Matching complexity is not enough

What a comparison of representations establishes, and what it cannot yet establish

Executive summary

Problem. Comparing two representations without controlling for their complexity conflates structure with capacity: what is attributed to the format is often nothing but regularisation.

Method. A single estimation family, a single marginal estimand, effective complexity matched through the trace of the smoothing operator, semi-synthetic data-generating processes calibrated on published covariate margins, protocol preregistered and timestamped before execution.

Main result. Matching removes part of the differences between representations, but not all of them. At comparable complexity budget, transformations belonging to the same estimation family retain different performances in estimating the causal estimand despite matched global complexity; the mechanistic analysis suggests that the expressivity of the function class accounts for part of this.

Generalisation. Over a distribution of worlds explored post-registration, the advantage is probabilistic rather than universal: P(continuous axis better than wide format) = 0.73 ± 0.02. In more than a quarter of worlds, another transformation does better.

Mechanistic result. The spectral concentration of complexity does not provide the expected explanation. The expressivity of the function class is more informative, without being universally predictive: it accounts for two regimes out of three and reverses sign in the third.

Identifiability boundary. No representation compensates for a loss of identifiability: under non-recoverable hidden confounding, all of them fail.

Boundaries. A single estimation family. Mechanisms whose structural support is defined by the authors. No clinical validation: clinical data serve solely to calibrate marginal distributions.

Conclusion. When the objective is to attribute a performance difference to the representation rather than to capacity, complexity matching is a necessary methodological condition. It is not sufficient to explain the residual differences.

Three levels of claim

A methodological article that superimposes its levels of evidence invites trivial refutation: it suffices to attack the weakest one to appear to have refuted the whole. This manuscript separates them before setting out anything else [25,27].

Level A: established within the prespecified experimental framework. Distinct representational transformations produce different biases despite the equalisation of their global effective complexity. Section 7.

Level B: supported, not isolated. These differences are more compatible with an explanation by the expressivity of the function class than with an explanation by capacity. The run is consistent with this reading; it does not isolate it from all alternatives. Sections 8 and 9.

Level C: tested, not retained. The mechanism behind this adequacy would be the alignment between the signal and the directions to which the smoothing operator assigns its complexity. This hypothesis has been tested and is not supported in the form proposed. Section 9.

The reader in a hurry may stop at A. The methodological reader will turn to Section 11, which maps what is controlled and what is not.

Key result

df_eff(R_1) = df_eff(R_2) implies neither **F_R_1 = F_R_2**, nor **Bias(R_1) = Bias(R_2)**.

Two pipelines with the same global effective complexity may operate over different function classes, and retain different performances in estimating the causal estimand.

1. Research question

The naive question opposes time carried as an axis to time split into columns. It is ill-posed. The defensible version is narrower:

"From the same source covariates, how do different representational transformations, paired with a single estimation family, modify the function class accessible at comparable effective complexity?"

That is to say: arranging the same data differently changes what a model can understand, even when it is given exactly the same room for manoeuvre.

The change is not terminological. The earlier formulation, "at identical input information", was **false in the statistical sense**. Our four transformations do not preserve the same information:

The ordered grid performs a many-to-one compression,

The wide format destroys part of the continuous metric,

The continuous axis preserves it and adds a quadratic basis,

The hierarchical form injects a structure absent from the source covariates.

Four operations are therefore conflated under the word "layout": arrangement, encoding, feature engineering, restriction of the function class. **We do not disentangle them.** What we establish is therefore not that "representation matters", but that the construction of the feature space remains decisive after equalisation of a global complexity budget. This is less spectacular and far harder to refute.

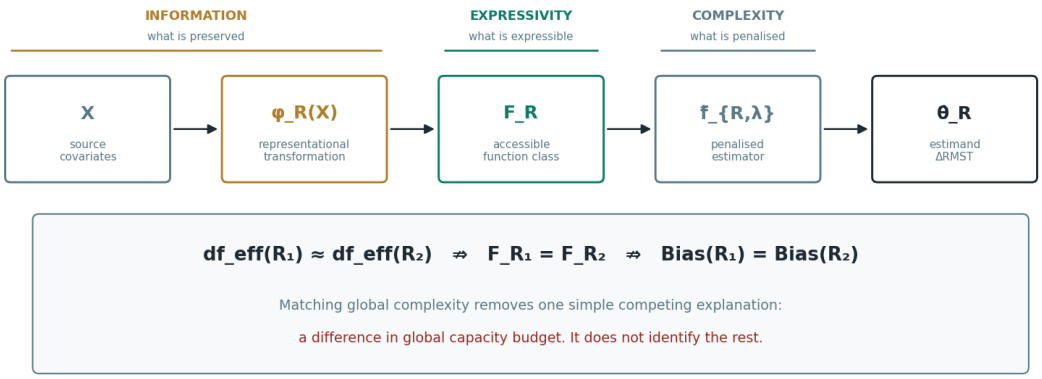

Figure 1. The representational chain and its three levels. Equalising effective degrees of freedom equalises neither the accessible function classes nor the biases.

2. Three levels covered by a single word

The word *symmetry* covers three distinct things, and their conflation is the principal conceptual weakness of the literature on inductive bias.

The first level is the **structure of the input space** (order, metric, neighbourhood, hierarchy) that the transformation fixes.

The second is the **functional hypothesis** (smoothness, locality, additivity) that the transformation-model pair makes cheap or otherwise.

The third alone is **symmetry in the strict sense**: invariance under an explicitly defined group,

$f(gx) = f(x)$ for $g \in G$.

Writing the model as the composition $f = h \circ \phi$, the effective invariance depends jointly on both terms. The transformation, which fixes ϕ, does not determine it alone.

Our experiments test the second level, not the third. We shall therefore speak of *structural adequacy*, and reserve *symmetry* for the theoretical framework.

3. State of the art

Six lines of work intersect here, none of which, on its own, names its object and is therefore self-sufficient.

The first is that of **inductive bias and invariance**, whose most accomplished formulation is geometric: an architecture encodes a hypothesis of invariance under a group, from which parameter sharing follows [1,2,3,4,5]. This vocabulary was forged for images and graphs.

The second is that of **deep survival models**, up to the piecewise exponential framework that grounds the family used here [6,7,8,9,10]. All of them vary the architecture; none varies the transformation at fixed architecture.

The third is that of **tabular learning**, which asks how to encode a variable, not what structure its transformation declares [11,12,13,14].

The fourth is that of **population-adjusted indirect comparison**. The reconstruction of individual data from published curves [15] feeds matched comparison [16,17,18]; that line consumes reconstructed data, whereas our testbed produces it.

The fifth is that of **functional analysis and temporal regularisation**: basis expansions, tensor-product smoothing [19], varying-coefficient models [20], regularisations imposing sparsity of differences [21,22]. The trade-off our testbed brings to light — separate parameters per interval against a shared smooth function — is an old and well-characterised problem there. Our contribution is not to discover it but to reread it in terms of transformation.

The sixth is that of **effective dimensions and spectral alignment**: kernel methods, effective dimensions, Bayesian priors, where the question is not how many degrees of freedom an estimator consumes but in which directions it spends them [29,30,31,32,33].

4. Data

4.1 What the reconstruction produces

The inversion of the Kaplan-Meier equations, following the method of Guyot and colleagues [15] accepted by NICE and the HAS, produces individual time-event-censoring datasets validated against published hazard ratios.

We used this method to recreate "source patients" from the survival curves presented in the articles reporting the ZUMA-7 and TRANSFORM trials, so as to obtain raw material for our testbed.

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

It does not restore the covariates: the inversion operates on an aggregate curve.

4.2 No clinical conclusion follows from this study

We use a **semi-synthetic simulation calibrated on published covariate margins** [28]. The presence of clinical cohort names in a methodological article spontaneously produces an impression of empirical validation. It is therefore necessary to be explicit.

Clinical data serve solely to calibrate certain marginal distributions of covariates. No conclusion about lymphoma, about the effectiveness of CAR-T, about treatment heterogeneity or about clinical decision-making follows from this work.

What comes from the TRANSFORM and ALYCANTE cohorts: the marginal distributions, nothing else. What does not come from them: the correlations, the causal mechanism, the form of the temporal heterogeneity, the interactions, the real censoring regime.

Binary covariates are drawn by thresholding a Gaussian copula at the published margins, continuous ones by inverse transform of a log-normal.

That is to say: we fabricate fictitious patients whose characteristics have the same proportions as those published, and remain plausibly linked to one another — the yes/no variables through a threshold set on the known proportion, the numerical values through a distribution that reproduces the usual shape of biological quantities.

5. Methods

5.1 The four transformations

The **wide format** discretises each continuous covariate into quantile bands encoded as indicators.

The **ordered grid** aggregates the covariates into a single score, discretised into levels.

The **continuous axis** keeps the covariates as continuous dimensions, augmented by their quadratic terms.

The **hierarchical form** crosses the covariates with a binary severity stratum whose encoding is declared and fixed.

5.2 A single estimation family, and what this implies

A discrete-time survival model fitted by penalised logistic regression in person-period format. This choice reduces inductive biases that would compete with the one under study, but this family is not opinion-free: it imposes an additive form, a link function, a quadratic penalty, an interval approximation.

It follows that the study bears on a slice of the space of questions:

R × DGP | M = $\mathbf{M_0}$

where

- R denotes the representation,
- DGP the mechanism,
- M the estimation family.

We vary the first two terms and fix the third. Whether the observed differences are intrinsic to the function spaces or stem from their interaction with this particular penalty remains open, and constitutes the natural extension of this work (Section 12).

5.3 Estimand

Marginal difference in restricted mean survival time at a fixed horizon, by g-formula:

$$\Delta\mathrm{RMST}(\tau) = \int_0^\tau S^1(t)\,dt - \int_0^\tau S^0(t)\,dt$$

A conditional coefficient depends on the covariates included, so that two transformations compared on their own coefficient do not aim at the same target. The marginal RMST is invariant to specification.

We measure the average lifetime gained over a fixed window — a quantity that remains the same whatever the way the model is built — rather than an effect "all other things being equal", which would change meaning with each layout and make the comparison impossible.

5.4 Complexity matching

The trace of the smoothing operator is the standard measure of the effective degrees of freedom of a penalised linear estimator [29,30]:

$$\mathrm{df}_{\mathrm{eff}} = \mathrm{tr}[(X^{\top}WX + \lambda P)^{-1}X^{\top}WX]$$

The target is the one reached by the **oracle**, a fit on the true latent confounder; the penalty parameter of each transformation is solved by dichotomic search so as to reach it before any reading of the bias.

Two reservations:

Matching equalises a **quantity, not a geometry**: two operators with the same trace may have very different spectra (Section 9).

And the oracle is **privileged information**; this reservation is lifted in Section 8.5, where a target selected out of sample reproduces the ranking.

5.5 Recoverability, design and preregistration

Recoverability is judged at the level of the condition, on the mean oracle bias, not replication by replication, otherwise the sampling noise of the oracle at finite sample size would be conflated with structural non-identifiability.

We check that a fabricated (synthetic) world is sound by looking at whether the benchmark model recovers the right answer *on average*, not at every draw: otherwise we would eliminate valid worlds on a simple stroke of bad luck, mistaking it for a defect of substance.

Partially factorial design: a core crossing two sets of margins, three confounding strengths, three regimes and three levels of spectral regularity (54 conditions); a robustness ring varying one factor at a time (6 conditions). Two thousand replications per condition, a number derived from a declared uncertainty constraint [26].

The code, the protocol and the analysis plan were deposited on OSF and timestamped before execution; the run verifies the code fingerprint at launch and refuses to execute if it differs.

6. Prior exploratory campaign (non-confirmatory)

No result in this section is retained as probative. It explains why the confirmatory design has the shape it has.

An initial experiment suggested an advantage for the shared basis; a design with matched degrees of freedom showed that this advantage was not attributable to continuity. An ancillary experiment gave the continuous axis as the winner under multiple and temporal confounding, the wide format under the non-linear regime, until an objection test reversed the reading: a continuous axis enriched with a quadratic term moved ahead everywhere. The wide format prevailed only because it was being compared with a misspecified continuous axis.

It is this test that produced the preregistered hypothesis: if the advantage rests on the adequacy between the expressible form and the real form, a transformation must win exactly in those regimes whose structure its function class matches.

7. Confirmatory results

Sixty conditions, two thousand replications, two sets of margins. Four conditions excluded from the primary analysis, all with hidden confounding. No core condition excluded.

7.1 Primary result

Confounding regime	n	Wide	Grid	Continuous	Hierarch.
Multiple	12	0.2318	0.2313	**0.2147**	0.2195
Temporal	12	1.0171	1.1129	**0.9265**	0.9720
Non-linear	12	0.0241	0.0339	**0.0223**	0.0850

Absolute value of the bias on ΔRMST, at matched effective complexity. Monte-Carlo standard error ≈ 0.004.

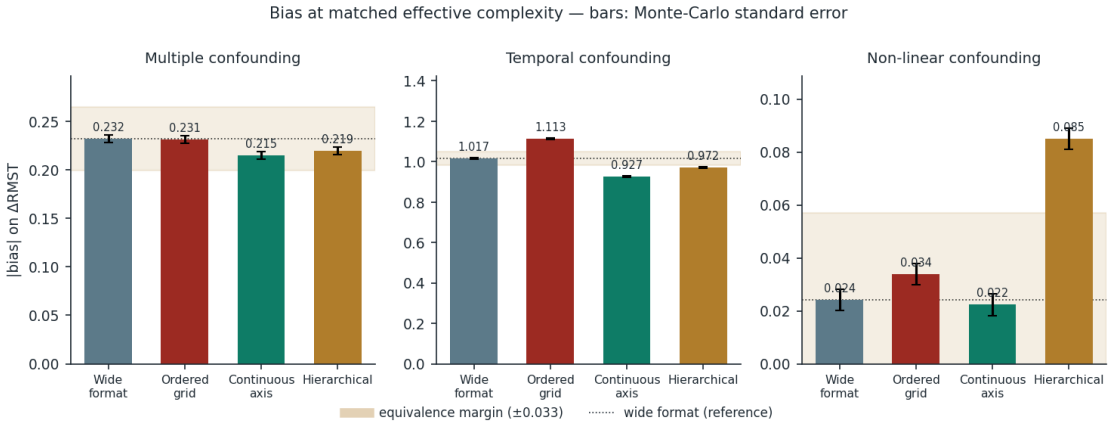

Figure 2. Bias by regime and transformation. Bars: Monte-Carlo standard error. Gold band: equivalence margin around the wide format.

7.2 Decision by equivalence margin

Regime	Comparison	Difference	> margin (0.033)?
multiple	continuous vs wide	0.017	no
multiple	hierarchical vs wide	0.012	no
temporal	continuous vs wide	0.091	**yes**
temporal	hierarchical vs wide	0.045	**yes**
temporal	grid, *worse* than wide	0.096	**yes**
non-linear	continuous vs wide	0.002	no
non-linear	hierarchical, *worse* than wide	0.061	**yes**

At two thousand replications, the multiplicity correction rejects equality in twelve conditions out of twelve for the continuous axis under multiple confounding. The corresponding difference is half the margin. We conclude on the margin, not on the rejection.

The origin of the margin must be stated without embellishment: it comes from the minimum detectable difference observed during the exploratory campaign. Neither a clinically grounded quantity, nor a declared fraction of the RMST. Its robustness is examined in Section 8.1.

7.3 Monte-Carlo stability

A run at one hundred and twenty replications on a separate machine gives the same ordering across the three regimes, to within less than three thousandths. The standard error goes from 0.0160 to 0.0040, a factor of 4.04 against a theoretical square root of 4.08.

This precision bears on the error conditional on each world. It says nothing about the average performance over a distribution of worlds (see Section 8.3).

7.4 Structural failure of the ordered grid

The wide format, the continuous axis and the hierarchical form reach the target trace in all conditions. **The grid reaches it in none.** Having reduced the covariates to a single score, its function class is capped at a complexity below that of the oracle, whatever the penalty.

The comparison is therefore not "four transformations at matched complexity": it is three matched transformations, and a fourth structurally under-parameterised one whose incapacity is itself a prespecified result.

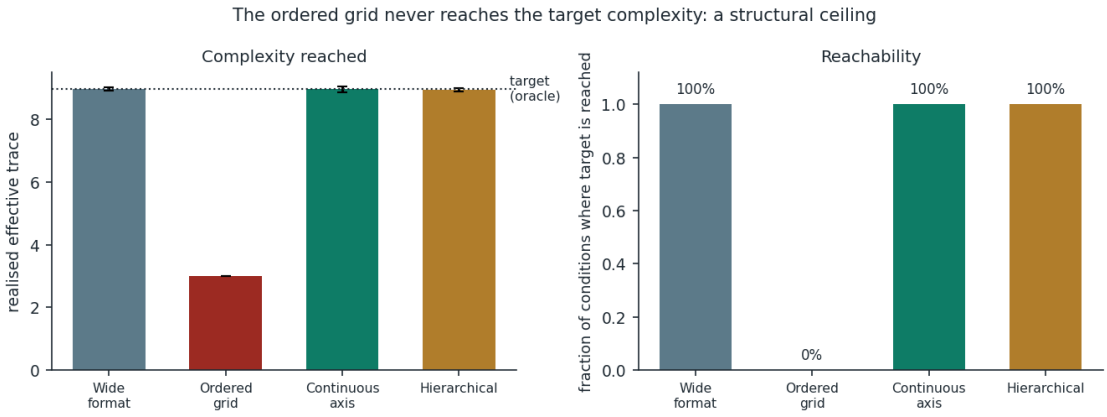

Figure 3. Realised complexity and fraction of conditions in which the target is reached.

8. Post-registration robustness and stress tests

The analyses in this section are **post-hoc and outside the preregistration**. They test the robustness of the result in Section 7; none adds any confirmatory claim to it.

8.1 Sensitivity to the margin

Over $\delta \in [0.020; 0.050]$, the verdict is refuted at every point, on both sets of margins. The decision boundary, by dichotomy, lies at $\delta \approx 0.140$: a margin **4.2 times larger** than the one retained would be required to preserve H1, that is, 26.8 Monte-Carlo standard errors.

δ	multiple: better / worse	temporal: better / worse	non-linear: better / worse
0.020	8 / 0	22 / 12	0 / 15
0.033	**0 / 0**	18 / 12	0 / 14
0.050	0 / 0	12 / 12	0 / 7

The multiple regime loses all of its exceedances between 0.020 and 0.033; the temporal one retains eighteen. The margin sits just above the noise of the first regime and well below the signal of the second (an adequacy observed, not constructed).

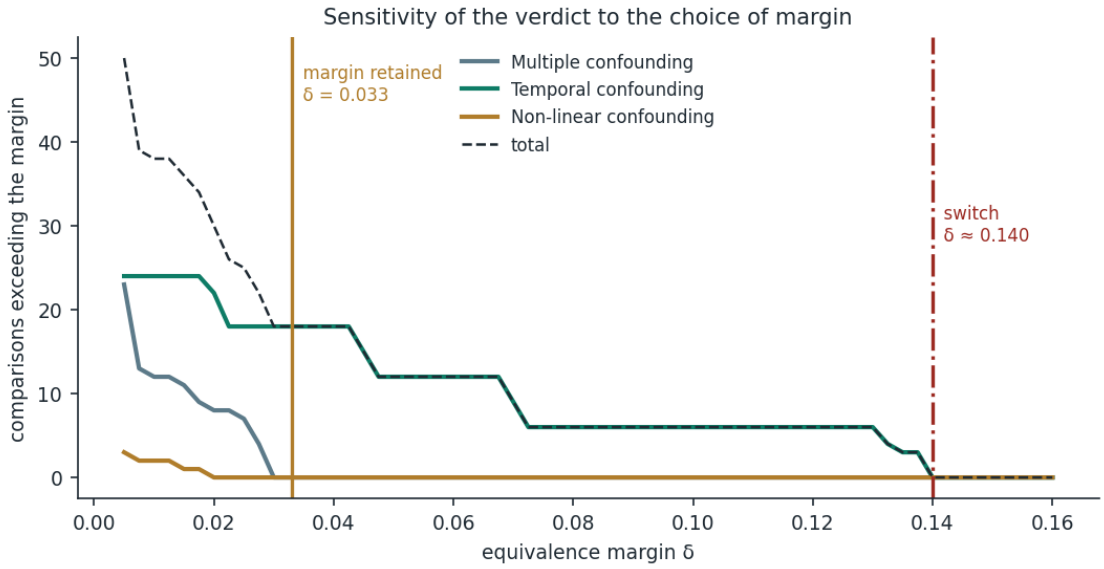

Figure 4. Comparisons exceeding the margin as a function of δ. Gold line: margin retained. Red line: switching boundary.

8.2 Bias-variance decomposition

Matching also equalises the variance: relative range of standard deviations across transformations from 1.6% to 3.6%. Nothing guaranteed this.

But the ranking switches under the non-linear regime: by bias, the continuous axis leads the wide format; by RMSE, the order reverses. **Any claim of dominance must name its criterion.**

Regime	Share of variance in the MSE	Dominated by
Multiple confounding	52.8%	variance
Temporal confounding	**5.6%**	**bias**
Non-linear confounding	93.0%	variance

Our strongest conclusions bear on the regime in which the choice of estimand is best justified; the weakest on those in which it is least so.

8.3 Between-world variability

Four hundred worlds drawn from a declared distribution, twenty-five replications each; seven quantities that the design treated as known become random here.

Transformation	Between-world var.	Within-world var.	Share between
Wide format	0.2935	0.0363	89.0%
Ordered grid	0.3362	0.0370	90.1%
Continuous axis	0.2441	0.0365	87.0%
Hierarchical	0.2836	0.0368	88.5%

88.6% on average. The standard deviation of performance between worlds is about 0.54, one hundred and thirty times the Monte-Carlo standard error of the confirmatory run.

Two thousand replications do not make up for sixty worlds.

The appropriate estimand is a probability of superiority, $\Pi(R_1, R_2) = P\_DGP(|bias(R_1)| < |bias(R_2)|)$:

Comparison	Π	Standard error
Continuous axis beats wide format	**0.730**	0.022
Continuous axis beats hierarchical	**0.800**	0.020
Hierarchical beats wide format	0.657	0.024

The average ranking is identical to that of the confirmatory run. But the continuous axis is the best in only **63.0%** of worlds: dominance is probabilistic, not universal.

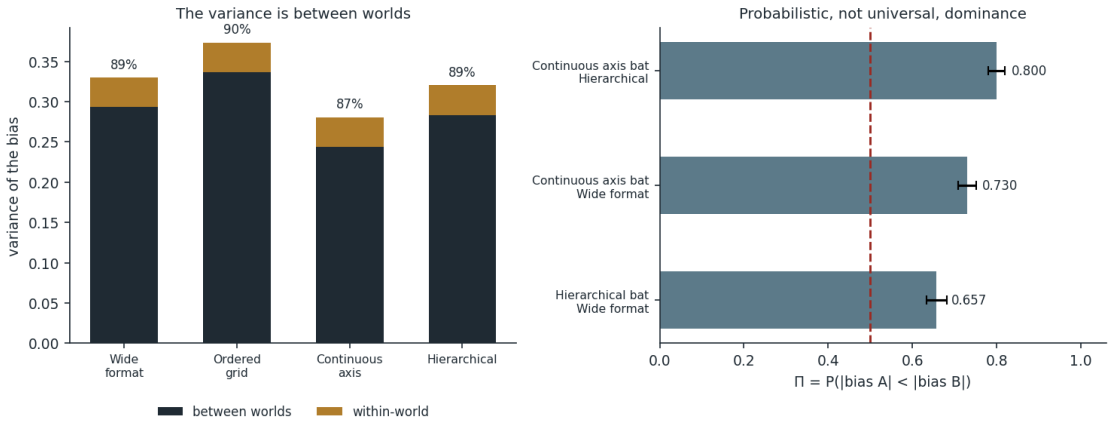

Figure 5. Left: decomposition of the variance. Right: probabilities of superiority over the distribution of worlds.

8.4 Sensitivity to sample size

Six sample sizes from 200 to 5,000, one hundred and twenty worlds, matched comparison. Π remains flat: 0.783 at 200, 0.800 at 5,000; slope against log n of +0.002. Both absolute biases do decrease with sample size (the estimators improve) but the probability that one beats the other does not move.

The aggregate masks three opposing directions: the advantage grows under multiple confounding (0.700 to 0.933), plateaus under temporal confounding, and declines towards chance under non-linear confounding (0.643 to 0.500).

8.5 Sensitivity to the complexity target

Over multiples of 0.5 to 1.5 times the oracle trace, Π varies from 0.770 to 0.790, against a standard error of 0.024: **the ranking does not depend on the level of complexity**, over a factor of three.

A target selected by cross-validation on the out-of-sample deviance, without privileged information, gives Π = 0.760 ± 0.025 against 0.770 ± 0.024 for the oracle, that is, a difference of −0.010 ± 0.035.

Matching therefore does not necessarily require an oracle target: in this testbed, a complexity target selected out of sample reproduces the ranking obtained with the oracle target.

This target grows with sample size (2.66, then 4.95, then 7.98) where the oracle target saturated (8.74 to 8.98). The channel invoked by inductive-bias theory is therefore open, and Π does not decrease for all that: 0.800, 0.720, 0.760. The persistence of the advantage observed in 8.4 is not an artefact of the capping of complexity.

8.6 Interval coverage

The frozen base does not produce intervals; we construct them by the delta method on the g-formula, with Wood's Bayesian variance for penalised estimation.

The variance is correctly estimated: se/sd ratio from 1.01 to 1.03. But coverage follows the bias-to-dispersion ratio.

Regime	bias / sd	Coverage
Non-linear	0.84	**94.8%**
Multiple	1.47	70.8%
Temporal	5.72	**7.6%**

The under-coverage does not come from a poor estimation of the variance; it results from the bias of the estimator. It nevertheless constitutes an **inferential failure** if these intervals are used to cover the causal estimand.

One tension must be declared: the temporal regime carries the conclusion of Section 7, and it is the one in which coverage is most degraded. The bias being *differential* between transformations, the comparison of biases across transformations remains interpretable; but in that regime, none of the four would provide a usable interval in practice.

8.7 Non-Gaussian dependence

Six copula families declared before execution (Gaussian as control, Clayton, Gumbel, Student, mixture of subpopulations, arm-dependent) over six hundred worlds. These families were defined independently of any question of representation.

Family	Π(continuous beats wide)	±	worlds
Gaussian (control)	0.696	0.046	102
Clayton (lower tail)	**0.589**	0.044	124
Gumbel (upper tail)	0.760	0.043	100
Student (heavy tails)	0.771	0.043	96
Mixture	0.802	0.042	91
Arm-dependent	0.701	0.049	87

Difference between the non-Gaussian families and the control: **+0.021 ± 0.050**, null within the noise. The continuous axis remains the best in all six families.

Two observations. The Gaussian control lies **below** three of the five irregular families, which makes less plausible the simple hypothesis that the Gaussian regularity of the testbed might have been chosen so as to favour the continuous axis. The support of the mechanisms nevertheless remains defined by the authors (Section 10.3). And Clayton is the only family below 0.60: a lower-tail

dependence is the structure that the function class of the continuous axis expresses least well, which indicates where to look for a counter-example.

8.8 Negative control: hidden confounding

The four conditions with hidden confounding are excluded from the primary analysis, the oracle failing to recover the truth there (mean biases from −0.72 to −2.14). In these conditions, **no transformation does usefully better than another**: all of them fail.

A transformation can improve functional approximation. It does not manufacture causal identifiability.

This control is essential to the reading of the whole article. It prevents an abusive extrapolation of the result towards representations, synthetic data or digital twins, where one might be tempted to believe that a good representation recovers the right mechanism. Operational reservation: in the real world, the analyst does not know on which side of this boundary they stand.

9. Mechanistic investigation

9.1 The spectral premise is established

Among the three genuinely matched transformations:

Transformation	Trace	Effective directions	Gini	|bias|
Wide format	8.989	15.5	0.181	0.445
Continuous axis	8.972	10.9	**0.457**	**0.408**
Hierarchical	8.941	15.8	0.371	0.448

Traces matched to within 0.53%, concentration index varying by 82%. Two pipelines of identical complexity distribute that complexity very differently. The effective degree of freedom therefore does not suffice to characterise an estimator.

Three ways of arranging the data receive exactly the same budget of flexibility but spend it very differently: the standard measure of complexity says how much a model can adjust, never in which direction.

9.2 The proposed mechanism is not supported

The correlation between spectral concentration and bias is null: −0.07, −0.06, +0.05 depending on the regime. A scalar statistic of concentration does not replace the trace.

A recovery was possible: perhaps what matters is not the *number* of directions but their *alignment* with the signal. We therefore projected the true confounding effect onto the eigenbasis of the operator, distinguishing **expressivity** (the share of the true effect falling within the design space) and **retention** (the share of what is expressible surviving the shrinkage).

We checked whether what matters is shining the light in the right place rather than broadly, by separating two questions: can the model represent the sought form at all, and does it preserve that form once found? It is the first that decides.

Transformation	Expressivity	Retention	Alignment	|bias|
Wide format	0.273	0.717	0.190	0.503
Continuous axis	0.270	**0.865**	0.230	**0.467**
Hierarchical	**0.363**	0.603	0.250	0.501

Within-regime correlations with the bias:

Regime	Expressivity	Retention
Multiple confounding	**−0.670**	+0.171
Temporal confounding	**−0.556**	+0.154
Non-linear confounding	+0.515	**−0.620**

The results are **more compatible with an explanation by the expressivity of the function class than with an explanation by spectral concentration, without any single descriptor accounting for all three regimes.** Expressivity predicts two regimes out of three and reverses sign in the third (the one that Section 8.2 shows to be dominated at 93% by variance).

This decomposition connects to an established literature: the generalisation error in kernel regression depends jointly on the spectrum and on the alignment of the target with the eigenfunctions [31,32]. A better-grounded descriptor exists and was not used, the cumulative power distribution of Canatar and colleagues [32]. A genuine test of the conjecture should go through it, and preregister it.

9.3 A cross-cutting result: aggregation masks the mechanism

Twice in this work, an average over heterogeneous families produced a conclusion that the detail contradicts. In Section 9.2, the global correlation between alignment and bias is positive whereas two regimes out of three give a negative correlation. In Section 8.4, Π is flat across sample size whereas the three regimes evolve in opposing directions.

This is not a methodological anecdote but a cross-cutting result:

Performances aggregated over heterogeneous families of mechanisms can mask precisely the interaction that the study seeks to characterise.

Formally, E_DGP[L(R, DGP)] does not suffice to characterise L(R, DGP) when the interaction R × DGP is precisely the object of interest. A global benchmark may then rank methods correctly while explaining incorrectly why they prevail.

An average of performance across a set of situations tells you which method prevails most often. It does not tell you in which situations, nor why. When it is precisely the interaction between method and situation that is under study, a global ranking can be correct and its explanation false.

This observation goes beyond our testbed and concerns any practice of comparison over heterogeneous collections of datasets.

10. Discussion

10.1 What matching does, and does not do

The methodological contribution amounts to a single gesture, inexpensive and consequential. Its scope must be stated exactly.

Matching global complexity removes one simple competing explanation: a difference in global capacity budget. It does not identify the remaining explanation.

This is what the whole body of work shows: the differences persist after matching, and our mechanistic investigation (Section 9) fails to reduce them to a single descriptor.

Two distinct questions must be separated here, and their conflation explains part of the disagreements about representation comparisons. *Which pipeline works best?* calls for a comparison at optimal configuration, where imposing equal complexity would be artificial. *What share of the difference comes from the representation?* on the contrary requires matching. This work addresses the second, and its conclusions do not transfer to the first.

10.2 In relation to classical statistical methods

The trade-off brought to light (separate parameters per interval against a shared smooth function) is an old problem of functional regularisation. We grant this without reservation. Our contribution is to show that it governs differences that the literature commonly attributes to "representation" or to "format", and to provide a protocol that separates the two.

10.3 Simulation bias: parametric uncertainty and structural uncertainty

The most serious objection is that the authors write both the mechanism and the representations intended to succeed within it. Four devices respond to it in part:

The covariate margins come from third-party publications,

The form of the confounder is sampled along an axis of regularity that includes forms no transformation can express,

The design is timestamped before execution,

The dependence structure is sampled across six families defined independently of any question of representation (8.7).

What remains must be formulated precisely, since the vague acknowledgement "the mechanism is still written by us" understates the problem:

The distribution of worlds is random **conditionally on a family of mechanisms whose structural support remains defined by the authors.**

We have randomised the parameters, the worlds, the covariance structures, the sample sizes and the complexity budgets. We have not randomised the space of admissible mechanisms. Parametric uncertainty is broad; structural uncertainty remains limited.

11. What is controlled, what is not

Dimension	Control or stress test	Residual limitation
Global complexity	matched by the trace, swept over a factor of three (8.5)	full spectrum of the operators not equalised
Monte-Carlo uncertainty	controlled, 2,000 replications	
Covariate margins	calibrated on third-party publications	actually observed dependences unknown
Sample size	tested from 200 to 5,000 (8.4)	general asymptotic regime not reached
Dependence structure	six families, three with tail dependence (8.7)	universal space of dependences not covered
Estimation family	fixed at M_0	no other family tested
Causal mechanism	three regimes, distribution of worlds (8.3)	structural support defined by the authors
Identifiability	oracle and negative control (8.8)	hidden confounding not detectable in practice
Interval inference	coverage measured (8.6)	intervals unusable under dominant bias

"Control or stress test" does not mean neutralisation: several dimensions are explored, not mastered. The right-hand column states what remains open in each case.

This map replaces a linear reading of the precautions scattered through the text. It is the honest complement to the executive summary.

A measure that does not measure what its name announces. The protocol provided for quantifying a generative uncertainty by resampling the generator. The implementation resamples the replications already produced, reproducing the Monte-Carlo variance: the ratio between the two quantities comes out at 1.01. We declare this as a limitation, not as a result.

12. Scope of the inference and next falsification

Result established within the prespecified experimental framework. In this penalised family of discrete-time survival models, at matched global complexity, distinct representational transformations produce different biases despite matched global complexity, and the margin exceedances occur in a structured way across regimes. Their interpretation through the expressivity of the function class belongs to Level B, not to Level A. Over a distribution of worlds, the advantage is probabilistic: the continuous axis produces a bias lower than the wide format in 73% of worlds, and it is the best of the three matched transformations in 63% of them. This advantage depends neither on the level of complexity retained — invariant over a factor of three — nor on access to privileged

information, nor on the regularity of the dependence structure of the covariates. It does not dissipate when sample size grows from 200 to 3,000 and the complexity target, selected by cross-validation, grows with it.

Methodological conjecture. The principle should extend to families for which a relevant notion of effective complexity can be defined. Conceptually plausible, empirically undemonstrated: in a deep network, the trace of the smoothing operator has no immediate analogue.

The priority independent test is not more simulations. The marginal return of a testbed whose authors define the structural support is decreasing, and ten thousand additional worlds would change nothing. The experiment that would settle matters is of another nature: a third party specifies the mechanisms, the parameters, the distributions, the interactions and the temporal dynamics, without knowing the results obtained by the representations; the pipeline is then executed blind. No single experiment will be "decisive" in the strict sense, but this one would bear on the only dimension our analyses cannot reach. This is the extension we identify as the priority, together with the extension to other estimation families — the R × M × DGP question of which this work studies only one slice.

Positioning. This work does not show that certain representations are intrinsically better. It contributes to the methodology of attributional comparisons of representations. Survival is here the experimental testbed, not the object of the doctrine.

Code availability

Item	Reference
Preregistration	OSF DOI 10.17605/OSF.IO/TPG5S
Frozen base	plasmode.py, SHA-256 9ca0aa6137c3a289…792592a
Confirmatory run	plasmode_e2.py, run_parallel.py
Post-hoc analyses	sensibilite_marge.py, decomposition_variance.py, analyse_spectrale.py, alignement_spectral.py, variabilite_mondes.py, effectif_complexite.py, cible_complexite.py, couverture.py, dgp_adversarial.py
Environment	Python 3.12.3, numpy 2.4.4, scipy 1.17.1, pandas 2.3.3

The confirmatory run verifies the fingerprint of the frozen base at launch and refuses to execute if it differs. No patient data are used at any stage.

References

Mitchell TM. The need for biases in learning generalizations. Rutgers University Technical Report CBM-TR-117; 1980.

Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798-1828.

Cohen T, Welling M. Group equivariant convolutional networks. In: Proceedings of ICML; 2016.

Battaglia PW, Hamrick JB, Bapst V, et al. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261; 2018.

Bronstein MM, Bruna J, Cohen T, Veličković P. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. arXiv:2104.13478; 2021.

Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv. BMC Med Res Methodol. 2018;18:24.

Lee C, Zame WR, Yoon J, van der Schaar M. DeepHit. In: Proceedings of AAAI; 2018.

Lee C, Yoon J, van der Schaar M. Dynamic-DeepHit. IEEE Trans Biomed Eng. 2020;67(1):122-133.

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.

Bender A, Rügamer D, Scheipl F, Bischl B. A general machine learning framework for survival analysis. In: Proceedings of ECML PKDD. Springer; 2021.

Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer. arXiv:2012.06678; 2020.

Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data. Adv Neural Inf Process Syst. 2021;34:18932-18943.

Somepalli G, Goldblum M, Schwarzschild A, Bruss CB, Goldstein T. SAINT. arXiv:2106.01342; 2021.

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.

Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data. BMC Med Res Methodol. 2012;12:9.

Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials. Pharmacoeconomics. 2010;28(10):935-945.

Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU Technical Support Document 18. Sheffield: NICE Decision Support Unit; 2016.

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.

Wood SN. Generalized additive models: an introduction with R. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2017.

Hastie T, Tibshirani R. Varying-coefficient models. J R Stat Soc Series B. 1993;55(4):757-796.

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.

Kim SJ, Koh K, Boyd S, Gorinevsky D. l1 trend filtering. SIAM Rev. 2009;51(2):339-360.

Cox DR. Regression models and life-tables. J R Stat Soc Series B. 1972;34(2):187-220.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-481.

Pawel S, Kook L, Reeve K. Pitfalls and potentials in simulation studies. Biom J. 2024;66(1):e2200091.

Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-2102.

Boulesteix AL, Lauer S, Eugster MJ. A plea for neutral comparison studies in computational sciences. PLoS One. 2013;8(4):e61562.

Schreck N, Slynko A, Saadati M, Benner A. Statistical plasmode simulations. Stat Med. 2024;43(9):1804-1825.

Buja A, Hastie T, Tibshirani R. Linear smoothers and additive models. Ann Stat. 1989;17(2):453-510.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2009. §7.6.

Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel-target alignment. Adv Neural Inf Process Syst. 2001;14.

Canatar A, Bordelon B, Pehlevan C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nat Commun. 2021;12:2914.

Kaufman S, Rosset S. When does more regularization imply fewer degrees of freedom? Biometrika. 2014;101(4):771-784.