Representation, regularization, and functional complexity: a controlled experimental disentanglement on survival data

Abstract

At identical input information, two layouts of the same tensor do not produce the same learning. We started from the hypothesis that a continuous representation of time would carry an advantage of its own; the bench submits that hypothesis to falsification. We argue, more cautiously, that the layout participates in the inductive bias by structuring the relations, neighbourhoods, and parameter sharing available to the model, effective invariance resulting from the representation-architecture-constraint triple rather than from the layout alone. We test this on survival data reconstructed from published curves, within a single family of penalized discrete-time survival models, through two experiments, shape fidelity and the estimation of a structural effect under simulated confounding. The design is frozen and hashed before the confirmatory run. The raw result suggests an advantage for the shared basis in fidelity; a factorial design with matched effective degrees of freedom shows that this advantage is not attributable to continuity, penalized discrete representations equalling or slightly exceeding the continuous spline, which ties the effect to regularization. Under simulated confounding, once uncertainty is quantified at full retention, no robust difference appears, the probability that the continuous beats the flat remaining close to chance and no cell surviving multiplicity correction. The contribution is therefore a methodological disentanglement: in the configurations tested, the apparent advantage of a continuous representation is explained by regularization and effective degrees of freedom, with no detectable advantage proper to continuity, and no robust estimation difference is detected under confounding. The experimental unit remains the complete representational pipeline, not the isolated layout.

1. Research question

The question is easy to state and hard to close. At identical input information, does the layout of the tensor change what a model learns? The superficial version opposes time carried as an axis to time scattered across columns, order against absence of order. The deep version, which is ours, asks whether the performance differences attributed to layout arise from the representational structure itself, or from the functional class, the parameter sharing, and the regularization that this structure induces within a given pipeline. It is the latter question that the bench tests.

Two experiments organize the work. Fidelity, does the layout change the reconstruction of the survival shape. Causal bias, does the layout change the inference of a treatment effect under confounding. A third avenue, data augmentation respecting the declared symmetry, is presented as a perspective and not executed here; we remove it from the core of the article so as not to announce a result we do not produce.

2. Tensorization is an invariance hypothesis

A representation is not a neutral container into which one pours already constituted data. It is a hypothesis about the structure of the problem. Before a model learns anything, the chosen layout has already declared what, in the phenomenon, is to remain invariant. The format does not transport information, it states its geometry.

Let us read each layout for what it makes available, without crediting it with more than it carries. The independent-column format imposes neither metric nor order between dimensions; it makes permutation invariance natural, but does not impose it, since that invariance exists only if the model parameters are jointly permuted as well, which a regression with a distinct coefficient per column does not do. The axis layout makes available an order, a neighbourhood, and parameter sharing along time; it does not thereby entail invariance to any monotone reparametrization of time, since a spline or a polynomial basis has its derivatives and smoothing penalties modified by such a transformation. The hierarchical layout makes a nesting into segments with reset at boundaries natural. Three layouts, three space structures made available, not three imposed invariances.

Three levels that the word symmetry tends to conflate must then be separated. The first is the structure of the input space, order, metric, neighbourhood, hierarchy, which the layout fixes. The second is the functional hypothesis, smoothness, locality, additivity, proportionality, which the representation-model pair makes cheap or not. Only the third is symmetry proper, invariance or equivariance under an explicitly defined group. Write the model as a composition of a representation and a head, $f = h \circ \varphi$, where φ maps the input into the representation space and h learns on top of it. The effective symmetry of f depends jointly on φ and h; the layout, which fixes φ, does not determine it alone. The defensible formulation is therefore not that choosing a layout amounts to choosing a symmetry group, but that the layout makes certain classes of symmetries, neighbourhoods, and parameter sharing more natural than others, invariance resulting from the triple representation, architecture, constraint. This is less spectacular and considerably more solid. The condensed image, tensorization as the moment when ontology becomes geometry, retains its value under this reading condition: the layout declares the available geometry, it does not by itself close what the model will learn.

This reading reformulates the empirical question and gives it its name. What is at stake is not axis against column, a surface opposition that hides the substance. The empirical stake is the agreement between the structure made accessible by the representation, the constraint effectively imposed by the model and its regularization, and the regularity of the phenomenon to be learned. When this triple agrees with the real structure of the effect, the model inherits a useful constraint and learns better; when it imposes a false constraint, it learns less well, at given capacity and regularization procedure. The layout is only one of the three pieces of this wager, and the wager can be lost. We shall see, moreover, that in our bench the bulk of the gap passes through regularization and not through layout alone.

3. State of the art

This work stands at the crossing of four lineages, none of which, alone, names its object.

The first is that of inductive bias and invariance. That learning without bias does not generalize has long been established [1], and representation theory has made it an explicit object [2]. Its most accomplished formulation is geometric, an architecture encodes an invariance hypothesis under a group of transformations, from which parameter sharing follows [3-5]. This vocabulary was forged for images and graphs; we transport it to the layout of covariates and time within a survival tensor.

The second is that of deep survival models, from DeepSurv [6] to DeepHit and Dynamic-DeepHit [7,8], to discrete-time neural survival [9] and the piecewise exponential framework that grounds the family used here [10]. All of them vary the architecture. None varies the layout at fixed architecture.

The third is that of tabular deep learning, TabTransformer, FT-Transformer, SAINT [11-13], with the reminder that gradient-boosted trees still often dominate [14]. These works ask how to encode a variable, not which invariance its layout declares, and they ignore the time-to-event structure.

The fourth is that of population-adjusted indirect comparison, on the applied side. The reconstruction of individual data from published curves [15] feeds matched and simulated comparison, with the cardinal distinction between anchored and unanchored comparison [16-18]. This lineage consumes reconstructed data; our generator is a producer of it, and the layout governs the fidelity of the synthetic arm produced.

A fifth lineage, which we must acknowledge because our fidelity results fall directly under it, is that of functional analysis and temporal regularization: basis expansions and tensor-product smoothers [19], varying-coefficient models [20], regularizations imposing sparsity of differences such as the fused lasso and trend filtering [21,22]. The trade-off our bench brings out, separate parameters per interval against a smooth function shared over time, is an old and well characterized problem there. Our contribution is not to discover it but to reread it in terms of layout; honesty requires naming this kinship rather than presenting as an ontological rupture what is a geometric reformulation of a known functional regularization.

The blind spot is shared. What goes into the tensor has been studied, and which architecture is run on top of it. The shape of the tensor itself has not been studied as an inductive bias, nor has this choice been systematically related to the functional regularization it induces. It is this link that we name, with the caution our results impose.

4. Data

The perimeter is the causal bench, two second-line trials in large B-cell lymphoma, TRANSFORM and ZUMA-7. No individual data are used. We reconstruct time-event-censoring datasets from the published Kaplan-Meier curves alone and their numbers-at-risk tables, by

the method of Guyot and colleagues [15]. The right term is *compatible*, not *identical*: the algorithm does not recover the real trajectories, it regenerates a distribution that reproduces the curve.

Validation re-estimates the hazard ratio on the reconstructed dataset and checks that it falls back on the published value to within three hundredths.

Source	Endpoint	Published HR	Reconstructed HR	Difference
TRANSFORM	event-free survival	0.375	0.359	−0.016
TRANSFORM	overall survival	0.757	0.751	−0.006
ZUMA-7	event-free survival	0.398	0.398	0.000
ZUMA-7	overall survival	0.730	0.729	−0.001

The time-event structure is therefore derived from published trials, with no confidential data (figure 3). Exclusive reliance on public sources makes the bench replicable and auditable by a third party; we credit it with no virtue of priority, which would be a matter of disclosure date and not of reproducibility. The backbone is overall survival, the only quantity homogeneous and recoverable from a registry; composite EFS remains within-source. For the causal experiment, a synthetic confounder with posited ground truth is added, whose law is written in §5.

5. Methods

5.1 The four layouts

We describe four layouts of the same information, at identical input information (figure 1). The *wide format* discretizes a variable into independent columns, with no exploited order. The *grid* preserves order through a normalized index. The *continuous axis* keeps the variable as a continuous dimension. The *hierarchical* form structures it by stratum. Each layout receives a frozen and declared expressivity budget, matched across layouts, with equal effective degrees of freedom, failing which one would be measuring capacity and not structure. Let us state at once, since review rightly demands it, that the results experiments empirically compare only two of these layouts, independent columns and the continuous axis; the ablation in §6 adds a variation in degrees of freedom that approaches the ordered grid, but the hierarchical form and a fully parametrized grid are not yet evaluated. This intermediate comparator is nonetheless crucial for separating order from continuity and smoothing, and its absence is a declared limitation: as it stands, the experiment opposes several properties at once rather than a single one.

5.2 Model family

A single family is used, a discrete-time survival model fitted by penalized log-linear regression, and a single learning procedure. This choice aims to reduce inductive biases competing with the one under study, but the opposite excess must be refused: this family is not opinion-free. It imposes an additive form, a link function, a penalization, an interval

approximation, and a particular way of sharing or not sharing parameters. The experimental unit is therefore not the layout alone but the complete representational pipeline, layout, parametrization, penalization, head. For robustness, a minimal single-hidden-layer perceptron, at matched capacity, verifies that the effect is not a pure artefact of linearity. We do not vary the architecture, so that the observed effect mixes the layout with its interaction with the model and the regularization; this reservation is central, not marginal. On budget matching, we equalize the number of raw parameters and the penalty strength across layouts, the continuous basis receiving a few polynomial degrees and the wide format an equivalent number of columns. This matching is parametric, not functional: two representations of equal dimension have neither the same functional class nor the same smoothing bias, and this imperfect equality is a declared limitation, not a guarantee. It is in fact the very core of what the bench studies, and it would be naive to claim to have neutralized it by a simple parameter count.

5.3 Design frozen before the confirmatory run

The confounding of the causal experiment is fabricated by us, with representations designed by us. The objection is the right one, and the answer is not to deny it but to remove the choice from the author's hands. Three devices. Expressivity budgets are frozen and hashed before any draw, so that no representation receives after the fact the term that saves it. The shape of the confounder is not chosen but sampled along a regularity axis, from rough to smooth, including shapes that no layout can express, so that nobody wins by construction. An oracle control verifies configuration by configuration that the truth is recoverable, and configurations where it is not are discarded by a rule declared in advance. We speak of a design frozen before the confirmatory run, not of preregistration: a file hashed locally before execution is not equivalent to a time-stamped third-party deposit, and only such a deposit, with version history and explicit separation of exploratory and confirmatory analyses, would make the guarantee fully enforceable.

5.4 The generating process of the confounder

The distinction governing everything is this: the estimator sees only a representation of the confounder, the oracle sees the confounder. The bias measures the gap between the two (figure 2).

For each subject i, a latent confounder $U_i \sim \mathcal{N}(0, \Sigma_U)$, $U_i \in \mathbb{R}^d$. Treatment depends on U_i, which creates confounding:

$$\Pr(A_i = 1 \mid U_i) = \sigma(\alpha_0 + \gamma\, w_A^\top U_i), \qquad \sigma(x) = (1 + e^{-x})^{-1},$$

the scalar γ setting the confounding strength, swept over a grid. Survival is posited in discrete time, in the language of the family used:

$$\lambda_i(t) = \Pr(T_i = t \mid T_i \geq t, A_i, U_i) = \sigma\big(\beta_0(t) + \tau\, A_i + g_s(U_i, t)\big),$$

where $\beta_0(t)$ is the baseline log-odds per interval, calibrated on the reconstructed shape, τ the treatment coefficient of the discrete-time model, and g_s the confounder effect. Let us specify the estimand, since the hazard ratio invites confusion: τ is here a structural conditional effect,

the coefficient of the generating mechanism, not a marginal effect. Since the hazard ratio is not collapsible, the comparison $|\hat{\tau} - \tau|$ is licit only because estimand and estimator live in the same conditional space. A marginal estimand, a difference in restricted mean survival or a fixed-horizon risk obtained by the g-formula, would be more robust and will be the object of an extension; we confine ourselves here to the structural coefficient, consistent with the family used.

The effect g_s is not fixed but drawn from a family indexed by regularity s:

$$g_s(u,t) = \sum_{k=1}^{K} a_k \cos(\omega_k^\top u + \nu_k t + b_k), \qquad a_k \propto (\| \omega_k \|^2 + \nu_k^2)^{-s/2}.$$

A large s gives a smooth surface, a small s a rough one. The three regimes are instances of it: multiple confounding ($d > 1$, Σ_U non-diagonal), temporal confounding ($\nu_k \neq 0$), non-linearity (high harmonics in u, of which the quadratic is the simple case). These three regimes cover three frequent sources of misspecification, dependence between components, temporal variation, and non-linearity; they do not constitute an exhaustive taxonomy of structural departures, which also include discontinuities, thresholds, high-order interactions, locally supported effects, or regime changes. Likewise, the regularity sweep over s traverses a continuum of smoothness, but it does not subsume differences in functional topology: a function with a break is not a more or less smooth spline in the sense relevant to every estimator. We therefore test a continuum of smoothness over three families of misspecification, not the set of all possible departures.

Censoring is non-informative, $C_i \perp T_i \mid A_i, U_i$, $Y_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}(T_i \leq C_i)$, which keeps τ identifiable. The estimand is τ, known, and the reported quantity is the bias $\hat{\tau} - \tau$.

The oracle adjusts on the true confounder within its class, and a configuration is retained only if

$$|\hat{\tau}_{\text{oracle}} - \tau| < \epsilon.$$

The two generators discarded at construction are instances of this, the one whose effect was not identifiable and the one whose censoring became informative, rejected by the rule, not by judgement. Let us specify what recoverable covers, since the term mixes several distinct things: theoretical identifiability, finite-sample estimability, correct specification of the oracle class, numerical convergence. Since the oracle fits within the exact class of the generator, its failure signals non-identifiability or informative censoring rather than misspecification; we therefore speak more exactly of failure of the oracle criterion, and §5.6 explicitly separates the structural part from the finite-sampling part.

Each layout gives the estimator only a representation $R(U_i)$ under a frozen budget, and the estimator fits $\hat{\tau}_R$ within the class $\mathcal{H}_R$ that the layout permits. Residual bias is then a function of agreement:

$$|\hat{\tau}_R - \tau| \text{ small} \Leftrightarrow g_s \in \overline{\mathcal{H}_R} \text{ (at the resolution of the budget).}$$

This is the a priori prediction of the bench, not an established fact: the layout should reduce causal bias only when the functional class it permits contains the structure of the confounder. The results in §6 state to what extent this prediction holds, and the answer, as will be seen, is more cautious than the prediction.

5.5 Metrics

For fidelity, the integrated squared error to the curve, complemented by the restricted mean survival error and integrated calibration, all reported and mutually consistent. For estimation under simulated confounding, we report the signed bias, the absolute error, and above all the paired difference in error between representations for the conditional coefficient τ. Concordance and the Brier score are set aside, being irrelevant for a structural coefficient on a two-arm reconstruction.

5.6 Recoverability filter, structural grid, and event regime

The oracle filter of the initial design conflated two distinct things, the recoverability of the truth, which is a property of the generator, and the sampling noise of the oracle estimator at finite sample size. A diagnostic shows it: under the initial regime the oracle is unbiased on average, of the order of zero, but its sampling standard deviation is about 0.10, far above the threshold ϵ of 0.05. The per-seed filter was therefore rejecting two thirds of the configurations for noise and not for non-recoverability, starving the causal cells. We correct this with two declared and re-hashed moves, motivated by this diagnostic.

First, the structural grid is decoupled from estimation. Recoverability is judged once per generator configuration, on a large sample where the oracle standard deviation becomes small, and its threshold is set at three times this residual standard deviation, so that it discards only configurations whose oracle bias persists at large sample size, the structurally non-identifiable ones, and not those that depart only through sampling. The finite-sample experiment then runs on all seeds of the admitted configurations, without re-filtering. The grid looks only at the generator, never at which layout wins, which guarantees that it raises power without biasing the contrast; this is the line between correcting power and settling the outcome.

Second, the regime is set on more events, with the baseline hazard raised and censoring lightened, power in survival resting on the number of events and not on raw sample size. The oracle standard deviation thus goes from 0.10 to 0.07. The combination restores full retention, eighteen to twenty seeds out of twenty, and the causal estimates cease to be noisy.

For traceability, the table of deviations from the initial protocol.

Element	Initial protocol	Amendment	Justification
Oracle filter	per seed	per configuration at large N	separate noise from non-recoverability
Event regime	initial	raised hazard, lightened censoring	lower the estimator variance

Element	Initial protocol	Amendment	Justification
Recoverability threshold	fixed (0.05)	3σ residual	control false rejection

The amendment was diagnosed after a first run judged too unstable; it is therefore informed by that run, while remaining blind to the winning layout. This status of confirmatory amendment, and not of initial protocol, is declared as such. The event-enriched regime modifies the simulated world, so we verified that the verdict does not depend on it. Varying the baseline hazard under the oracle grid and the same seeds as the main analysis, from the rich regime, about fifteen per cent of events per interval, to the poor regime, about eight per cent, the probability that the continuous beats the flat remains between 0.46 and 0.61 and the paired difference stays within six thousandths, all intervals crossing zero. The null result is therefore not an artefact of the amended event regime. A full grid crossing events, censoring, and confounding strength would remain desirable for a demanding review.

5.7 Detail of the factorial design

The factorial design is described here so that it can be reconstructed without the code. The continuous representation is a penalized spline, a locally supported cubic B-spline basis together with a penalty on the second differences of its coefficients, a P-spline; a global degree-five polynomial, used initially and mechanically disadvantaged on breaks, was discarded in favour of this true spline. The discrete representation is an indicator per interval together with a penalty on the second differences of neighbouring log-odds, without a continuous basis; the grid is a linear ordered index. The effective degrees of freedom are the trace of the smoothing operator, $tr[(X^\top WX + \lambda P)^{-1} X^\top WX]$, evaluated at the penalized solution obtained by iteratively reweighted least squares. Matching fixes a target degrees-of-freedom value and adjusts λ by bisection search to reach it. Estimation is on a training set, evaluation on a disjoint test set in equal halves, over twenty-five seeds; the smooth truth is a squared-sine hazard, the break truth a plateau followed by a jump. The discrete against spline contrast is paired by seed and its ninety-five per cent interval is a paired bootstrap with two thousand replications, consistent with the arm under confounding. Matching by the trace equalizes a global effective complexity, not the full geometry of the operators, two operators of equal trace being able to differ in spectrum or boundary behaviour; we therefore speak of comparable global effective complexity, not of strictly equal flexibility. The design is thus a controlled ablation at matched effective complexity, not a fully crossed factorial design, and we do not over-qualify it.

6. Results

The overall logic of the bench, from the raw effect to the methodological principle, is summarized in figure R0, which the hurried reader can read on its own. Results are reported with their uncertainty, and deliberately read down. The confirmatory run operates at full retention, ninety to ninety-eight per cent of admitted configurations. This high retention reduces the risk of substantial selection, without excluding it, and we therefore report its

distribution: oracle rejection rates run from zero to five per cent depending on regularity and regime, with one exception, temporal confounding, where they reach fifteen per cent. That is the only cell where residual selection warrants caution, and we flag it there; elsewhere the low rejection rate limits, without excluding, the risk of orientation.

Fidelity (figures R1, R4, and R5). Representing time by a shared basis rather than by independent columns modestly reduces the reconstruction error, the gap increasing with the number of intervals, where independent columns, one per interval, overfit, something objectified by a gap between training and test deviance. A controlled ablation at matched global effective complexity clarifies its nature. At the same effective complexity measured by the trace of the smoothing operator, we compare a discrete representation with a second-difference penalty, which shares information between neighbouring intervals without a continuous basis, and a true locally supported penalized spline. The choice of comparator matters, and we correct it: a global polynomial, used initially, is mechanically disadvantaged on a break and exaggerated the gap. With a penalized spline, ISE differences drop by an order of magnitude, around one millionth, and the paired bootstrap contrast shows no superiority for the continuous representation: on smooth truth at six degrees of freedom the difference is nil, elsewhere it favours the discrete by a minute margin. Restricted mean survival, integrated calibration, and test deviance are identical between the two. The exact formulation is therefore more cautious than "equivalent": at comparable global effective complexity, no superiority proper to continuity is observed, and the residual gaps, small and often oriented towards the discrete, are without practical relevance. The fidelity effect is a matter of the bias-variance trade-off and of regularization, not of a geometry of continuity. This is the clearest result of the bench, and one must guard against replacing a pro-continuous doctrine with a pro-discrete mirror image: at matched complexity, the two behave in the same way.

The table gives the four prespecified metrics, mutually consistent, by representation and truth, the continuous representation being the penalized spline.

Representation (truth)	eff. df	ISE	RMST	calibration	test deviance
indep. columns (smooth)	48.0	$4.2\cdot10^{-4}$	$9.3\cdot10^{-3}$	$1.5\cdot10^{-2}$	0.259
grid df=2 (smooth)	2.0	$1.3\cdot10^{-3}$	$1.3\cdot10^{-2}$	$2.8\cdot10^{-2}$	0.238
fused discrete df=6 (smooth)	6.0	$3.5\cdot10^{-4}$	$9.6\cdot10^{-3}$	$1.4\cdot10^{-2}$	0.233
penalized spline df=6 (smooth)	6.0	$3.5\cdot10^{-4}$	$9.8\cdot10^{-3}$	$1.4\cdot10^{-2}$	0.233
indep. columns (break)	47.0	$5.8\cdot10^{-4}$	$1.1\cdot10^{-2}$	$1.9\cdot10^{-2}$	0.230
grid df=2 (break)	2.0	$1.6\cdot10^{-3}$	$1.4\cdot10^{-2}$	$2.9\cdot10^{-2}$	0.205
fused discrete df=6 (break)	6.0	$5.5\cdot10^{-4}$	$1.2\cdot10^{-2}$	$1.8\cdot10^{-2}$	0.205
penalized spline df=6 (break)	6.0	$5.6\cdot10^{-4}$	$1.2\cdot10^{-2}$	$1.8\cdot10^{-2}$	0.205

The table also carries the substantive lesson: the ISE gaps between discrete and spline, statistically resolved but of the order of one millionth, are invisible on RMST, calibration, and deviance. Statistical significance does not amount to practical relevance, and no prespecified margin of clinical relevance would be crossed.

Estimation error of the structural effect under simulated confounding (figures R2 and R3). We rename this arm, since it does not estimate a clinical causal effect but the recovery error of a conditional coefficient in a simulated world. The reported unit is the paired difference between layouts on the same seeds, resampled by a seed-paired bootstrap, the resampling unit being the seed and not the subject, two thousand replications, percentile intervals. The verdict is null, and it is now quantified rather than asserted. None of the eight cells survives a Holm correction, the smallest p value being 0.038 against a corrected threshold of 0.006; the paired sign tests are all non-significant; the probability that the continuous beats the flat remains between 0.43 and 0.59. The $s = 2$ cell, marginal before correction, resists neither Holm nor the sign test: it is a multiplicity false positive. We also characterize what this null covers, failing which it would be over-interpreted. The minimum detectable difference at eighty per cent power is about 0.022 in log-odds, and the data are compatible with the absence of any gap greater than 0.033, this value being the widest ninety-five per cent confidence interval bound observed across cells, expressed in log-odds per interval, whose practical interpretation is a negligible shift in conditional hazard, that is, on a baseline hazard of five to twenty per cent, an absolute gap of the order of two to five tenths of a percentage point. It is therefore a null within a margin, not a proven equivalence: an effect finer than the detectable difference cannot be excluded, and the reported uncertainty is that of the Monte Carlo conditional on the bench, not that of transport to other trials. This null is stable across the event regime, the sensitivity analysis in §5.6 confirming it from rich to poor hazard, but its point level remains sensitive to the draw of configurations, so that it reads in the sense of multiplicity correction and not as a stable point. A control under an external generator, a Weibull accelerated failure time outside the estimation family, illuminates its scope: when representations are not matched in complexity, the rich representation removes residual confounding that the coarse representation leaves, and the gap is then clear and systematic; but this is an effect of functional class and capacity, not of continuity, which confirms the disentanglement rather than an advantage proper to the format. An external control at matched complexity remains to be conducted. The hypothesis plan is declared and hierarchical: main contrast integrated over s, then a global representation against regularity test, then secondary contrasts by regime, and finally cell-by-cell exploratory analyses, the Holm correction applying only to this last exploratory level.

Replication on independent generators (figure R6). To answer the reproach of an endogenous generator, we apply the same protocol to three structurally independent generator families: a Cox proportional hazards model, a Weibull accelerated failure time, and the discrete logistic model. In each, a smooth confounder acts on treatment and on outcome, and we measure the recovery error of the treatment coefficient when this confounder is represented in three ways at matched effective degrees of freedom, independent columns, penalized discrete, continuous spline, the reference being adjustment on the true confounder. The result is the same from one generator to the next, which is the essential point: the ordering of representations is preserved, the continuous spline recovers the effect at least as well as the discrete, better than independent columns, on Cox as on Weibull as on the discrete logistic. The external robustness of the qualitative verdict is therefore established outside the estimation family.

Two nuances that honesty imposes. First, contrary to the result on the representation of time, where discrete and continuous were equivalent, the spline here retains a slight advantage for a smooth confounder. The explanation is instructive and consistent with the principle of the paper: matching by the trace does not equalize the approximation quality of a smooth target by a step basis. Second, this advantage narrows when the grain of the discrete representation is refined, from twenty to ninety-six bins the gap is reduced by half to two thirds without disappearing, which shows that discrete grain is a lever distinct from effective degrees of freedom. The lesson is therefore not that one representation wins universally, but that a comparison of representations must control both effective complexity and functional class, failing which the gap attributed to the format mixes the two. This is exactly the contribution the bench defends. Let us insist in order to remove an appearance of contradiction: R6 does not contradict R5. R5 concerns the representation of time, where the grain was already fine and where discrete and continuous were equal; R6 concerns the representation of a smooth confounder, where a coarse discrete grain approximates less well than a spline. These are two distinct experimental questions, and in both the gap, when it exists, is explained by effective complexity and grain, not by a virtue proper to continuity.

Overall reading. The bench establishes a modest fidelity effect, regularizing in nature, and detects no robust estimation difference under simulated confounding. The fidelity result, independent of synthetic confounding, carries the burden of proof, but the factorial design reduces its reading to a phenomenon of effective degrees of freedom rather than of continuity. The null result reduces the fear of a construction mechanically producing the expected superiority, without constituting proof of the bench's neutrality, since a null is also compatible with two equally well specified methods, a noisy metric, insufficient power, or a poorly discriminating DGP. An increase in sample size, and above all in independent clinical contexts, could reveal an effect currently below the detectable difference, but nothing in the present data allows this to be asserted.

Augmentation. Reserved avenue, not executed, presented as a perspective in §7.2.

7. Discussion

7.1 Against classical statistical methods

Classical methods are not the outside of the thesis, but an excessive extension of the vocabulary must be avoided. A Cox proportional hazards model [23] is not a tensorization; it imposes a structural constraint, proportionality of hazards and an additive form over covariates. Kaplan-Meier [24] rests on independent censoring and homogeneity of the survival mechanism within strata. Fine-Gray specifies a model for the subdistribution hazard. These are model assumptions, and sometimes invariances, not necessarily group symmetries. The link with our framework is more modest thus, and more accurate: each fixes in advance a structure and a class of compatible transformations, where the bench treats this structure as a variable. The danger, which we name in order to avoid it, is that of a concept so extensive that it would encompass every statistical assumption and would then cease to discriminate anything. The operational message is methodological and not a result of this study, which the bench did not test on a strictly proportional generator: when a classical

method already fixes the right structure, a single clock and a proportional effect for Cox, the choice of layout should carry little weight; it is when the structure of the effect departs from the posited assumption that the layout becomes a potential lever again, subject to the reservation our results add, no advantage proper having been detected beyond regularization.

7.2 Perspectives

One extension is reserved, with no result at this stage: data augmentation respecting the declared structure could help more than blind augmentation, but only where this structure coincides with that of the effect; we do not execute it here.

7.3 Application

The possible applied consequence concerns synthetic arms and indirect comparisons: the layout could influence the fidelity of the twin to the joint distribution of covariates and the stability of reweighting. It is not tested here, no measure of balance, overlap, effective sample size, or MAIC bias being produced, and a dedicated experiment would be necessary before any assertion.

7.4 Simulation bias and neutral comparison

A poorly constrained comparative simulation makes it possible to claim the superiority of almost any method [25], and this is the fundamental objection addressed to any bench whose author fabricates the world. We answer it by design rather than by protestation. The plan follows the ADEMP standard for simulation studies [26] and aims to be a neutral comparison in the sense of Boulesteix and colleagues [27]. The safeguards are the protocol frozen and hashed before any draw, the regularity grid prespecified so that no layout expresses it by construction, and the oracle recoverability grid decoupled from estimation and blind to the winning layout. We do not claim that these safeguards prove neutrality: a generator may favour a method in a regime-dependent way if the regimes themselves are chosen to coincide with the theory, and regime dependence is therefore not internal proof of neutrality. Neutrality is established only by the frozen protocol, neutral comparators, ablations, a plasmode anchoring on real covariates [28], and, in time, external replication. That the causal result is null here, no layout being favoured, is moreover consistent with a bench that does not lean.

8. Limitations

The domain of validity is explicit and it bounds the conclusions, more narrowly than earlier versions suggested. A single model family is used, and the unit actually tested is the complete representational pipeline, not the isolated layout; the observed effect mixes layout, parametrization, and penalization, and separating them would require several control architectures, Cox, flat discrete model, penalized spline, gradient boosting, minimal perceptron. Budget matching is parametric and not functional, two representations of equal dimension having neither the same functional class nor the same smoothing bias. The causal

generating mechanism is written in the language of the estimation family, which favours representations compatible with this parametrization; a robust validation would use several generators, Cox, accelerated failure time, non-proportional or non-monotone hazards, a simulator independent of the estimated family, or even a semi-real anchoring on real covariates. The estimand is a non-collapsible conditional coefficient, and a difference in restricted mean survival would be more robust. The first-order fidelity experiment, R1, and the Guyot reconstruction do not yet have the clean train-test separation that the R5 factorial design already benefits from, nor a sensitivity analysis to curve digitization and numbers-at-risk tables. Validating the reconstruction by proximity of the hazard ratio validates a scalar, not the complete structure, censoring times, non-proportionality, restricted mean survival. The empirical support reduces to two trials, one pathology in cellular immunotherapy. Finally the causal result is null at this sample size: we cannot exclude a fine effect, but neither can we assert it. These limitations do not weaken the theoretical thesis on representation, they severely bound what the empirical verification allows one to conclude, and this bound is declared rather than passed over in silence.

Code availability and reproducibility

The bench is reproducible end to end from public sources. Reconstruction operates through `guyot.py` on the published Kaplan-Meier curves of TRANSFORM and ZUMA-7. The design frozen before the confirmatory run, matched expressivity budgets, a DGP family parametrized in regularity, and the oracle filter, is fixed in a protocol hashed before any draw, so that neither budgets nor shapes can be adjusted after seeing the data; this local hashing does not, however, amount to a time-stamped third-party deposit. The confirmatory run applies the filter amendment described in §5.6, a recoverability grid decoupled from estimation and a regime with more events, and reports uncertainty through paired differences, probabilities, and bootstrap intervals. No individual data are used, only the published curves and their numbers-at-risk tables are inputs, which makes the bench replicable and auditable by a third party.

Figures

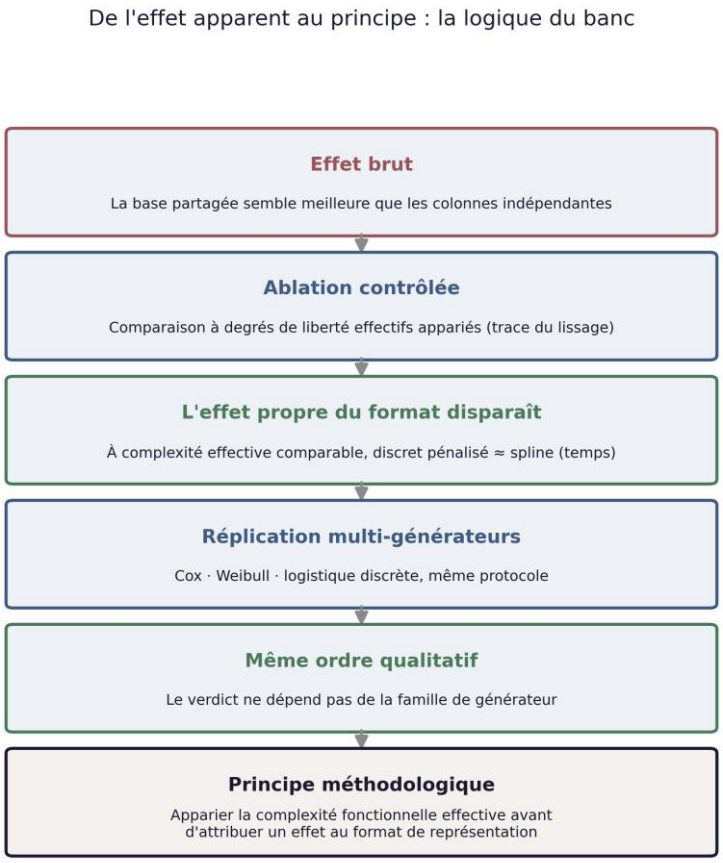

Figure R0. Summary of the logic of the bench: from the raw effect to the methodological principle, by way of ablation, complexity matching, the disappearance of the effect, and multi-generator validation.

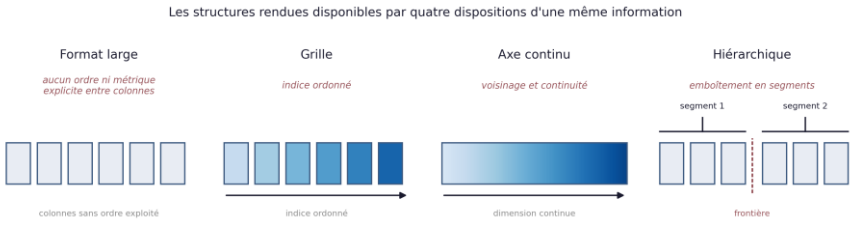

Figure 1. The structures made available by four layouts of the same information; the labels describe the structure of the input space, not an imposed invariance.

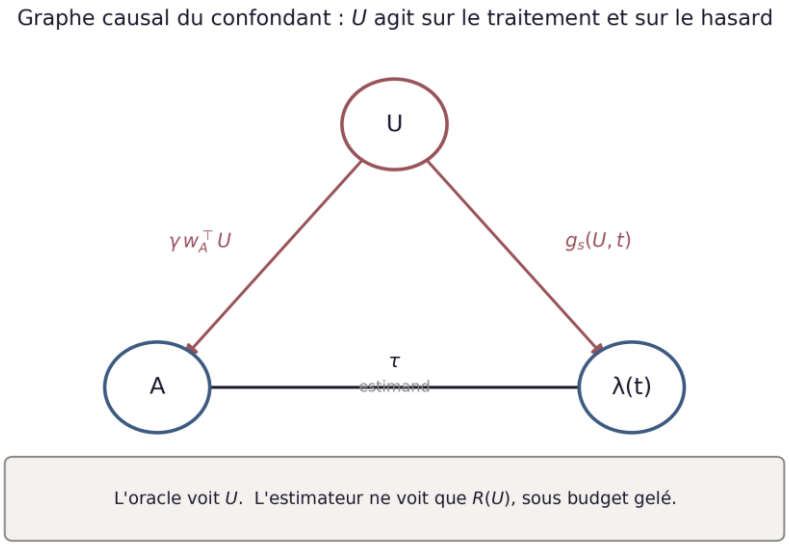

Figure 2. Causal graph of the confounder, U acting on treatment and on the hazard, the oracle seeing U and the estimator $R(U)$.

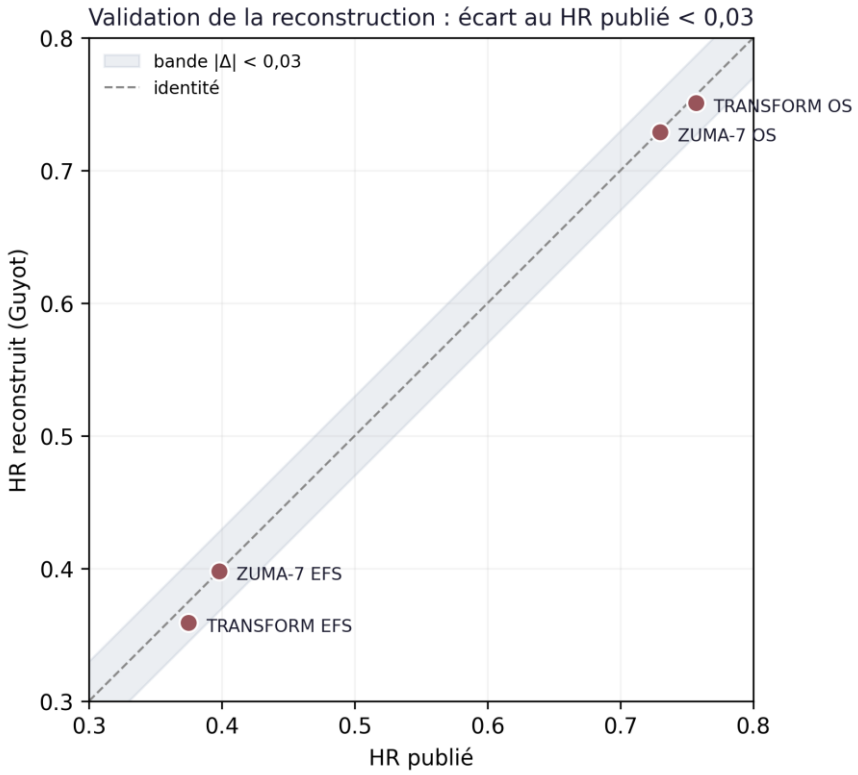

Figure 3. Validation of the HR on the reconstructed data, difference from the published hazard ratio below three hundredths for the four curves; validation of one scalar per curve, not of the complete structure.

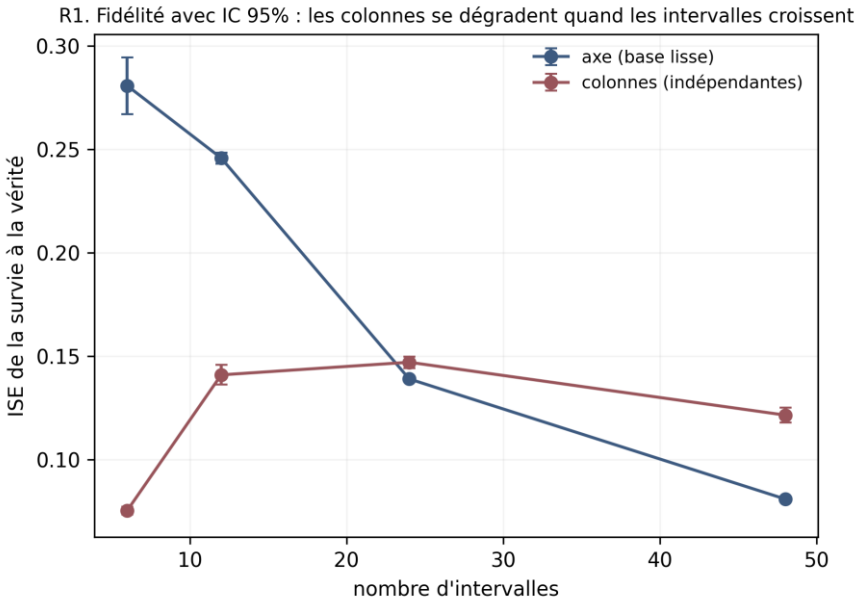

Figure R1. Fidelity: at fixed penalization, the error of independent columns increases relative to the shared basis as granularity grows, 95% confidence intervals. R1 measures the effect of granularity at given parametrization, R5 that of flexibility at fixed granularity.

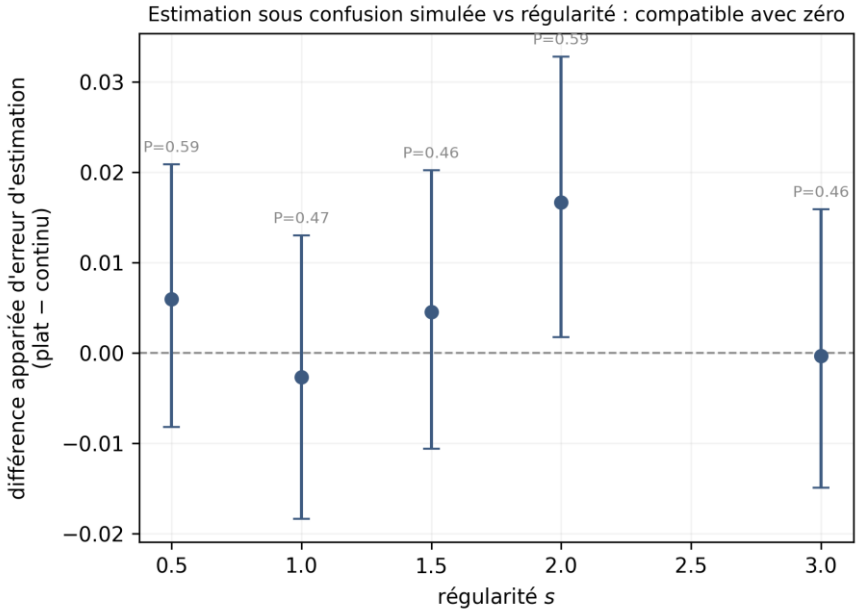

Figure R2. Paired difference in estimation error of the structural coefficient, flat minus continuous, as a function of regularity *s*, 95% bootstrap CI, compatible with zero.

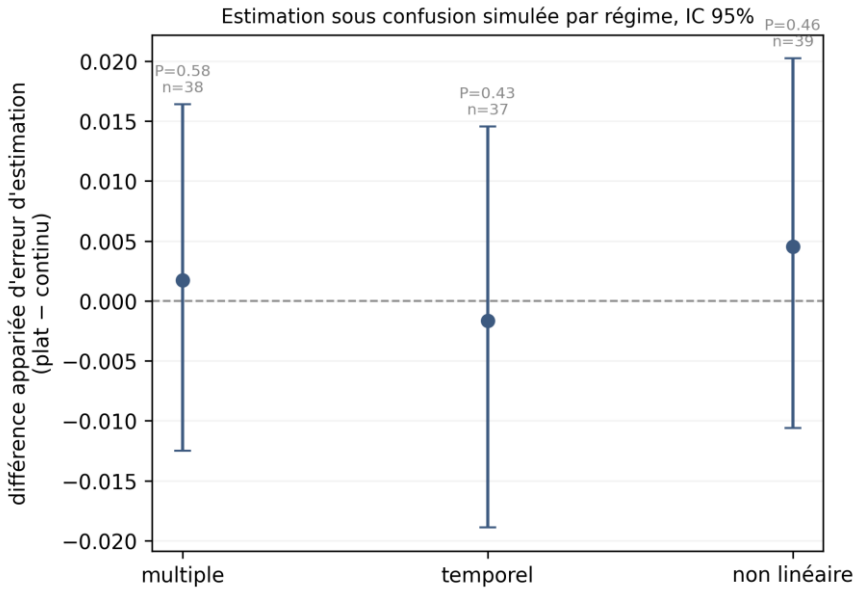

Figure R3. Paired difference in estimation error of the structural coefficient by regime, 95% CI, sample size and probability annotated.

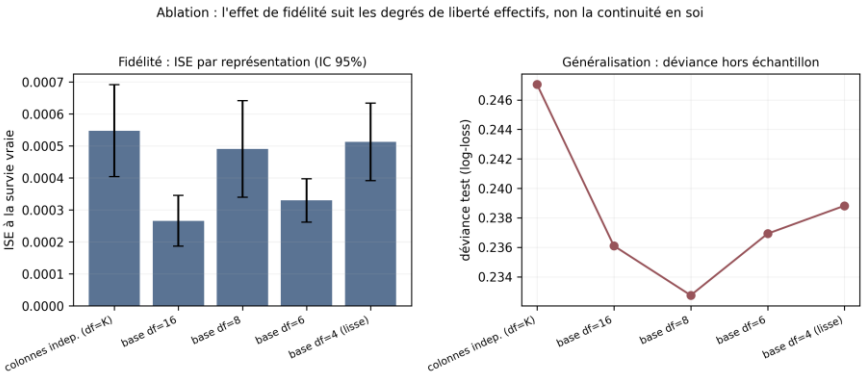

Figure R4. Fidelity ablation, ISE and out-of-sample deviance as a function of the degrees of freedom of the temporal representation, showing that the effect follows effective flexibility and not continuity per se.

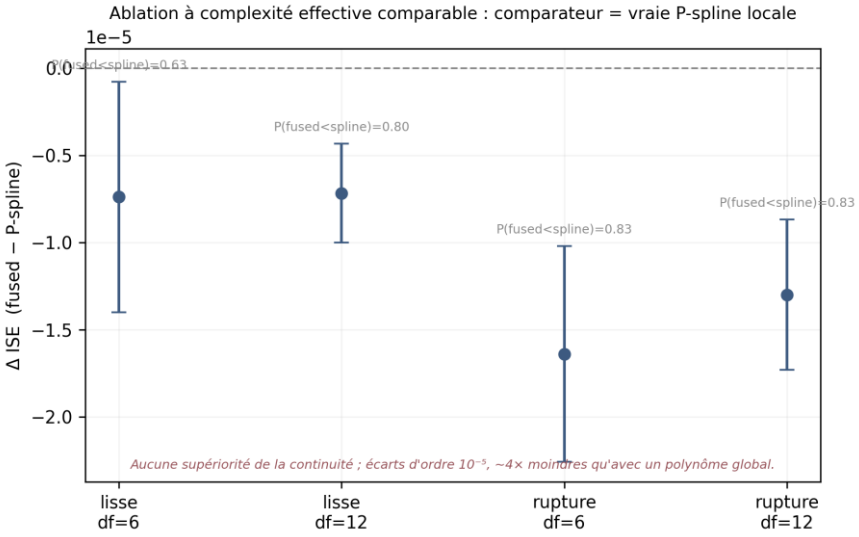

Figure R5. Controlled ablation at comparable global effective complexity, paired contrast of discrete against penalized spline on smooth and break truths: no superiority proper to the continuous representation, the residual gaps being an order of magnitude smaller than those of a global polynomial.

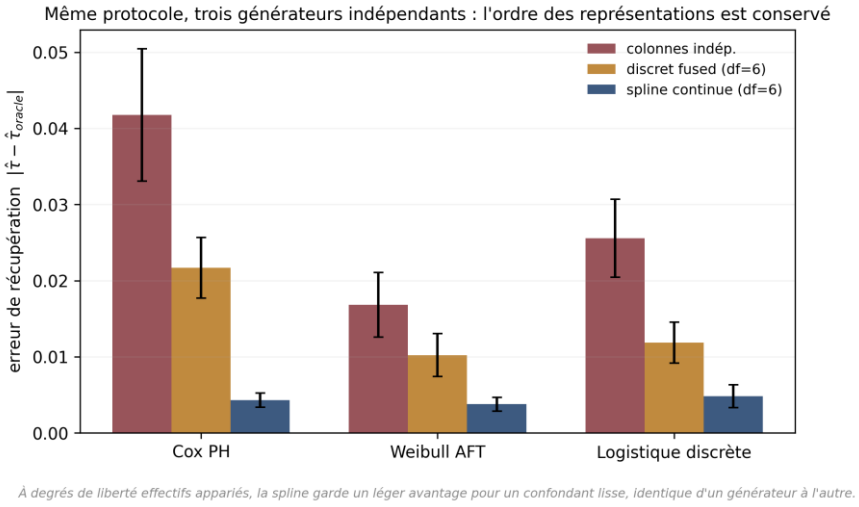

Figure R6. Same protocol on three independent generators, Cox proportional hazards, Weibull accelerated failure time, discrete logistic: recovery error of the treatment coefficient by confounder representation at matched effective degrees of freedom. The ordering of representations is preserved from one generator to the next.

9. Scope of inference

Rather than a defensive list, we explicitly bound each claim by its proof and its limit. This table stands in for a conclusion: it fixes what the bench allows one to conclude, and nothing more.

Claim	Proof in this bench	Limit
The choice of representation and penalization modifies the accessible functional class	factorial ablation, fidelity	attribution conditional on the pipeline
Parameter sharing modestly reduces reconstruction error on a fine grid	R1, R4	small effect, effective df mechanism
No effect proper to continuity detected	R5, penalized discrete equals or exceeds the continuous at matched df	smooth and break configurations only
The qualitative verdict is robust to the generator family	R6, three independent generators, same ordering	modest spline advantage for a smooth confounder, dependent on discrete grain
No estimation difference under confounding	R2, R3, Holm, sign tests	null within a margin of 0.033, not equivalence
The layout is an invariance hypothesis in the strong sense	none	not established, doctrinal formula
Superiority of the continuous axis, reduction of causal bias	none	not supported by the results, not detected within the margin
Generalization to other pathologies, DGPs, architectures	none	out of scope
Improvement of a MAIC or of a synthetic arm	none	untested transfer hypothesis

The novelty is not the layout as an invariance hypothesis, a thesis this bench does not support, but a methodological disentanglement: showing that effects attributed to layout can be confounded with functional regularization, and proposing a decomposition between representation, functional class, architecture, constraint, and effective complexity. The defensible formulation is therefore this one: at identical nominal information, the representational pipeline modifies the functional class and the regularization accessible to the estimator; in this survival bench, the apparent advantage of a continuous representation is explained by effective degrees of freedom rather than by continuity itself, and no robust estimation difference is detected under simulated confounding. From this follows a design principle, more transportable than the local result without claiming that the latter generalizes: any comparison of representations must explicitly match or characterize effective functional complexity and induced penalization, failing which the effect attributed to the format may be nothing but a disguised regularization effect. This principle exceeds survival analysis. It holds wherever representations of unequal capacity are opposed, tabular data, text, image, graphs, and it is probably the widest reach of the work: not a property of continuity, but a requirement of experimental control for any comparison of formats. We state it here without claiming to demonstrate it beyond the survival bench, but we do not undersell it either.

Two reservations of scope that honesty imposes. First, forty or a thousand seeds around four curves drawn from two trials do not make forty or a thousand clinical contexts; the reported uncertainty is simulational, not an empirical transport, and conceptual pseudo-replication remains. Second, the neutrality of the bench is not proved by the frozen protocol alone: representations, regimes, and metrics were defined according to the theory, and strong validation, generators built by an independent team, an external challenge dataset, plasmodes on real covariates, replication by a second implementation, blinded analysis of labels, is not an optional extension but a condition of external validity. Passage to a demanding methodological review presupposes this programme, plus the addition of several generators independent of the estimated family, a grid of event and censoring rates, a marginal estimand of the restricted mean survival type, and evaluation of the grid and hierarchical layouts.

References

1. Mitchell TM. The need for biases in learning generalizations. Rutgers University Technical Report CBM-TR-117; 1980.
2. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798-1828.
3. Cohen T, Welling M. Group equivariant convolutional networks. In: Proceedings of the International Conference on Machine Learning (ICML); 2016.
4. Battaglia PW, Hamrick JB, Bapst V, et al. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261; 2018.
5. Bronstein MM, Bruna J, Cohen T, Veličković P. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. arXiv:2104.13478; 2021.
6. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18:24.
7. Lee C, Zame WR, Yoon J, van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2018.
8. Lee C, Yoon J, van der Schaar M. Dynamic-DeepHit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng. 2020;67(1):122-133.
9. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. J Mach Learn Res. 2019;20(129):1-30.
10. Bender A, Rügamer D, Scheipl F, Bischl B. A general machine learning framework for survival analysis. In: Proceedings of ECML PKDD. Springer; 2021.
11. Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer: tabular data modeling using contextual embeddings. arXiv:2012.06678; 2020.
12. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data. Adv Neural Inf Process Syst. 2021;34:18932-18943.

13. Somepalli G, Goldblum M, Schwarzschild A, Bruss CB, Goldstein T. SAINT: improved neural networks for tabular data via row attention and contrastive pre-training. arXiv:2106.01342; 2021.
14. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Adv Neural Inf Process Syst; 2022.
15. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2012;12:9.
16. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons. Pharmacoeconomics. 2010;28(10):935-945.
17. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU Technical Support Document 18: methods for population-adjusted indirect comparisons in submissions to NICE. Sheffield: NICE Decision Support Unit; 2016.
18. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. Med Decis Making. 2018;38(2):200-211.
19. Wood SN. Generalized additive models: an introduction with R. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2017.
20. Hastie T, Tibshirani R. Varying-coefficient models. J R Stat Soc Series B. 1993;55(4):757-796.
21. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J R Stat Soc Series B. 2005;67(1):91-108.
22. Kim SJ, Koh K, Boyd S, Gorinevsky D. l1 trend filtering. SIAM Rev. 2009;51(2):339-360.
23. Cox DR. Regression models and life-tables. J R Stat Soc Series B. 1972;34(2):187-220.
24. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-481.
25. Pawel S, Kook L, Reeve K. Pitfalls and potentials in simulation studies: questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. Biom J. 2024;66(1):e2200091.
26. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-2102.
27. Boulesteix AL, Lauer S, Eugster MJ. A plea for neutral comparison studies in computational sciences. PLoS One. 2013;8(4):e61562.
28. Schreck N, Slynko A, Saadati M, Benner A. Statistical plasmode simulations: potentials, challenges and recommendations. Stat Med. 2024;43(9):1804-1825.