A2: Overview of the study design. For each modality a standard backbone maps the input to a scalar risk score, and the loss is the only free variable apart from the output head, which emits K logits for the discrete-time methods and one scalar for the rest. The SCL is computed over comparable patient pairs within each minibatch.

defines the time-adaptive pairwise ranking term as

$$l_{\mathrm{TAPR}} = \frac{1}{|\mathcal{A}^1|} \sum_{i \in \mathcal{A}^1} \sum_{j \in R(i)} \exp\big(\sigma\big[(\mathrm{risk}_i - \mathrm{risk}_j) - \rho\,(t_j - t_i)\big]\big),$$

Table A5: Per-cohort results on the four TCGA pathology cohorts, under the protocol of Table A3.

Dataset	Loss	Harrell C-index	Trunc. Uno	Time-dep. AUC	Coupling	Int. Brier
TCGA-BRCA	**SCL**	0.646 ± 0.030	0.618 ± 0.025	0.633 ± 0.026	0.989	0.143
	Squared-hinge	0.616 ± 0.020	0.600 ± 0.016	0.607 ± 0.018	0.401	0.143
	Cox	0.628 ± 0.033	0.611 ± 0.030	0.613 ± 0.036	0.247	0.140
	MTLR	0.580 ± 0.019	0.572 ± 0.026	0.573 ± 0.029	0.478	0.149
	DeepHit	0.606 ± 0.029	0.592 ± 0.033	0.586 ± 0.033	0.297	0.144
	TripleSurv	0.621 ± 0.037	0.614 ± 0.033	0.603 ± 0.032	0.302	0.149
TCGA-GBMLGG	**SCL**	0.804 ± 0.014	0.795 ± 0.018	0.830 ± 0.016	0.988	0.136
	Squared-hinge	0.808 ± 0.014	0.798 ± 0.015	0.835 ± 0.014	0.324	0.135
	Cox	0.812 ± 0.010	0.803 ± 0.012	0.838 ± 0.011	0.217	0.133
	MTLR	0.809 ± 0.013	0.800 ± 0.016	0.834 ± 0.015	0.382	0.137
	DeepHit	0.813 ± 0.007	0.803 ± 0.007	0.836 ± 0.010	0.245	0.153
	TripleSurv	0.800 ± 0.011	0.792 ± 0.014	0.824 ± 0.013	0.343	0.133
TCGA-LUAD	**SCL**	0.563 ± 0.017	0.566 ± 0.017	0.568 ± 0.020	0.991	0.197
	Squared-hinge	0.557 ± 0.020	0.567 ± 0.021	0.562 ± 0.024	0.375	0.195
	Cox	0.557 ± 0.016	0.565 ± 0.016	0.565 ± 0.022	0.257	0.194
	MTLR	0.557 ± 0.030	0.560 ± 0.023	0.562 ± 0.035	0.383	0.197
	DeepHit	0.567 ± 0.017	0.565 ± 0.015	0.567 ± 0.022	0.301	0.199
	TripleSurv	0.577 ± 0.015	0.586 ± 0.015	0.585 ± 0.015	0.352	0.191
TCGA-UCEC	**SCL**	0.704 ± 0.013	0.686 ± 0.016	0.720 ± 0.016	0.985	0.094
	Squared-hinge	0.694 ± 0.029	0.679 ± 0.029	0.712 ± 0.029	0.367	0.094
	Cox	0.684 ± 0.029	0.671 ± 0.034	0.703 ± 0.024	0.204	0.093
	MTLR	0.698 ± 0.034	0.687 ± 0.029	0.712 ± 0.039	0.355	0.098
	DeepHit	0.617 ± 0.042	0.606 ± 0.039	0.617 ± 0.054	0.703	0.136
	TripleSurv	0.648 ± 0.035	0.653 ± 0.034	0.654 ± 0.043	0.274	0.100

Table A6: Convex linear survival ranking SVM against the deep SCL, per tabular dataset. The convex counterpart of Section 4 is solved to global optimality with a quasi-Newton method on the same matrices and the same pooled five-fold protocol. Harrell C-index, mean $\pm$ standard deviation over ten fold partitions.

Dataset	Convex linear ranking SVM	Deep SCL
METABRIC	0.646 ± 0.008	0.689 ± 0.007
SUPPORT	0.853 ± 0.000	0.878 ± 0.001
GBSG	0.677 ± 0.003	0.670 ± 0.011
ROTTERDAM	0.695 ± 0.001	0.707 ± 0.002
FLCHAIN	0.796 ± 0.001	0.796 ± 0.001
WHAS500	0.768 ± 0.003	0.750 ± 0.010
SEER	0.730 ± 0.004	0.720 ± 0.009
PBC	0.812 ± 0.006	0.799 ± 0.010
LUNG	0.629 ± 0.006	0.611 ± 0.017
TCGA-GBM	0.732 ± 0.006	0.733 ± 0.007
Mean	**0.734**	**0.735**

with $R(i) = \{j : t_j > t_i\}$ and $\text{risk} = 1 - \sum_k p_k t_k^{\text{bin}}$, and Equation 6 combines the three terms as $l = -\alpha\, l_{\text{likelihood}} - \beta\, l_{\text{TAPR}} + \gamma\, l_{\text{calibration}}$. The ranking term therefore enters with a minus and is maximized, which on a comparable pair drives risk_i above risk_j, the concordant direction the surrounding text asks for, and stays finite because the same paper bounds the risk to $[\frac{1}{2K}, \frac{2K-1}{2K}]$. Read on its own, Equation 5 looks like a quantity to minimize, and that is how we first read it: we added it with a plus, which optimizes every comparable pair toward the anti-concordant corner. The paper is self-consistent; the error was ours.

Table A9 reports what it cost. Under the identical protocol, backbone, partitions and epoch budget, with the ranking term as the only difference, the published form beats our earlier one on every dataset. We also report a third form, $\exp(-\sigma[\cdot])$ added with a plus, which is concordant but is a saturating margin rather than the published objective; we list it because an earlier draft of this paper

Table A7: Pooled test C-index under the two checkpoint-selection rules: the per-fold checkpoint chosen by the best validation C-index, and by the minimum validation loss. The discrepancy column is their signed difference and is the quantity plotted in Figure 3; a negative value means the loss-selected checkpoint tested better. This is a paired within-run comparison computed on the checkpoint-selection runs. On the tabular and pathology cohorts those are the same runs as the benchmark, and the select-by-C-index column agrees with Table 1 to within 0.0004 on every cell; on the imaging cohorts both columns come from the canonical execution over ten fold partitions, which is a separate execution from the benchmark runs behind Table 1.

Cohort	Method	Select by val C-index	Select by val loss	Discrepancy
HECKTOR (PET/CT)	**SCL**	0.623	0.622	+0.000
	Squared-hinge	0.619	0.633	-0.015
	Cox	0.612	0.615	-0.003
	MTLR	0.629	0.662	-0.033
	DeepHit	0.552	0.592	-0.039
	TripleSurv	0.601	0.623	-0.022
CRLM (liver CT)	**SCL**	0.581	0.585	-0.004
	Squared-hinge	0.579	0.593	-0.015
	Cox	0.569	0.574	-0.005
	MTLR	0.576	0.599	-0.023
	DeepHit	0.572	0.591	-0.019
	TripleSurv	0.584	0.578	+0.006
UCSF-PDGM (MRI)	**SCL**	0.691	0.697	-0.006
	Squared-hinge	0.701	0.702	-0.001
	Cox	0.686	0.703	-0.017
	MTLR	0.701	0.724	-0.023
	DeepHit	0.704	0.708	-0.005
	TripleSurv	0.665	0.666	-0.001
UPENN-GBM (MRI)	**SCL**	0.621	0.626	-0.005
	Squared-hinge	0.628	0.631	-0.003
	Cox	0.619	0.623	-0.004
	MTLR	0.619	0.644	-0.025
	DeepHit	0.637	0.639	-0.002
	TripleSurv	0.638	0.618	+0.020
Pathology (4 cohorts, mean)	**SCL**	0.679	0.679	-0.000
	Squared-hinge	0.669	0.681	-0.012
	Cox	0.670	0.683	-0.012
	MTLR	0.661	0.669	-0.008
	DeepHit	0.651	0.672	-0.021
	TripleSurv	0.661	0.677	-0.015
Tabular (10 datasets, mean)	**SCL**	0.735	0.736	-0.001
	Squared-hinge	0.730	0.729	+0.001
	Cox	0.730	0.733	-0.003
	MTLR	0.728	0.728	-0.000
	DeepHit	0.698	0.709	-0.011
	TripleSurv	0.713	0.718	-0.005

used it and because it is the natural thing to reach for when the sign is in doubt. All TripleSurv numbers elsewhere in this paper use the published form.

Two limitations of the baseline remain and we state them rather than let them be found. The paper sets σ and ρ by validation performance and β for magnitude matching; we fix $\sigma = \rho = \beta = 1$, as we fix every other baseline's hyperparameters, so the comparison is controlled rather than tuned. And we have not reproduced a published TripleSurv number, there being no reference implementation to check against.

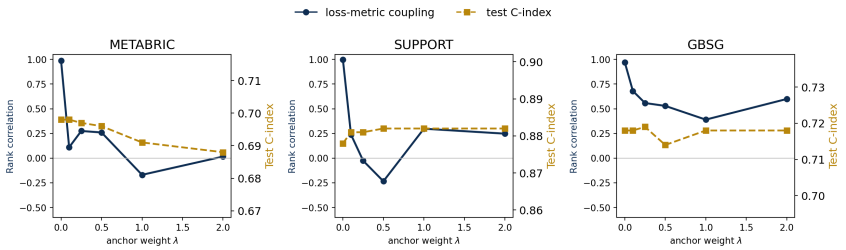


Figure A3: Effect of the likelihood-anchor weight λ in the hybrid loss. This sweep uses each dataset's own fixed train/validation split, two seeds and 120 epochs, rather than the five-fold protocol over ten fold partitions of the main results. At every positive anchor weight we tested, the coupling falls sharply, from 0.97–1.00 at $\lambda = 0$ to 0.11–0.68 at the smallest positive weight ($\lambda = 0.1$), and it turns negative at intermediate weights on two of the three cohorts, reaching -0.23, while the test concordance index is flat in λ to within 0.010. The sweep does not establish behaviour for arbitrarily small positive weights. The dependence is not monotone. Coupling is worst at intermediate weights, where the two terms most contend, and recovers only partially at large weights as the objective approaches a pure likelihood.

Table A8: Loss families on the two axes that define our claim, on the ten tabular datasets. Discrimination is the pooled Harrell C-index and coupling the mean absolute loss-metric rank correlation, both over the canonical ten fold partitions for the rows the benchmark runs cover; the ReLoss rows come from their own run over ten partitions, and the survival SVM, X-CAL and survival-games rows from their own runs over three. The survival SVM is convex-solved and has no training trajectory, so coupling is undefined. The likelihood row summarizes Cox, MTLR, DeepHit, and TripleSurv (Table 1).

Category	Loss	Harrell C	Coupling
Likelihood	Cox / MTLR / DeepHit / TripleSurv	0.70–0.73	0.36–0.57
Ranking, margin	Survival SVM (scikit-survival)	0.725	n/a
Ranking, sigmoid	**SCL (ours)**	**0.735**	**0.97**
Ranking, multi-horizon	SCL per horizon (ours)	0.729	0.79
Ranking, listwise	GRAFT	0.722	0.53
Learned alignment	ReLoss (alone / with Cox)	0.618 / 0.729	0.22 / 0.50
Calibration	X-CAL / survival games	0.653 / 0.698	0.27 / 0.19

B.8 When the by-loss rule wins, and why it is not about value-monotonicity

Switching the checkpoint rule from validation C-index to minimum validation loss helps the baselines on some cohorts and hurts them on others, and the asymmetry deserves an explanation rather than a shrug. It is a property of the variance of the selection statistic, not of the loss. A validation C-index computed on a fold carrying few events is noisy, so selecting its peak fits that fold's noise; the loss is a mean over every comparable pair or subject and is the smoother statistic. Where events are scarce the smoother rule wins whatever the objective was.

The prediction is that the baselines' by-loss gain should fall as the number of events per validation fold rises. It does, in direction, at Spearman -0.40 over the eighteen datasets, which at this many datasets does not reach significance ($p = 0.099$). The largest gains sit on the smallest cohorts, $+0.026$ on TCGA-UCEC with fifteen events per validation fold and $+0.022$ on HECKTOR with thirty; the three cohorts where the by-loss rule costs the baselines most carry 33, 98 and 160 events. We therefore offer the event count as the most plausible account of a scattered effect and not as an established one. What the table does establish is the scatter itself, since the gain changes sign across cohorts with no regularity a practitioner could exploit in advance. The SCL is exempt from the whole effect because the two rules select the same checkpoint for it on most folds. Its own discrepancy never exceeds 0.0061 against 0.064 for the baselines, and we do not report its correlation with the event count, since every value sits at the fourth decimal and any ordering among them is noise.

Table A9: TripleSurv under three readings of its ranking term, on the ten tabular datasets: the form this repository used before Equation 6 was read, the published form, and the saturating margin an earlier draft of this paper used. One protocol, one set of partitions, the ranking term the only difference. Best per row in bold.

Dataset	Added with a plus	As published	Saturating margin	Published − plus
FLCHAIN	0.555	0.745	**0.748**	+0.190
GBSG	0.615	0.647	**0.649**	+0.032
LUNG	0.588	**0.593**	0.593	+0.005
METABRIC	0.606	0.672	**0.673**	+0.065
PBC	0.734	0.787	**0.791**	+0.053
ROTTERDAM	0.659	0.695	**0.697**	+0.036
SEER	0.638	**0.692**	0.684	+0.054
SUPPORT	0.853	0.863	**0.867**	+0.010
TCGA-GBM	0.693	**0.715**	0.714	+0.023
WHAS500	0.676	0.726	**0.732**	+0.050
Mean	0.662	0.713	0.715	+0.052

Two things follow for how the result should be read. A reader who finds the by-loss rule ahead on some dataset has found a fact about that dataset's event count, and one who averages such differences into a second ranking is averaging effects of opposite sign, which is why we report the paired discrepancy rather than a by-loss ranking in the main text. The claim that survives is the one the mechanism does not touch: the SCL is the only loss whose two rules agree everywhere, so it is the only one for which the choice of rule never has to be made.

B.9 Does the property survive a multi-horizon extension?

A reader will ask what happens if the surrogate is applied at every horizon, since the SCL scores a single scalar risk and the integrated Brier scores in Table 1 come from a Breslow baseline hazard fitted on the training fold rather than from the loss itself. We ran that extension. It is a hypothesis we tested, not a method we propose, and the theory of Section 3 predicts its outcome before any of it is measured. Writing h_k for logistic-hazard logits and $R_k = 1 - \prod_{j \le k} \big(1 - \sigma(h_j)\big)$ for the cumulative incidence at bin k, the multi-horizon loss is

$$L_\tau^{\mathrm{curve}}(f) \;=\; \frac{1}{K} \sum_{k=1}^{K} L_\tau\big(R_k\big),$$

the mean of K sigmoid concordance losses, one per horizon, on a model that emits a full discrete survival curve. Nothing else changes: the same backbone, temperature, optimizer, fold partitions and epoch budget.

The theory predicts that this costs value-monotonicity, and says why. Theorem 1 makes one SCL term an approximation of $1-C$ for the score it is applied to, so the mean of K terms approximates $1-\frac{1}{K}\sum_k C(R_k)$, the average concordance of the K per-horizon risk scores. What a reader compares the curve against is the concordance of the reported risk, which is the C-index of the *averaged* score $\frac{1}{K}\sum_k R_k$. The average of the concordances and the concordance of the average agree only to the extent that the horizons order patients alike, and they need not, since a model may rank a pair one way early and the other way late. The tolerance therefore widens with the disagreement among horizons, a slack the single-horizon loss does not carry.

The measurements follow that prediction in direction, and they leave the property recognisable rather than destroyed. Coupling falls from 0.97 to 0.79 on the tabular pool, and by 0.15 and 0.13 on the pathology and imaging cohorts, yet the multi-horizon form still clears 0.5 on every one of the eighteen datasets and reverses sign on two per cent of folds, against a fifth to a third for the likelihood losses. Discrimination is not what separates the two forms. The C-index differs by 0.007 and the time-dependent AUC by 0.006, within the spread between fold partitions. Calibration, which is the only reason to want the curve directly, is where the extension does worst. Its integrated Brier score is 2.3 times the scalar loss's, and on nine of the ten datasets it is worse than predicting the marginal Kaplan-Meier curve for every patient. SEER is the clearest case. The multi-horizon model reaches a

Table A10: The single-horizon SCL against its multi-horizon extension on the ten tabular datasets. One invocation of one script, the same ten fold partitions, the loss the only difference. Coupling is the mean absolute per-fold rank correlation between the negative validation loss and the validation C-index. The last row counts the datasets whose integrated Brier score exceeds that of the marginal Kaplan-Meier curve, which uses no covariates at all; that reference is fitted on the whole cohort while the models are scored out of fold, so it is optimistic and losing to it is the conservative reading.

	SCL (scalar, Breslow curve)	SCL applied per horizon
Harrell C-index	**0.735**	0.729
Time-dependent AUC	**0.778**	0.772
Integrated Brier score	**0.146**	0.333
Coupling: mean $\|r\|$ [range]	**0.97** [0.91, 1.00]	0.79 [0.62, 0.95]
Datasets with coupling below 0.5	**0/10**	**0/10**
Datasets worse than the KM reference	**1/10**	9/10

C-index of 0.707 and an integrated Brier score of 0.435 against the covariate-free reference's 0.078, ranking patients well while placing their risk at the wrong level entirely. The calibration loss is not universal, and we report where it is absent: on the four imaging cohorts the multi-horizon score is 0.25 against the scalar loss's 0.16, while on the four pathology cohorts the two are level at 0.145 and 0.143, and on TCGA-GBMLGG the multi-horizon form is the better calibrated of the two. That exception is evidence against the account below rather than a small-sample nuisance, since the Brier score does separate the methods on those cohorts elsewhere in Table 1, and we do not have an explanation for it. On the two pools where the loss does appear, the mechanism is the one the loss makes plain, since each per-horizon term depends on the scores only through their differences and is therefore indifferent to the level of the hazard, so nothing in the objective anchors the predicted probabilities. The Breslow baseline that the scalar loss relies on supplies exactly that anchor.

We therefore report the multi-horizon form as a limitation rather than a recommendation, and it is why the main text keeps to the scalar loss with a Breslow baseline where a survival curve is needed. It is also the sharpest available test of the account in Section 3, because the manipulation is confined to the loss. The hybrid of Section 4 shows that adding a likelihood term collapses the coupling; the multi-horizon form shows that averaging concordance surrogates over horizons erodes it by a smaller amount. Both act through the same mechanism of putting distance between the loss value and any single concordance index, and the size of the erosion tracks how much distance is introduced.

B.10 When can the loss value be read as the metric?

Theorem 1 bounds the approximation error by $2B\tau$, where B bounds the density of comparable-pair score gaps near zero. The practical question is what to measure in place of B. An earlier version of this paper suggested keeping the standard deviation of the risk well above τ. That is the wrong quantity in principle: a single outlier inflates the spread while leaving most comparable pairs nearly tied. Six uncensored subjects with risks $(0.004, 0.003, 0.002, 0.001, 0, -100)$ have standard deviation 40.8 at $\tau = 0.1$ and a concordance index of 1, yet the SCL is 0.330 rather than near zero, because two thirds of the comparable pairs have a gap below τ.

The quantity the theorem actually controls is the mass of comparable-pair gaps near zero, which is estimated directly by the *near-tie fraction*, the proportion of comparable pairs with $|f_i - f_j| < \tau$. Table A11 reports it for every training loss across eight tabular cohorts and the four imaging cohorts of the canonical run, alongside the observed approximation error $|L_\tau - (1 - C)|$ computed on the same predictions. The near-tie fraction predicts the error closely, with a rank correlation of $+0.97$ over the 72 cells, and it separates cleanly. On the 39 cells with a near-tie fraction below 0.10 the approximation error never exceeds 0.005, while on the 13 cells above 0.50 it ranges from 0.031 to 0.120. A practitioner can compute it in one pass over the validation predictions and use 0.1 as a working threshold.

In our runs the risk standard deviation happens to predict the error equally well, with a rank correlation of -0.96, because none of these models produces the outlier configuration that defeats it. We report the near-tie fraction rather than the spread because it is the quantity the bound is stated in and

cannot be defeated that way, and we present the threshold as an empirical working rule rather than a guarantee.

Table A11: Near-tie fraction and observed approximation error by training loss, over eight tabular cohorts and the four imaging cohorts of the canonical run, giving 72 cells. The last column gives how many cells each row averages, since the squared hinge, the hybrid and the multi-horizon variant were run on a subset of the cohorts. The near-tie fraction is the proportion of comparable pairs whose absolute score gap is below $\tau = 0.1$. The two losses whose value cannot be read as the metric are exactly the two that emit a bounded derived risk.

Training loss	Risk sd	Near-tie fraction	$\lvert L_\tau - (1 - C)\rvert$	Cells
MTLR	30.2	0.017	0.0003	12
Cox	1.74	0.050	0.0019	12
SCL	1.32	0.064	0.0024	12
Hybrid (ours)	1.19	0.068	0.0022	4
Squared-hinge	0.742	0.113	0.0055	4
SCL-curve (ours)	0.263	0.234	0.0160	4
TripleSurv	0.163	0.441	0.0508	12
DeepHit	0.118	0.548	0.0708	12

B.11 Does the choice of training window decide the coupling result?

The coupling in the body is measured over the second half of training, on the argument that the early epochs are a period in which the loss and the metric move together for reasons that say nothing about either. That is a choice, and on a cohort that converges early it is a choice a reader may suspect of doing the work. Table A12 recomputes the same correlation on five windows of the same runs. It covers the ten tabular datasets rather than all eighteen, because the measurement needs the loss and the metric at every epoch and the benchmark runs store only the summary; the tabular pool is also where the concern bites hardest, holding the smallest cohorts in the study and therefore the ones that converge earliest.

The ordering of the losses is the same in every window and the SCL's own value barely moves, between 0.96 and 0.98. The window that flatters every loss is the one running to the minimum of the validation loss, which is also the window a practitioner actually occupies: the loss falls across almost all of it by construction, so anything moving with it correlates well, and Cox rises from 0.54 on the first half to 0.79 there while DeepHit rises from 0.33 to 0.56. The SCL still leads it by 0.11 and is still the only loss above 0.9. The second half is not the window most favourable to us; what widens the gap there is the baselines falling away, MTLR from 0.72 to 0.46, which is the behaviour Section 3 predicts once the ranking has stabilised.

Table A12: Mean absolute rank correlation between the negative validation loss and the validation C-index, by the window it is measured on, over the ten tabular datasets. From the 150-epoch trajectory captures, which record every epoch, so these are one partition per dataset rather than the ten behind Table A2; the question here is whether the window changes the conclusion, not what the value is. Best per column in bold.

Loss	To loss minimum	First half	Second half	All epochs	After loss minimum
SCL	**0.96**	**0.97**	**0.98**	**0.98**	**0.97**
Squared-hinge	0.85	0.81	0.73	0.79	0.79
Cox	0.79	0.54	0.63	0.63	0.67
MTLR	0.63	0.72	0.46	0.62	0.50
TripleSurv	0.63	0.39	0.48	0.52	0.53
DeepHit	0.56	0.33	0.27	0.34	0.32

The practical form of the same question is where each rule would stop training. Table A13 gives the epoch at which the validation loss is lowest against the epoch at which the validation C-index is highest. For the SCL they are the same epoch on seven of the ten datasets and never more than six apart, a mean gap of 1.4 epochs. Every other loss is between 30 and 54, and the direction is

not consistent: Cox stops a median of 32 epochs early while MTLR keeps descending for a median of 44 epochs after the C-index has peaked. DeepHit is the case that a difference of averages would hide, since its two mean epochs are 1.6 apart while its mean per-dataset gap is 37.8. None of this is visible from the loss curve.

Table A13: Epoch of lowest validation loss against epoch of highest validation C-index, over the ten tabular datasets. The gap is how far a practitioner stopping on the loss lands from the best checkpoint, and it is summarised over the per-dataset gaps rather than taken between the two averaged epochs: a loss that stops early on some datasets and late on others cancels in the latter and looks accurate when it is not.

Loss	Lowest val. loss	Highest val. C-index	|gap| per dataset		
	epoch	epoch	mean	median	worst
SCL	48.9	48.7	**1.4**	**0.0**	**6**
Squared-hinge	29.4	57.5	30.5	6.0	142
Cox	20.5	52.2	32.1	31.5	83
MTLR	122.4	84.0	42.8	44.5	135
TripleSurv	53.4	68.7	54.1	60.0	119
DeepHit	50.8	49.2	37.8	26.0	86

B.12 Is it the training objective or the monitoring signal?

The experiments in the main text confound two things: training with the SCL, and having a concordance-based quantity available to read. We separate them. For every training loss we additionally evaluate the SCL on the validation predictions at each checkpoint, on the same scalar risk the C-index is computed from, and select the checkpoint by that value instead. A third variant standardizes the validation risk to zero mean and unit variance before computing the monitoring SCL. Table A14 reports all of it over ten tabular cohorts, six training losses, and the same ten fold partitions as the main results.

Which value is read is what does the work, and it does not do all of it. Selecting on a model's own loss deviates from the reference rule by up to 0.064 of C-index and by more than 0.01 on 14 of the 50 non-SCL cells. Selecting on the validation SCL leaves the worst case slightly higher, 0.072, while cutting the count to 10 cells; standardizing the risk first brings both down, to 0.027 and 3 cells. Cox's worst deviation falls from 0.0094 to 0.0054 and MTLR's from 0.0163 to 0.0021, with coupling rising from 0.51 to 0.96 and from 0.29 to 0.98. A practitioner who keeps the Cox loss and changes only which number they read therefore recovers most of the selection quality, and the two cases where reading the SCL instead makes matters worse are the ones the next paragraph explains.

The cases where monitoring fails are the ones Remark A1 predicts, and the near-tie fraction of Appendix B.10 identifies them. DeepHit and TripleSurv emit a bounded derived risk on which 0.55 and 0.44 of comparable pairs fall within τ of a tie, so the sigmoid never saturates, the surrogate hovers near $1/2$, and its coupling is only 0.50 and 0.69. On DeepHit the monitored value is worse than the model's own loss, 0.0724 against 0.0643. Standardizing the risk lifts every one of the six worst cells, from the range $[0.21, 0.42]$ to the range $[0.62, 0.99]$, for example 0.207 to 0.616 on SUPPORT with DeepHit and 0.257 to 0.884 on TCGA-GBM with TripleSurv. The opposite extreme also fails, in the opposite direction. MTLR's raw risk has a mean standard deviation near 10^6, which saturates the sigmoid so completely that the SCL equals $1 - C$ and coupling reaches 0.999, and standardizing *down* to unit variance costs it, dropping to 0.429 on SUPPORT. The rule is not to standardize but to keep the near-tie fraction small, which is the practical form of Remark A1 and which standardizing achieves in one direction and destroys in the other.

We draw the honest consequence. Pushed to its limit that rule makes the monitoring surrogate the concordance index itself, and monitoring requires no differentiability, so reading the SCL is not better than computing the C-index when a validation set is available. What this experiment establishes is therefore narrower than the framing of a new loss: the negative result stands, in that which of the two signals is read moves the deployed checkpoint by up to 0.064 of C-index unpredictably; the remaining distinction of the SCL is that the training objective and the monitoring signal are the same object, so the value a practitioner already watches is readable as the metric without a second eval-

uation pathway; and the near-tie condition is a new and testable criterion for when a concordance surrogate can be read at all.

Table A14: Separating training from monitoring, over ten tabular cohorts $\times$ six training losses $\times$ ten fold partitions. "Risk sd" is the mean standard deviation of the validation risk, to be read against $\tau = 0.1$. The three deviation columns give the worst absolute difference in pooled test C-index between the reference rule (select by validation C-index) and selecting by the model's own validation loss, by the validation SCL, and by the validation SCL on a standardized risk. The three coupling columns are the corresponding mean loss-metric rank correlations.

		Worst |deviation| selecting by			Mean coupling of		
Training loss	Risk sd	own loss	val SCL	std. SCL	own loss	val SCL	std. SCL
SCL	1.85	0.0026	0.0026	0.0021	$+0.97$	$+0.97$	$+0.94$
Squared-hinge	0.59	0.0117	0.0061	0.0008	$+0.64$	$+0.91$	$+0.95$
Cox	1.44	0.0094	0.0054	0.0024	$+0.51$	$+0.96$	$+0.95$
MTLR	996286.71	0.0163	0.0021	0.0132	$+0.29$	$+0.98$	$+0.75$
DeepHit	0.10	0.0643	0.0724	0.0265	$+0.18$	$+0.50$	$+0.92$
TripleSurv	0.18	0.0254	0.0232	0.0034	$+0.29$	$+0.69$	$+0.91$

B.13 Paired bootstrap of the C-index difference on the imaging cohorts

Two things make a reported difference uncertain, and an interval has to cover both. Patients are a sample, and the pipeline is not reproducible, so retraining the identical configuration lands somewhere else. An interval for one method alone covers neither, which is what an earlier version of this analysis reported.

The canonical run covers each cohort with ten event-stratified fold partitions, every method scoring the same patients within a partition. Each bootstrap replicate therefore resamples two levels: the patients once with replacement, with every method scored on that same sample, and the ten partitions with replacement. The partition is the larger of the two sources of spread, 0.015 against 0.006 between repeated executions of the identical configuration (Appendix B.16), and an earlier version resampled the executions while averaging over the partitions, which held the dominant component fixed and gave intervals that were too narrow. The percentile interval over $10{,}000$ replicates therefore covers patient sampling and the partition together, leaving only the smaller retraining component outside it. We raised the replicate count from 1000 after finding that a comparison on CRLM changed verdict between 1000 and 2000 replicates.

Four of the twenty comparisons against a published loss resolve, and they do not all go one way. The SCL is ahead of DeepHit on HECKTOR by 0.070 and of TripleSurv on UCSF-PDGM by 0.026; DeepHit and TripleSurv are both ahead of it on UPENN-GBM, by 0.015 and 0.017. The remaining sixteen intervals contain zero. The table also lists eight comparisons against our own two variants, which that count excludes. This is weaker than earlier versions of the same analysis reported, and the reasons are instructive: we now propagate the fold partition as well as the patients, and we corrected the sign of the TripleSurv ranking term (Appendix B.7). The honest summary is that imaging discrimination is unresolved at these cohort sizes, and that the two cohorts where it does resolve point in opposite directions.

We record this because it is the reason for running the analysis at all. Point estimates on these cohorts move by more than the differences between methods, so any ordering read off them is not reproducible, and we do not report one.

B.14 Does preprocessing cross the cross-validation boundary?

It does, and we measured the consequence rather than assuming it away. The stored feature matrices for the tabular cohorts were standardized once over the whole cohort before any splitting, and our folds are drawn from that pooled matrix, so subjects that end up in a test fold contributed to the preprocessing statistics. Only feature means and variances leak, not outcomes or event times, which bounds how much can go wrong but does not make it correct. The same holds for the clinical

Table A15: Paired bootstrap of the C-index difference on the four imaging cohorts, $10{,}000$ replicates, resampling the patients and the ten fold partitions together. Positive favours the SCL. An asterisk marks an interval excluding zero. Four of the twenty comparisons against a published loss do, two favouring the SCL and two favouring the baseline, and none of the eight against our own variants.

Cohort	Baseline	SCL − baseline	95% CI
HECKTOR	Cox	$+0.0104$	$[-0.0157, +0.0376]$
	MTLR	-0.0061	$[-0.0306, +0.0183]$
	DeepHit	$+0.0703$	$[+0.0345, +0.1054]^{*}$
	TripleSurv	$+0.0215$	$[-0.0099, +0.0529]$
	Squared-hinge	$+0.0040$	$[-0.0157, +0.0230]$
	Hybrid (ours)	-0.0031	$[-0.0366, +0.0302]$
	SCL-curve (ours)	-0.0081	$[-0.0429, +0.0275]$
CRLM	Cox	$+0.0123$	$[-0.0160, +0.0409]$
	MTLR	$+0.0050$	$[-0.0294, +0.0385]$
	DeepHit	$+0.0089$	$[-0.0247, +0.0423]$
	TripleSurv	-0.0024	$[-0.0370, +0.0318]$
	Squared-hinge	$+0.0025$	$[-0.0171, +0.0218]$
	Hybrid (ours)	$+0.0168$	$[-0.0132, +0.0475]$
	SCL-curve (ours)	$+0.0158$	$[-0.0143, +0.0458]$
UCSF-PDGM	Cox	$+0.0050$	$[-0.0183, +0.0280]$
	MTLR	-0.0098	$[-0.0304, +0.0102]$
	DeepHit	-0.0130	$[-0.0380, +0.0119]$
	TripleSurv	$+0.0259$	$[+0.0037, +0.0474]^{*}$
	Squared-hinge	-0.0101	$[-0.0289, +0.0084]$
	Hybrid (ours)	-0.0120	$[-0.0329, +0.0091]$
	SCL-curve (ours)	-0.0079	$[-0.0357, +0.0201]$
UPENN-GBM	Cox	$+0.0025$	$[-0.0096, +0.0140]$
	MTLR	$+0.0026$	$[-0.0130, +0.0195]$
	DeepHit	-0.0154	$[-0.0273, -0.0033]^{*}$
	TripleSurv	-0.0165	$[-0.0310, -0.0020]^{*}$
	Squared-hinge	-0.0065	$[-0.0190, +0.0063]$
	Hybrid (ours)	-0.0082	$[-0.0206, +0.0049]$
	SCL-curve (ours)	-0.0125	$[-0.0271, +0.0021]$

covariates of the imaging cohorts, which are loaded from a tensor standardized over the whole cohort; the voxel intensities are normalized per scan and do not cross the boundary.

To size it we took the three cohorts whose raw, unstandardized features are available from scikit-survival, and ran the same model on the same folds twice: once with the scaler fitted over the whole cohort before splitting, reproducing our situation, and once with the scaler refitted inside each training fold, which is the correct protocol. Over nine dataset-and-loss cells the difference in pooled out-of-fold C-index has mean $+0.0002$ and largest absolute value 0.0066, and it has no consistent sign: it favours the pooled fitting on seven cells and the per-fold fitting on two. The largest single deviation, -0.0066 for MTLR on GBSG, favours the correct protocol. This is an order of magnitude below the run-to-run variability of Appendix B.16 and does not change any ordering in the paper. We measured it only on the three cohorts whose raw features scikit-survival exposes, so the figure is an estimate of the magnitude on cohorts of this kind and not a direct measurement on the imaging or pathology cohorts, whose raw features we cannot reconstruct. The loader now exposes an unscaled path for callers that re-partition the data, and the diagnostic is `experiments/leakage_diagnostic.py`.

B.15 Does pooling out-of-fold scores distort the C-index?

The protocol trains one model per outer fold and concatenates the five test-fold risk vectors into one out-of-fold vector, on which each metric is computed once. Scores from different folds come from independently trained models and share no common scale or offset, so pooling creates cross-fold comparisons whose concordance depends in part on those offsets rather than on any within-fold ranking. We measured the size of the effect by computing, from the same predictions, the pooled

Table A16: Effect of fitting the feature scaler over the whole cohort rather than inside each training fold, on the three cohorts whose raw features are loadable. This asks whether the protocol distorts an estimate rather than comparing losses, so it runs on three of them. Pooled out-of-fold Harrell C-index, mean over three seeds. A positive bias means the leaked fitting reported the higher number.

Dataset	Loss	Scaler fitted per fold	Fitted over the cohort	Bias
GBSG	SCL	0.6590	0.6599	$+0.0009$
	Cox	0.6559	0.6604	$+0.0045$
	MTLR	0.6556	0.6490	-0.0066
WHAS500	SCL	0.7419	0.7424	$+0.0004$
	Cox	0.7361	0.7378	$+0.0017$
	MTLR	0.7414	0.7421	$+0.0007$
FLCHAIN	SCL	0.7921	0.7924	$+0.0003$
	Cox	0.7934	0.7935	$+0.0002$
	MTLR	0.7888	0.7887	-0.0001

estimate the paper reports, the mean of the C-index computed inside each test fold, and the pooled estimate after standardizing each fold's scores to zero mean and unit variance. Table A17 reports the result over eight tabular cohorts, three losses, and three seeds.

Pooling biases the C-index *downward* by a mean of 0.006, it does so in 70 of the 72 cells, and the largest single deviation is 0.037. Fourteen of the cells deviate by more than 0.01 and three by more than 0.02. Per-fold standardization removes almost all of it for the two scalar-risk losses, from -0.0052 to -0.0010 for the SCL and from -0.0072 to -0.0007 for Cox, and does not help MTLR, whose deviation stays at -0.0049; averaged over the three the mean deviation falls from -0.0058 to -0.0022. For the scalar-risk losses the mechanism is therefore the cross-fold offsets rather than anything about the ranking. Two properties make this harmless for the comparisons in the paper. The bias is small relative to the run-to-run variability of Appendix B.16, and it is nearly the same for every loss, -0.0052 for the SCL against -0.0072 for Cox and -0.0052 for MTLR, so it shifts all methods together rather than favouring one. The pooled numbers we report are therefore slightly conservative rather than optimistic, and no ranking in the paper turns on the difference.

Table A17: Effect of pooling out-of-fold scores across folds, over eight tabular cohorts $\times$ three losses $\times$ three seeds. This too is a question about the protocol rather than about which loss wins, which is why it does not carry the full set. "Pooled" is the estimate the paper reports; "within" averages the C-index computed inside each test fold; "standardized" pools after z-scoring each fold's scores. Deviations are pooled minus within. A negative value means the pooled estimate is the lower one.

Loss	Mean deviation	Worst $\lvert$deviation$\rvert$	Mean after standardizing
SCL	-0.0052	0.0369	-0.0010
Cox	-0.0072	0.0340	-0.0007
MTLR	-0.0052	0.0284	-0.0049

B.16 Two sources of variability, and their sizes

No deterministic-algorithm flags are set, so 3D convolution and reduction order vary between runs and the pipeline does not reproduce bitwise even at identical seeds. A standard deviation over three seeds drawn from one run therefore does not describe all the uncertainty in a reported number. We repeated the entire protocol four times on each imaging cohort and separated the two components: the spread over seeds within a run, and the spread of the run-level means across runs.

The within-run seed spread is the larger of the two. Averaged over the twenty cohort-and-loss cells it is 0.0152, against 0.0066 between runs, and it is the larger on nineteen of the twenty. The seed standard deviation usually reported was therefore not understating the uncertainty here; it was missing a smaller second component. Combining them in quadrature gives the totals in Table A18, which range from 0.009 to 0.033 and are what the paired bootstrap of Appendix B.13 propagates.

The practical consequence is that no imaging difference below roughly 0.02 of concordance index is resolvable at these cohort sizes, which is why Section 5 reports paired intervals rather than a single run's ordering.

Table A18: How much of the imaging uncertainty is retraining nondeterminism rather than the fold partition, measured by repeating the three-seed protocol four times. The main results use a single execution over ten partitions instead; this table exists to size the component that design does not separate. "Seed" is the within-run standard deviation over three seeds, averaged across runs; "run" is the standard deviation of the four run-level means; "total" combines them in quadrature. TripleSurv is absent because its repeated executions predate the sign correction of Appendix B.7, and we do not mix the two implementations.

Cohort	Method	Mean C	Seed sd	Run sd	Total sd
HECKTOR	**SCL**	0.643	0.014	0.010	0.017
	Cox	0.619	0.016	0.015	0.022
	MTLR	0.650	0.017	0.004	0.017
	DeepHit	0.560	0.014	0.007	0.016
	Squared-hinge	0.618	0.017	0.007	0.019
CRLM	**SCL**	0.590	0.009	0.003	0.009
	Cox	0.590	0.014	0.008	0.016
	MTLR	0.601	0.031	0.004	0.031
	DeepHit	0.560	0.019	0.002	0.019
	Squared-hinge	0.585	0.011	0.005	0.012
UCSF-PDGM	**SCL**	0.693	0.022	0.007	0.023
	Cox	0.680	0.033	0.004	0.033
	MTLR	0.704	0.012	0.009	0.016
	DeepHit	0.695	0.021	0.006	0.022
	Squared-hinge	0.700	0.012	0.007	0.014
UPENN-GBM	**SCL**	0.625	0.011	0.006	0.012
	Cox	0.623	0.009	0.007	0.011
	MTLR	0.616	0.009	0.013	0.016
	DeepHit	0.637	0.008	0.005	0.009
	Squared-hinge	0.626	0.008	0.004	0.009

B.17 The loss value as a readable quantity

The introduction claims that a likelihood value carries no reading as a level of concordance. Table A19 makes this concrete. For a fixed trained model and a fixed cohort we vary only the batch size over which the loss is averaged and record the mean minibatch value, alongside the value computed on the full cohort. The SCL value is stable across two orders of magnitude of batch size. It agrees with the full-cohort value to within 0.019 at the smallest batch and within 0.007 once $B \geq 64$, and it tracks the minibatch $1 - C$ to within 0.035 at $B = 8$ and 0.007 from $B = 64$ upward, the residual at small B being the bias of the pair average over few pairs rather than a dependence of the loss on the batch size. The Cox partial-likelihood value changes by a factor of 4.2 to 6.9 over the same range, because each term normalizes over a risk set whose size grows with the batch, and it reaches its own full-cohort value on one of the five cohorts. Two runs of the same model on the same data therefore report different Cox losses if they use different batch sizes, and neither number can be compared with a value from another cohort.

B.18 GRAFT across all four modalities

Table A8 reports GRAFT (Ashhad et al., 2026) on the tabular pool, where it sits alongside the other loss families. Table A20 gives the comparison across the fifteen cohorts it was run on, from the canonical benchmark: five-fold event-stratified cross-validation, pooled out-of-fold predictions, ten fold partitions, and one backbone per modality. GRAFT is close behind the SCL on discrimination, by 0.013 on the tabular pool, 0.027 on pathology and 0.027 on HECKTOR, and it couples at 0.53, above every likelihood baseline and far below the SCL's 0.97.

Table A19: The loss value as a function of the averaging batch size, for one fixed trained model on each cohort, the model trained with the SCL. The SCL column gives the range of the mean minibatch value over all batch sizes tested, next to the value on the full cohort and $1 - C$ on the full cohort. The Cox column gives the same range for the negative log partial likelihood. The SCL value is batch-invariant and equals $1 - C$; the Cox value is neither.

			SCL			Cox partial likelihood	
Cohort	n	Batch sizes	over batches	full	$1 - C$	over batches	full
SUPPORT	9105	8–1024	0.114–0.123	0.122	0.121	0.87–5.66	8.18
FLCHAIN	7874	8–1024	0.187–0.206	0.204	0.204	1.47–10.09	19.95
METABRIC	1904	8–1024	0.288–0.308	0.307	0.305	0.93–5.68	6.31
TCGA-BRCA	950	8–512	0.367–0.389	0.371	0.369	1.23–5.12	5.73
TCGA-GBMLGG	573	8–512	0.173–0.190	0.181	0.180	1.06–5.41	5.52

Table A20: GRAFT (Ashhad et al., 2026) against the SCL and the likelihood baselines on the fifteen cohorts GRAFT was run on: ten tabular, four pathology and HECKTOR. Every column comes from the canonical benchmark run behind Table 1, so the two tables agree cell by cell. Coupling is the mean absolute loss-metric rank correlation over all fifteen cohorts. GRAFT couples more tightly than any likelihood baseline but substantially less tightly than the SCL, consistent with the distinction between a Spearman rank correlation and pairwise concordance.

	Harrell C-index			Coupling
Loss	Tabular	Pathology	Imaging	(all cohorts)
SCL (ours)	**0.735**	**0.679**	0.611	**0.97**
GRAFT	0.722	0.652	0.584	0.53
Cox	0.730	0.670	0.612	0.45
MTLR	0.728	0.661	**0.642**	0.49
DeepHit	0.698	0.651	0.557	0.38
TripleSurv	0.713	0.661	0.610	0.39

B.19 Risk stratification and robustness to censoring

Figure A4 splits patients at the median out-of-fold predicted risk of the SCL on the four imaging cohorts. Figure A5 adds synthetic censoring to the training data of two tabular cohorts by randomly re-labelling a fraction of the training subjects as censored.

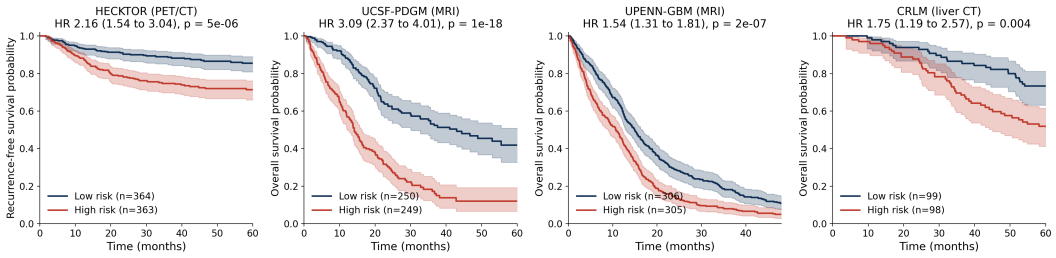


Figure A4: Kaplan-Meier curves for patients split at the median out-of-fold predicted risk of the SCL on the four imaging cohorts. Hazard ratios compare the high-risk with the low-risk group and come from a univariable Cox model on that binary split. The groups separate significantly on every cohort by the log-rank test, with hazard ratios of 2.16 on HECKTOR for recurrence-free survival (95% confidence interval 1.54 to 3.04), 3.09 on UCSF-PDGM (2.37 to 4.01), 1.54 on UPENN-GBM (1.31 to 1.81), and 1.75 on CRLM (1.19 to 2.57) for overall survival.

B.20 Loss-metric coupling on the remaining cohorts

These figures repeat the trajectories of Figure 2 on nine additional cohorts. The dashed line is the data-fit training loss, the solid line the data-fit validation loss, and the gold line the validation C-

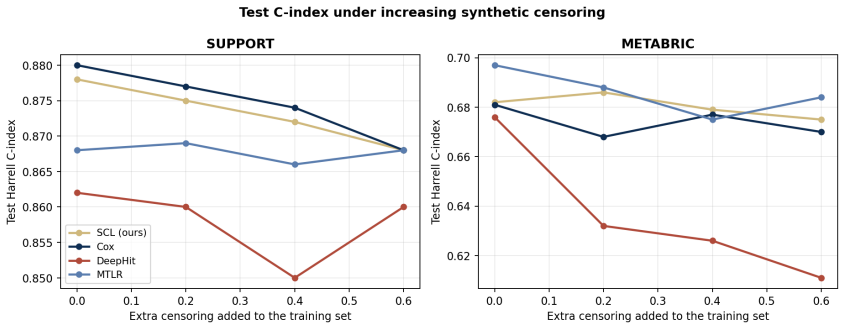


Figure A5: Test C-index as increasing synthetic censoring is added to the training data of two tabular cohorts, three seeds per point. The cohorts separate the methods differently. On SUPPORT, whose base event rate is 0.68, every method loses at most 0.012 across the whole range and the SCL is not the most robust. On METABRIC, whose base event rate is 0.42, the SCL varies by 0.011 and ends 0.007 below where it started, while DeepHit loses 0.065, the Cox likelihood 0.011 and MTLR 0.013. The pattern is consistent with the loss being defined only over comparable pairs, which uses the censoring structure directly rather than modelling the full event distribution, but one cohort of each kind is not enough to establish it.

index; the gold star and navy circle mark the checkpoint that would be selected by the best validation C-index and by the minimum validation loss. Across these cohorts the SCL stays closely coupled to the C-index, while the likelihood and margin losses couple substantially less tightly or unstably.

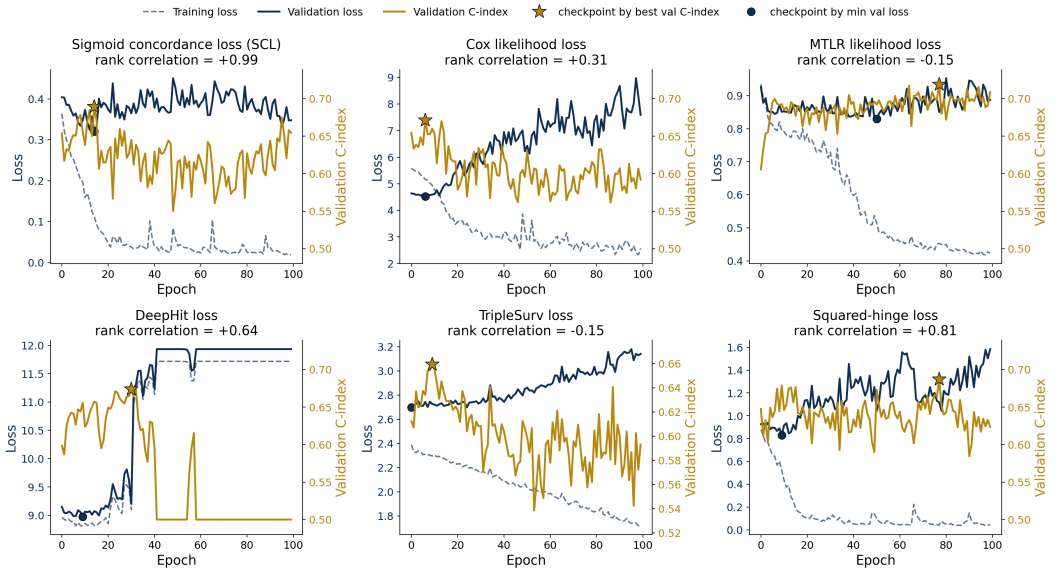


Figure A6: Loss-metric coupling on HECKTOR head-and-neck CT and PET (727 patients). Curves and markers are as in Figure 2.

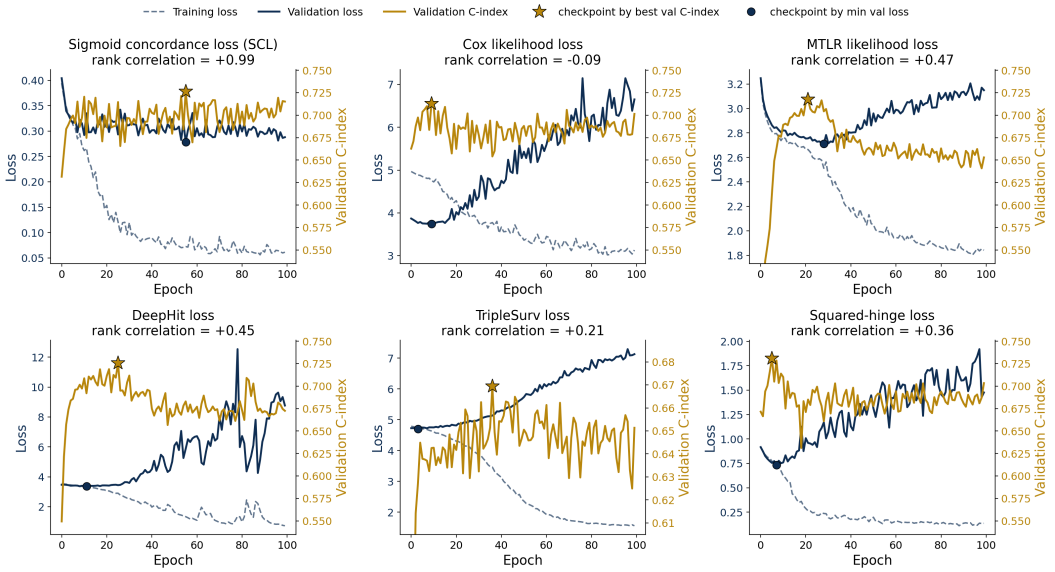


Figure A7: Loss-metric coupling on UPENN-GBM MRI (611 patients). Curves and markers are as in Figure 2.

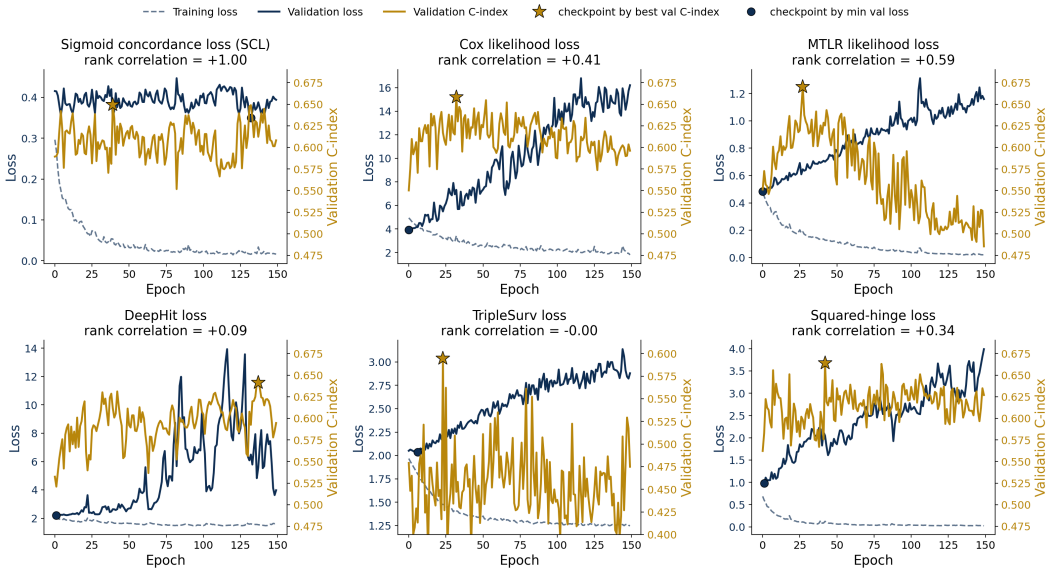


Figure A8: Loss-metric coupling on TCGA breast invasive carcinoma pathology (950 patients). Curves and markers are as in Figure 2.

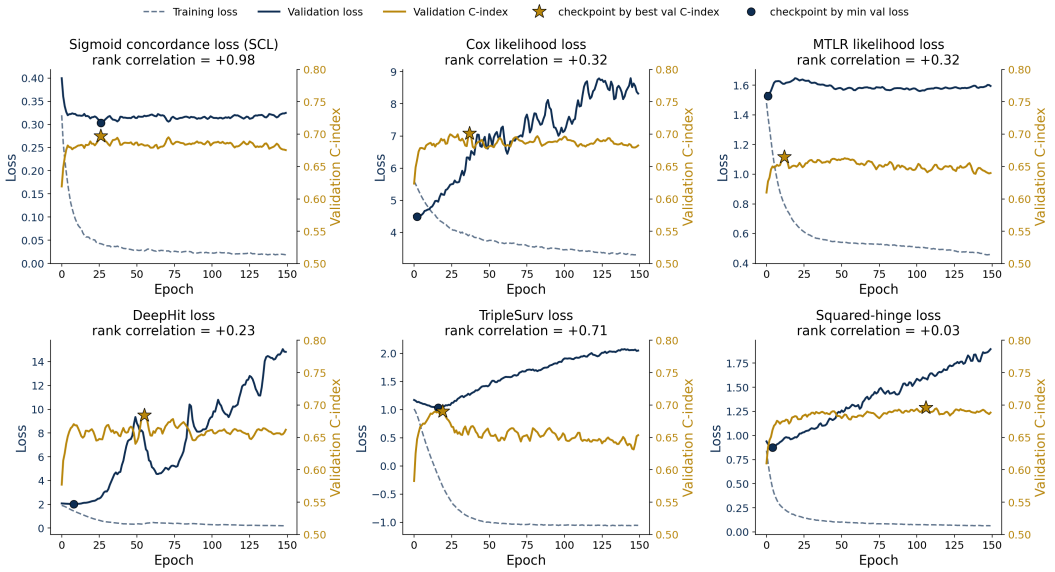


Figure A9: Loss-metric coupling on METABRIC (1904 patients). Curves and markers are as in Figure 2.

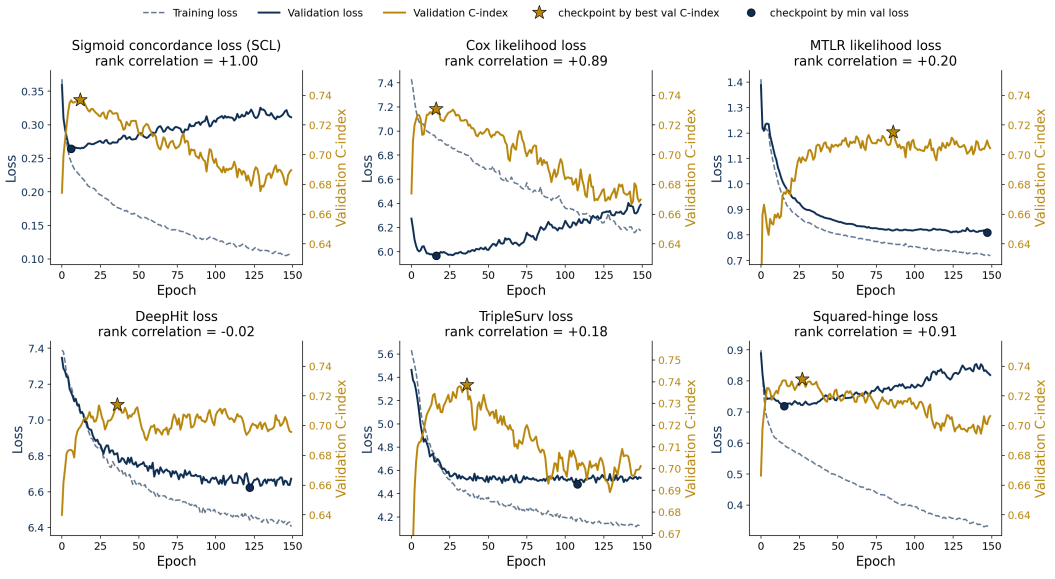


Figure A10: Loss-metric coupling on SEER (4024 patients). Curves and markers are as in Figure 2.

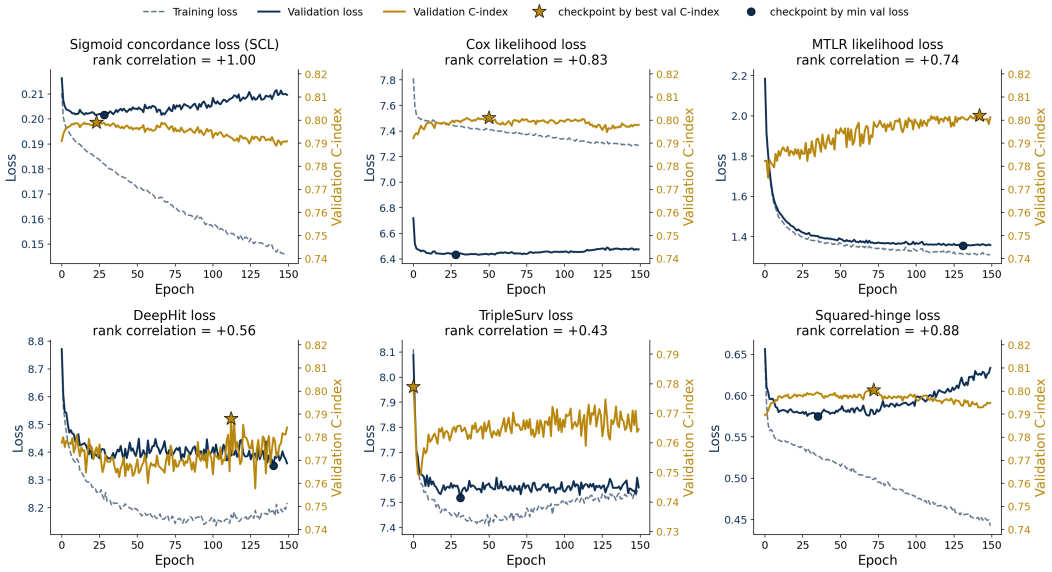


Figure A11: Loss-metric coupling on FLCHAIN (7874 patients). Curves and markers are as in Figure 2.

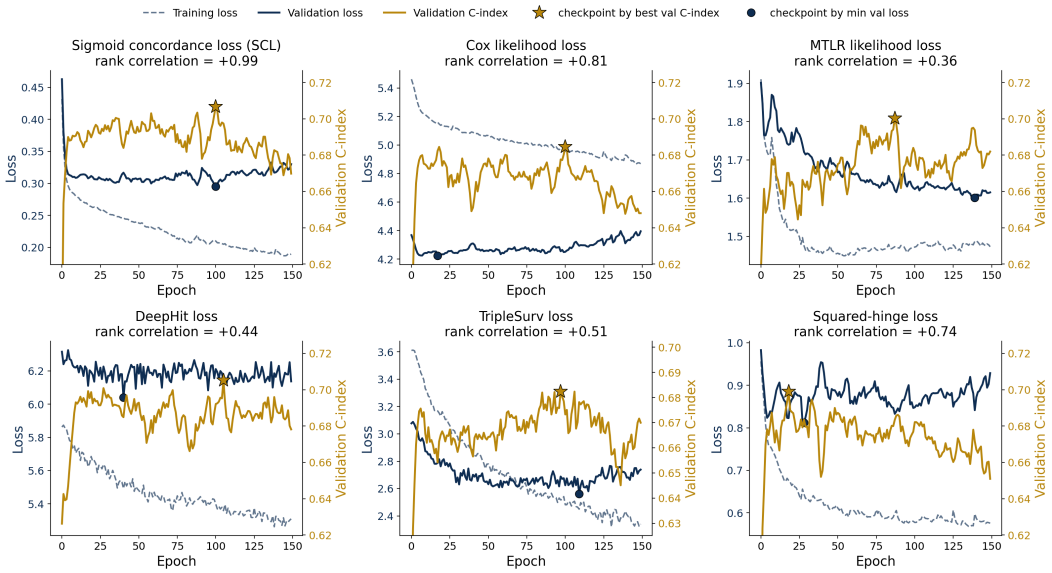


Figure A12: Loss-metric coupling on GBSG (686 patients). Curves and markers are as in Figure 2.

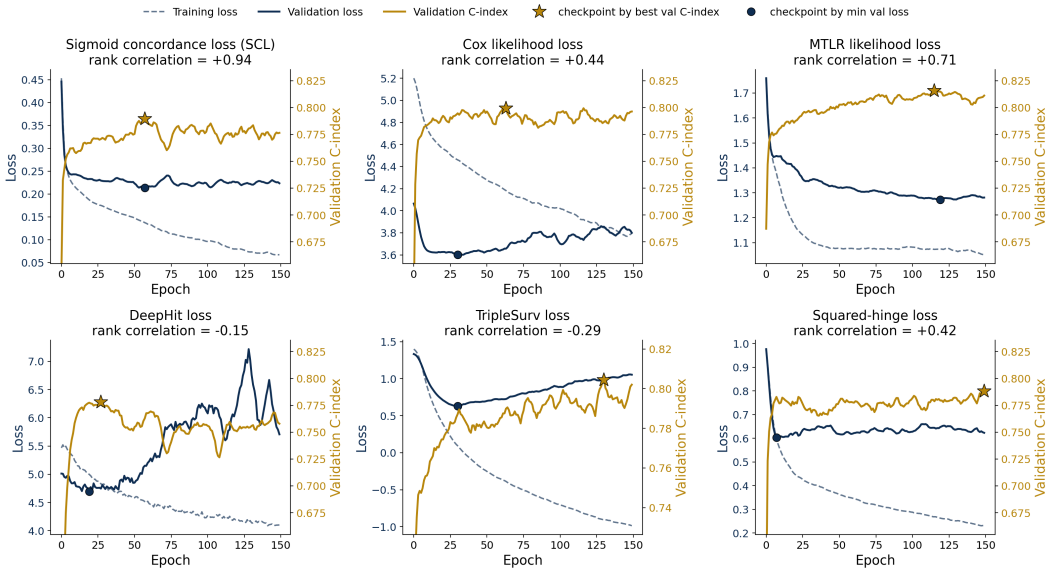


Figure A13: Loss-metric coupling on WHAS500 (500 patients). Curves and markers are as in Figure 2.

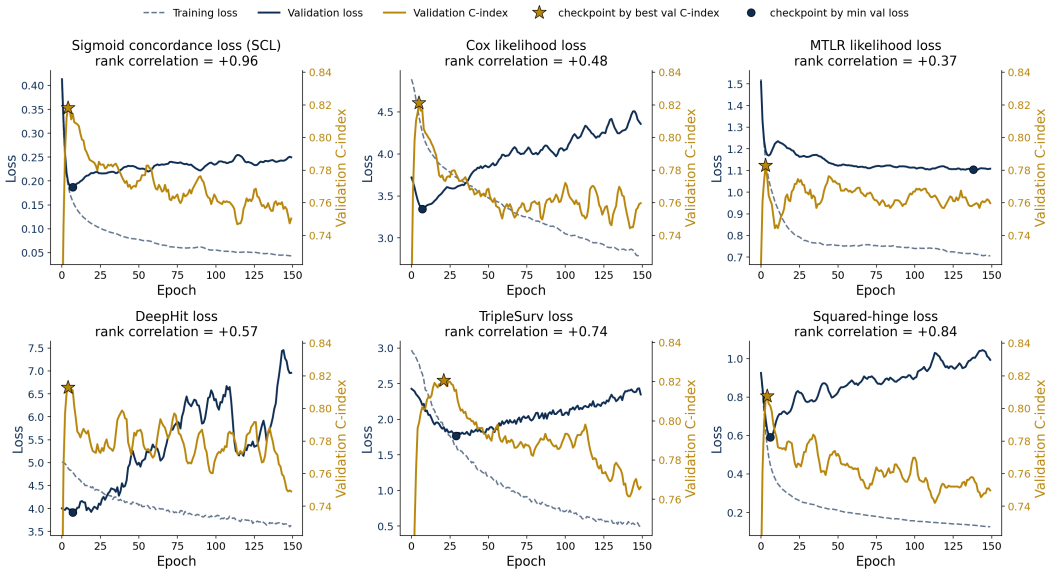


Figure A14: Loss-metric coupling on PBC (418 patients). Curves and markers are as in Figure 2.