LEARNING AND UNCERTAINTY IN GENERALIZED BAYESIAN DISCRETE-TIME SURVIVAL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

External survival models can improve prediction in target cohorts with few events, but the inferential consequences of borrowing from a fixed teacher remain insufficiently characterized. Existing discrete-time survival distillation methods primarily focus on point estimation; how teacher guidance affects survival-probability uncertainty requires further analysis. We develop a generalized Bayesian framework for discrete-time survival distillation that transfers interval-level conditional event probabilities from a fixed external teacher to a flexible neural student through a prediction-level Kullback–Leibler divergence, allowing teacher and student models to differ in architecture and parameterization. The resulting update combines the target survival likelihood, teacher guidance, and structural prior regularization to define a posterior over individualized survival predictions. Our local analysis quantifies how borrowing shifts fitted survival probabilities and distinguishes curvature-based posterior variance from sampling variance, which also depends on patient-level score covariance. We derive a generalized subject-deletion predictive criterion for selecting the borrowing weight. Experiments under compatible and imperfect teacher guidance show that borrowing can improve survival prediction and reduce sampling variability, whereas excessive reliance on an imperfect teacher can worsen prediction and introduce substantial bias. Real-data analyses show improved agreement with a pooled-data reference in breast cancer survival and predictive gains from borrowing across kidney-transplant populations. The framework supports flexible survival borrowing while clarifying its inferential benefits and limitations.

1 INTRODUCTION

Estimation of patient-specific survival probabilities can be imprecise in target cohorts with few observed events or substantial censoring. Prediction-level transfer can borrow from an external model trained on a larger cohort without accessing its training records (Hickey et al., 2024). Conditional event probabilities on a common time grid allow teacher and student models to differ in architecture and parameterization while expressing their predictions as survival curves. Survival distillation combines these predictions with a target likelihood on observed risk sets (Deng et al., 2026). We study patient-specific survival-probability inference under imperfect external guidance.

We study patient-specific survival probabilities, or equivalently absolute event risks, and their estimation uncertainty in finite, censored cohorts (Riley et al., 2025; Kompa et al., 2021). Although Bayesian survival models already quantify this uncertainty (Feng & Zhao, 2021), the inferential effects of external borrowing require further analysis. We therefore examine how borrowing from a fixed external teacher affects these estimates and their uncertainty, and how to select borrowing strength using censored target outcomes. Specifically, we develop a generalized Bayesian formulation that combines the target survival likelihood, KL guidance from a fixed teacher, and a regularizing prior. Posterior draws yield survival curves, providing posterior means and pointwise credible intervals for patient-specific survival probabilities. From this posterior, we derive a criterion for selecting the borrowing weight η by predicting each patient's observed outcome after removing their contributions from both the likelihood and guidance losses.

In an idealized exact-guidance benchmark, the teacher supplies true hazards representable by the student. Under the local conditions in Section 4 with a fixed, finite borrowing weight, our asymptotic

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

analysis shows that borrowing reduces both sampling and posterior variances relative to target-only inference, but sampling variance contracts more strongly. Borrowing increases curvature without adding leading score variability (Miller, 2021), explaining this difference and the interval rescaling needed for asymptotically nominal pointwise coverage in this benchmark.

With imperfect teacher, potential sampling-variance reduction must be balanced against teacher-induced bias. For specified local discrepancies in calibration, covariate effects, or temporal patterns, we derive first-order bias in survival-probability estimates, accounting for student curvature. We identify borrowing weights for which variance reduction outweighs added squared bias, specializing the distillation trade-off studied by Menon et al. (2021) to survival-probability estimation. Because student flexibility limits which teacher predictions can be represented, our simulations examine student structure and teacher discrepancy separately.

Our contributions are threefold. First, we formulate discrete-time survival distillation as a generalized posterior for patient-specific survival probabilities under censoring. Second, we characterize beneficial borrowing under exact guidance and local teacher–target mismatch, distinguishing posterior spread from sampling variability. Third, we derive a predictive borrowing criterion that removes each patient's likelihood and guidance contributions. Linear controls check the theoretical approximations; neural simulations assess prediction and interval coverage after selection. Breast-cancer and transplant analyses show predictive gains and narrower intervals, with improved agreement with a pooled-data reference in breast cancer.

2 Related Work

External information and survival distillation. External borrowing depends on what source information is available. In their standard shared-parameter form, power priors discount an external-data likelihood (Ibrahim & Chen, 2000; Ibrahim et al., 2003; 2015), while CoxTL combines density-ratio and importance weighting to address covariate and coefficient shifts using source records (Lu et al., 2026). Our external input consists only of teacher probabilities at target covariates, requiring neither its training likelihood nor a correspondence between teacher and student parameters. RECaST supports prediction-based borrowing through Bayesian calibration (Hickey et al., 2024). For survival outcomes, DiSKD aligns external hazards with a discrete target likelihood on observed risk sets and analyzes estimation risk under teacher heterogeneity (Deng et al., 2026). We extend this loss-based formulation to survival-probability posteriors, their uncertainty, and patient-level predictive selection of borrowing strength.

Censoring, discrete hazards, and Bayesian distillation. Prediction-level distillation permits different teacher and student architectures (Hinton et al., 2015). Bayesian Knowledge Distillation connects a classification distillation objective to a posterior mode and quantifies class-probability uncertainty (Fang et al., 2024). Adapting this construction to survival requires an observed-data likelihood for censored outcomes and a way to align time-indexed teacher guidance with that likelihood. Bayesian neural survival models have been developed in both continuous and discrete time (Eleuteri et al., 2007). BDNNSurv uses regression on pseudo-values to estimate survival probabilities and quantify their uncertainty (Feng & Zhao, 2021). We combine the discrete censored likelihood, Bernoulli hazard guidance, and a prior in an explicit generalized posterior for patient-specific survival probabilities.

Generalized Bayes and borrowing properties. Loss-based updating constructs a posterior from a loss and a prior (Bissiri et al., 2016). Under likelihood misspecification, posterior covariance can differ from the sampling covariance of the estimator (Kleijn & van der Vaart, 2012); Miller (2021) develops asymptotic theory for generalized posteriors. Sandwich adjustment and credible-region calibration adjust posterior intervals (Shaby, 2014; Syring & Martin, 2019). For survival distillation, uncertainty assessment must also account for how teacher error changes the estimated survival probability. Building on distillation risk analyses (Menon et al., 2021; Deng et al., 2026), we derive posterior variances for survival probabilities and sampling variances for their posterior means, accounting for the covariance between each patient's likelihood and guidance scores. The resulting formulas describe how exact guidance reduces each variance and give local conditions under which borrowing from an imperfect teacher lowers MSE.

3 BAYESIAN SURVIVAL DISTILLATION

Our discrete-time construction uses interval event probabilities to combine a censored target likelihood with a teacher-induced pseudo-prior on student parameters. Inference conditions on the deployed teacher; its retraining variability is not included.

3.1 TEACHER-GUIDED POSTERIOR

Let $D = \{(X_i, Y_i, \Delta_i)\}_{i=1}^n$ contain independent subjects, with covariates, observed follow-up, and single-event indicators. On a fixed grid $0 = t_0 < \cdots < t_K$, let A_{ik} indicate an observed at-risk interval and $\delta_i(t_k)$ an event in that interval. Student logits $f_{\theta,k}$ define

$$\lambda_{\theta,k}(x) = \operatorname{expit}\{f_{\theta,k}(x)\}, \qquad S_\theta(t_k \mid x) = \prod_{\ell \le k} \{1 - \lambda_{\theta,\ell}(x)\}. \tag{1}$$

Here $\lambda_{\theta,k}(x)$ is event probability conditional on reaching interval k, and $1 - S_\theta(t_k \mid x)$ is absolute event risk. The teacher supplies fixed hazards $\widetilde{\lambda}_{ik}$, collectively $\widetilde{P}$. A teacher survival curve gives $\widetilde{\lambda}_{ik} = 1 - \widetilde{S}(t_k \mid X_i)/\widetilde{S}(t_{k-1} \mid X_i)$ wherever the denominator is positive. A Cox, parametric, or neural survival teacher can therefore guide a linear or neural student through probabilities for the same endpoint and time intervals, without matching model parameters. Under conditionally independent censoring, define $r_i = -\log L_i$ and q_i by

$$\begin{aligned} r_i(\theta) &= -\sum_k A_{ik}\big[\delta_i(t_k)\log\lambda_{\theta,k}(X_i) + \{1-\delta_i(t_k)\}\log\{1-\lambda_{\theta,k}(X_i)\}\big], \\ q_i(\theta) &= \sum_k A_{ik} D_{\mathrm{KL}}\{\mathrm{Bern}(\widetilde{\lambda}_{ik}) \| \,\mathrm{Bern}(\lambda_{\theta,k}(X_i))\}. \end{aligned} \tag{2}$$

The likelihood is exact for native-discrete observations and approximates grouped continuous follow-up (Appendix E). The mask A_{ik} restricts both losses to observed at-risk intervals. Inference and predictive deletion treat patients as independent units.

The role of teacher guidance follows directly from the Bernoulli cross-entropy $c(u,p) = -u\log p - (1-u)\log(1-p)$. Define the soft event target $\bar{\delta}_{ik,\eta} = \{\delta_i(t_k) + \eta\widetilde{\lambda}_{ik}\}/(1+\eta)$. Then

$$r_i(\theta) + \eta q_i(\theta) = (1+\eta)\sum_k A_{ik}\, c\{\bar{\delta}_{ik,\eta}, \lambda_{\theta,k}(X_i)\} + C_i, \tag{3}$$

where C_i is independent of θ. Guidance changes the event targets. At a common student logit, the cell loss has curvature $(1+\eta)\lambda_{\theta,k}(1-\lambda_{\theta,k})$: borrowing can shift fitted survival probabilities while increasing local logit curvature.

We use a teacher-independent base prior $\pi_{\sigma,\tau}(\theta) \propto \exp\{-P_{\sigma,\tau}(\theta)\}$, combining a proper Gaussian base with optional smoothing of adjacent hazard logits. Smaller σ strengthens shrinkage of designated parameter blocks, and smaller τ strengthens temporal smoothing; Appendix A.1 gives the construction and the experiment-specific scales. Prior-predictive and paired sensitivity checks show how the prior affects survival-scale beliefs and borrowing comparisons (Appendix Tables 1 and 17).

For borrowing weight $\eta \ge 0$, combine the base prior with KL guidance to form a *teacher-induced pseudo-prior*, following the prior interpretation of distillation (Fang et al., 2024):

$$\pi^T_{\eta,\sigma,\tau}(\theta \mid \widetilde{P}, X, A) \propto \exp\left\{-\eta\sum_i q_i(\theta)\right\}\pi_{\sigma,\tau}(\theta),$$

where $X = (X_i)_{i=1}^n$ and $A = (A_{ik})$. It favors student hazards close to teacher predictions, with η controlling the teacher's influence relative to the target likelihood. Its dependence on observed target risk sets motivates the term pseudo-prior. Multiplication by the target likelihood gives

$$\begin{aligned} \Pi_{\eta,\sigma,\tau}(d\theta \mid D, \widetilde{P}) &\propto \left\{\prod_i L_i(\theta)\right\}\pi^T_{\eta,\sigma,\tau}(\theta \mid \widetilde{P}, X, A)\, d\theta \\ &\propto \exp\left[-\sum_i \{r_i(\theta) + \eta q_i(\theta)\}\right]\pi_{\sigma,\tau}(\theta)\, d\theta. \end{aligned} \tag{4}$$

Equation (4) defines a teacher-guided generalized posterior. Its loss-based Bayesian justification is the minimization of expected distillation loss plus KL divergence from the base prior over parameter distributions (Bissiri et al., 2016) (Appendix A.2). We retain the likelihood scale and factor $1 + \eta$; at $\eta = 0$, the pseudo-prior reduces to $\pi_{\sigma,\tau}$ and the update recovers target-only Bayes.

Each posterior draw yields a survival curve. We report $\widehat{S}_B(t_k \mid x) = \mathbb{E}_\Pi S_\theta(t_k \mid x)$ and pointwise posterior quantiles for the survival probability. The Metropolis-adjusted Langevin algorithm (MALA) (Roberts & Rosenthal, 1998) samples all student parameters; full-network inference jointly samples representation and prediction-head parameters (Appendix A).

3.2 Target-predictive borrowing selection

Borrowing weight η controls the teacher-induced pseudo-prior and trades target sampling variance against teacher-induced bias. The local MSE-optimal weight in Section 4.2 depends on unknown population quantities, so we choose one weight by prediction of held-out target outcomes, at fixed base-prior settings (σ, τ). Let Π_{-i} remove subject i from both losses while holding the teacher, preprocessing, time grid, and base prior (including its reference covariates) fixed. Define its conditional predictive ordinate $\mathrm{CPO}_i = \mathbb{E}_{\Pi_{-i}} L_i$. With $\widehat{\mathrm{CPO}}_i$ estimated from posterior draws, select

$$\widehat{\eta} \in \arg\max_{\eta \in \mathcal{G}} \sum_{i=1}^{n} \log \widehat{\mathrm{CPO}}_i(\eta; \sigma, \tau), \tag{5}$$

where the prespecified grid includes zero and selection uses only target training-cohort outcomes. Prior profiling, when used, evaluates every declared prior–weight pair and gives the target-only comparator the same prior menu. Deleting patient i removes r_i from the likelihood loss and q_i from the teacher-induced pseudo-prior, giving log importance weight $\eta q_i - \ell_i$, where $\ell_i = \log L_i$. PSIS stabilizes these weights (Vehtari et al., 2024) to estimate deleted-outcome prediction without a separate refit for every patient. Full-network PSIS reweights joint draws of representation and prediction-head parameters. The held-out score evaluates L_i alone. Appendix C gives the deletion identity, its relation to likelihood-only WAIC, and sensitivity assessment.

4 Borrowing Gains and Survival-Probability Uncertainty

For a fixed patient and horizon, let $\phi(\theta) = S_\theta(t_k \mid x)$, $\phi_0 = S_0(t_k \mid x)$, and $\bar{\phi}_{\eta,n} = \mathbb{E}_{\Pi_{\eta,n}} \phi(\theta)$. This first posterior moment is our squared-error point summary. Its repeated-cohort MSE combines squared bias with sampling variance $V_{\mathrm{samp},\phi} = \mathrm{Var}_D(\bar{\phi}_{\eta,n})$, while $V_{\mathrm{post},\phi} = \mathrm{Var}_\Pi\{\phi(\theta)\}$ describes spread within a fitted posterior. We compare *exact guidance*, $\widetilde{\lambda}_k(x) = \lambda_{0,k}(x)$, with *local teacher–target mismatch* of order $n^{-1/2}$. Exact guidance isolates the gain from borrowing, while local mismatch introduces bias on the scale of sampling error.

Throughout, η is fixed and finite, the student is correctly specified and locally regular and identifiable, and the base prior has negligible root-n influence. Local posterior normality and the moment and smoothness conditions in Appendix B.2 link posterior-mean functionals to local modes and justify the variance and MSE limits below.

4.1 Exact teacher guidance: estimation gains and uncertainty

Let θ_0 be the true student parameter. Under exact guidance, each patient has zero teacher score at truth. Both losses use the weights A_{ik}, so borrowing increases curvature from $H = \nabla^2 \mathbb{E} r_i(\theta_0)$ to $(1 + \eta)H$ without adding leading score variability. For $g = \nabla\phi(\theta_0)$ and $v_\phi = g^\top H^{-1} g > 0$, Proposition 3 gives

$$nV_{\mathrm{samp},\phi} \simeq v_\phi/(1+\eta)^2, \qquad nV_{\mathrm{post},\phi} \simeq v_\phi/(1+\eta). \tag{6}$$

The posterior mean's leading MSE equals its sampling variance here: exact guidance improves estimation, with diminishing absolute gains across fixed weights. Posterior spread contracts more slowly. Appendix B.6 connects the two spreads to interval interpretation. Appendix Tables 14 and 15 check the contraction rates and bias–spread coverage approximation numerically.

4.2 Local teacher–target mismatch: the beneficial borrowing range

Consider the teacher sequence $\widetilde{\lambda}_{n,k}(x) = \operatorname{expit}\{\operatorname{logit} \lambda_{0,k}(x) + h d_k(x)/\sqrt{n}\}$, with fixed direction d and finite amplitude h. Let B be the derivative at zero of $\mathbb{E}\nabla q_i(\theta_0)$ under a teacher-logit perturbation εd, and set $c_\phi = -g^\top H^{-1} B$. Corollary 5 yields

$$n \operatorname{MSE}(\bar{\phi}_{\eta,n}) \longrightarrow \mathcal{R}_\phi(\eta) = \frac{v_\phi + h^2 c_\phi^2 \eta^2}{(1+\eta)^2}. \tag{7}$$

The numerator reflects target sampling variation and squared teacher-induced bias. The curvature-adjusted coefficient hc_ϕ captures teacher effect on survival-probability estimates; its sign and magnitude can vary across patients and horizons. Posterior variance retains the leading expression in Equation (6). With $\kappa_\phi = h^2 c_\phi^2 / v_\phi$ measuring squared mismatch relative to target noise, strict leading-order MSE improvement over $\eta = 0$ and the local MSE-optimal weight satisfy

$$\eta(\kappa_\phi - 1) < 2 \quad (\eta > 0), \qquad \eta_\phi^* = \kappa_\phi^{-1} \quad (\kappa_\phi > 0). \tag{8}$$

Within this local $n^{-1/2}$ regime, sufficiently small positive weights improve leading-order MSE for every finite h. When $\kappa_\phi > 1$, the beneficial range ends at $2/(\kappa_\phi - 1)$. For $0 < \kappa_\phi \leq 1$, every fixed positive weight improves on $\eta = 0$, although MSE rises after η_ϕ^*. Only $\kappa_\phi = 0$ gives monotone improvement as η increases. This specializes the distillation bias–variance trade-off to survival probabilities (Menon et al., 2021). Equation (8) defines the local MSE-optimal weight for a specified survival probability. The implemented method chooses one weight over the declared grid by maximizing the estimated leave-one-patient-out log predictive score (Section 3.2), which evaluates the predictive likelihood of held-out target outcomes under censoring.

5 Simulation Experiments

We study how teacher guidance affects survival prediction and probability intervals across student models, and how teacher discrepancy alters the useful borrowing range. Simulation I varies student and event-law structure under compatible guidance. Simulation II varies teacher mappings and borrowing weights, with predictive selection.

5.1 Experimental settings

Each experiment uses 100 paired target cohorts of $n = 500$, with discrete outcomes and observed-event probability 0.628. Compatible teachers are fitted to 50,000 independent source subjects under the target event law. Main comparisons use full-state posterior sampling, a common reference prior, and fixed-weight paths over $\eta \in \{0, 0.25, 1, 3, 10, 30\}$. PSIS-selected predictions use target-training outcomes and the study-specific candidate menus detailed in Appendix D.3. We assess posterior-mean survival probabilities against truth on a fixed reference panel, and evaluate concordance, predictive deviance, and IBS on independent test cohorts. Figures compare borrowing paths and PSIS-selected predictions with target-only inference. Aggregate panels report cohort means and SDs. Appendix D gives design and metric details; supplementary analyses cover selection, interval interpretation, and numerical precision (Appendices D.3, D.6, and C.6.2).

5.2 Simulation I: Compatible Guidance across Student Models

Simulation I crosses linear and nonlinear event laws with Linear, shared-risk neural, and interval-specific MLP students. Figure 1(a,b) shows how accurate guidance reduces the estimation burden on a limited target sample. Under linear generation, borrowing brings the neural students close to Linear. Under nonlinear generation, their flexibility allows them to learn risk structure that Linear cannot represent: both neural paths continue toward lower RMSE as the Linear path levels off. The two neural students approach similar errors at stronger borrowing, suggesting that a suitable representation matters more than added complexity alone.

Panels (c,d) show prediction error and interval width decreasing together for the neural students. Under nonlinear generation, Linear also concentrates while retaining its approximation error. Panels (e–g) show accompanying mean improvements in held-out prediction after PSIS selection. The

Linear control links these gains to Section 4: independent calculations reproduce the decline in MAP risk and the distinct contraction of sampling variation and posterior spread (Appendix D.4). Borrowing therefore provides both more accurate risk estimates and a more concentrated posterior in the well-matched settings, while student structure determines the remaining error. Full summaries, time profiles, and the high-borrowing pilot appear in Appendices D.2 and D.8.

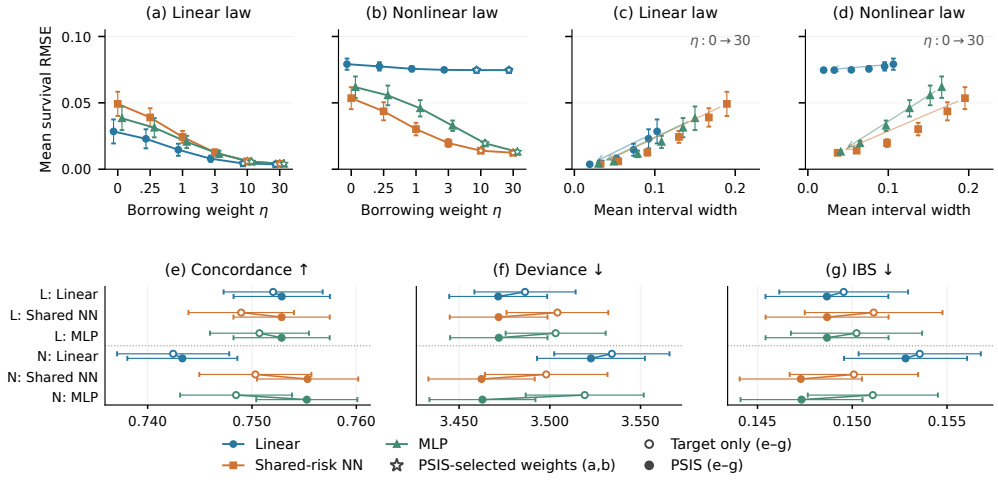


Figure 1: **Borrowing gains and interval concentration.** Mean RMSE versus (a,b) η and (c,d) mean raw interval width; stars mark PSIS-selected weights, and arrows trace $\eta = 0$ to 30. (e–g) Prediction metrics: target only (open) versus PSIS (filled), under linear (L) and nonlinear (N) laws. Bars show SDs across 100 cohorts. Students are evaluated at the six displayed weights, equally spaced in (a,b).

In the Linear control, as η increases from 0 to 30, observed MAP MSE falls from 88.97 to 1.39 in units of 10^{-5}, compared with independent predictions of 85.14 and 1.32. Across all six weights, MSE is underestimated by 4.3–10.3% and sampling variance by about 4–5%; posterior-variance predictions agree within approximately 1.1%. The full 100-cohort comparison and Monte Carlo precision are in Appendix D.4.

5.3 Simulation II: Imperfect Teacher and Useful Borrowing Ranges

Simulation II fixes the nonlinear law and shared-risk student, then introduces calibration shifts of $+0.10$ and $+0.50$, a risk-magnitude change, and a late temporal perturbation. PSIS selects over the six displayed weights. Figure 2(a) shows sustained gains under compatible guidance and initial gains that plateau under the mild calibration shift. Larger calibration and risk-magnitude errors narrow the useful borrowing range: RMSE initially falls but rises with stronger borrowing, illustrating the prediction-risk balance in Section 4.2. Appendix Table 16 separates the observed bias and sampling-variance changes. The colored stars locate the PSIS choices for the cohort illustrated in panel (b); circle sizes summarize selection frequencies across all 100 cohorts.

For the patient and cohort in panel (b), compatible and late-temporal guidance yield narrower CIF intervals than target-only inference, with posterior means close to truth. With the larger calibration shift, posterior means remain above truth and intervals contract less. These patterns are consistent with the center–spread distinction between curvature-driven concentration and teacher-induced displacement. In a separate fixed-representation population-MAP check, numerically integrating the borrowing derivative reconstructs survival-probability changes from $\eta = 1$ to 3 with 6.7–8.0% relative RMS error for the evaluated calibration and temporal perturbations (Appendix D.7.1).

PSIS-selected intervals have 94.17–98.46% mean empirical coverage and 13–70% smaller mean widths than $\eta = 0$ in four of five teacher settings; calibration $+0.10$ gives 79.62% coverage, and late-temporal final-horizon coverage is 83.18% (Appendix Table 11).

Appendix Table 9 shows higher mean concordance and lower mean deviance and IBS with PSIS than with target-only inference across all five teacher conditions. Gains are larger under compati-

ble, mildly shifted, and late-temporal guidance; larger discrepancies yield smaller gains with limited borrowing. These results suggest that predictive selection can retain useful guidance despite teacher discrepancy. Full outcomes and selection frequencies are in Appendix D.3. Population-center responses are examined in Appendix D.7.1.

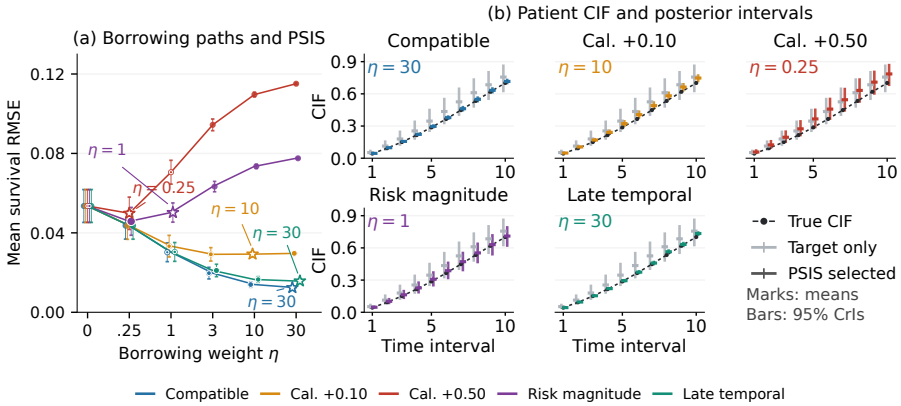


Figure 2: **Useful borrowing under imperfect teacher.** (a) Fixed-weight mean RMSE and cohort SDs; circle sizes show PSIS selection frequencies across 100 cohorts. Stars locate the example cohort's choices from (b) on these mean paths, not selected-procedure mean errors. (b) CIF estimates for one fixed patient in that cohort, at each teacher's PSIS choice. Horizontal marks show posterior means; vertical bars show pointwise 95% credible intervals. Facets share truth (black) and target-only estimates (gray); colors identify teachers. Within-interval offsets separate estimates. Both panels use full-network inference.

At the same teacher, prior, and cohort-specific PSIS-selected weight, posterior averaging reduces mean RMSE relative to the matched local MAP from 0.0632 to 0.0505 under calibration $+0.50$ and from 0.0607 to 0.0469 under risk-magnitude mismatch. Without borrowing, the corresponding reduction is from 0.0742 to 0.0535. Under compatible guidance, MAP and posterior-mean RMSE are close (0.0149 and 0.0151), indicating that the additional averaging benefit varies with teacher quality. Appendix D.5 reports every condition, cohort SDs, and numerical qualifications.

6 Real-Data Experiments

We examine two questions about borrowing from teacher predictions when target events are scarce. METABRIC asks whether a student fitted to a small target sample can approach a pooled-data reference without accessing source records. Kidney transplantation asks whether borrowing improves prediction across recipient populations. In both applications, held-out outcomes assess prediction and pointwise posterior intervals describe uncertainty conditional on the fixed teacher. These data do not provide ground-truth individual survival probabilities for assessing interval coverage.

6.1 Experimental settings

We use the released METABRIC survival benchmark (Katzman et al., 2018). Three event-stratified partitions each contain 1,446 source-training, 77 target-training and 381 test subjects, with 29, 35 and 31 target-training deaths through 120 months. Teachers are fitted to source-training records. For comparison, a separate pooled analysis fits the same student architecture to source plus target-training records under the reference prior. Its posterior mean provides an estimated reference based on more training data, not ground truth. Neither fit uses test outcomes.

Both applications use MALA to sample all student parameters. The METABRIC student has a 16-unit SiLU hidden layer and $K = 10$ interval-specific hazard logits. The generalized subject-deletion score, estimated with PSIS from target-training data, jointly selects the borrowing weight and prior (Section 3.2). To separate borrowing from changes in prior regularization, the main borrowing path holds the selected prior, which has no temporal smoothing, fixed across all weights, including

$\eta = 0$. Independently prior-selected target-only Bayes and matched-architecture point comparators are reported in Appendix E.

We evaluate survival-probability prediction with the inverse-probability-of-censoring-weighted integrated Brier score (IBS), risk ranking with Antolini's time-dependent concordance, and observed-outcome prediction with deviance, defined as twice the mean negative log predictive density. All methods within a dataset use common evaluation times and target-training censoring estimates. IBS is restricted to follow-up supported by those estimates. On METABRIC, reference RMSE and mean 95% interval width use identical test subjects and time points; the transplant width comparison uses one-year risk. Complete evaluation horizons and additional time-specific widths are given in Appendix E.

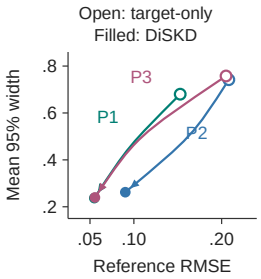
6.2 Prediction and pooled-reference agreement in a small target cohort

PSIS selects $\eta = 30$, the largest evaluated weight, in all three METABRIC partitions. Relative to matched-prior target-only inference, mean IBS decreases from 0.1592 to 0.1263 and mean concordance increases from 0.594 to 0.641 (Figure 3a). These values approach the teacher's IBS of 0.1253 and concordance of 0.650: borrowing largely closes the predictive gap between the small-target fit and the source-trained teacher.

(a) Borrowing path

η	100 IBS ↓	C^{td} ↑
0	15.92 (1.73)	.594 (.025)
0.25	14.60 (1.60)	.604 (.030)
1	13.93 (1.16)	.610 (.036)
3	13.53 (0.82)	.616 (.036)
10	12.97 (0.71)	.632 (.035)
30	12.63 (0.61)	.641 (.032)
Teacher	12.53 (0.55)	.650 (.024)

(b) Reference agreement



(c) Patient-level comparison

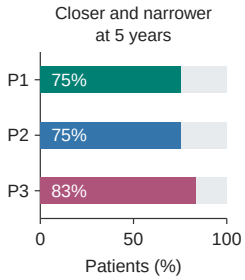


Figure 3: **Prediction, pooled-reference agreement and posterior widths on METABRIC.** All student fits use the PSIS-selected prior without temporal smoothing. (a) Prediction along the borrowing path; entries are mean (SD) across three overlapping partitions. The source teacher is a benchmark, and $\eta = 0$ is the matched-prior target-only fit. (b) Arrows connect target-only (open) to PSIS-selected $\eta = 30$ fits (filled), pairing pooled-reference RMSE with mean 95% interval width on the same subject/time grid. P1–P3 denote partitions; movement left and down means both quantities decrease. (c) Percentage of the 381 test patients in each partition with both smaller absolute five-year reference error and narrower 95% intervals. Gray denotes all other patients. The pooled posterior mean is an estimated reference, not ground truth.

The pooled-reference comparison asks whether posterior narrowing is accompanied by a shift toward the larger-data fit. Reference RMSE and mean interval width both decrease in all three partitions (b). At five years, 75–83% of test patients have both smaller absolute error relative to the pooled posterior mean and narrower 95% intervals (c). Mean paired width reductions are 57.3–61.4 percentage points under this matched-prior comparison (Appendix Table 18). Thus, at this horizon, borrowing brings most patients' predictions closer to the pooled fit, rather than merely producing narrower intervals.

6.3 Borrowing across transplant populations

Unlike METABRIC, the transplant analysis uses source and target cohorts that differ in transplant type. From the Scientific Registry of Transplant Recipients (Scientific Registry of Transplant Recipients, n.d.), we use 35,725 kidney-alone recipients as the source and kidney–pancreas recipients as the target. The target contains 1,935 training recipients with 83 events and 484 test recipients with 21 events within one year. The endpoint is death or kidney graft failure. The student has 16 hidden units, 20 intervals and independent $N(0, 1)$ parameter priors.

(a) Held-out prediction

η	100 IBS ↓	C^{td} ↑	Deviance ↓
0	3.39 (.70)	.614 (.055)	.745 (.140)
0.3	3.34 (.70)	.586 (.057)	.647 (.113)
1	3.31 (.70)	.586 (.059)	.627 (.107)
3	3.27 (.70)	.612 (.063)	.618 (.105)
10	3.27 (.69)	.619 (.057)	.622 (.104)
30	3.26 (.69)	.611 (.068)	.613 (.103)
Teacher	3.27 (.70)	.611 (.067)	.614 (.106)

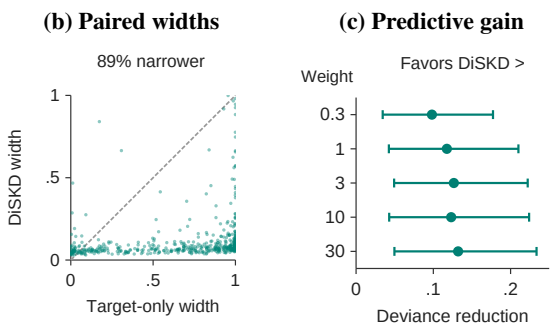


Figure 4: **Borrowing paths from kidney-alone to kidney–pancreas recipients.** (a) Test estimates (paired bootstrap SD) on 484 recipients; deviance is per subject. (b) One-year 95% posterior risk-interval widths at $\eta = 10$ versus $\eta = 0$ for all test recipients. Points below the diagonal have narrower intervals with borrowing. (c) Deviance reductions with pointwise 95% intervals from 2,000 paired test-subject bootstrap resamples conditional on fitted models. Positive values favor borrowing. Additional configurations and patient examples are in Appendices E.3 and E.3.1.

All evaluated positive weights have lower held-out IBS and deviance than $\eta = 0$, although the IBS differences are small (Figure 4a). At PSIS selected $\eta = 10$, deviance decreases from 0.745 to 0.622, while concordance changes from 0.614 to 0.619. Concordance does not consistently improve across weights, so the clearest gain is in predictive deviance rather than risk ranking. At this same fixed weight, one-year 95% posterior risk intervals are narrower for 88.8% of recipients (b); panel (c) reports paired deviance differences across the borrowing path.

Together, these results suggest that cross-population borrowing can improve probabilistic prediction without consistently improving risk ranking. Positive borrowing weights yield better held-out log scores alongside narrower patient-specific risk intervals, indicating that teacher guidance can remain useful across transplant populations. Also, these gains are realized by the PSIS-selected model.

7 Discussion

Two matched comparisons distinguish the effects of teacher guidance and posterior averaging. In the correctly specified Linear control, independent center and covariance approximations reproduce the principal changes in MAP MSE and the two variances under highly accurate estimated guidance (Appendix D.4). In the neural student, the largest observed averaging gains occur for target-only inference and larger teacher discrepancies, while compatible-guidance predictions are close to the matched MAP (Appendix D.5).

The local theory explains how estimation gains simultaneously with center displacement and quantifies a useful borrowing range by balancing squared teacher-induced bias against sampling-variance reduction. Its discrepancy measure links teacher error to survival-probability bias. The Linear checks distinguish posterior spread from repeated-sampling variability. Estimating teacher discrepancy from target data and extending the analysis to adaptive selection remain future directions.

Predictive selection affects how much borrowing benefit is realized. Compatible guidance substantially improves flexible students, while a misspecified linear student retains appreciable error under nonlinear generation. With imperfect teacher, the useful borrowing range depends on teacher discrepancy. Appendix D.3 evaluates selection performance. Fixed-configuration covariance approximations alone do not establish uncertainty properties of the full learning and selection procedure.

The applications highlight complementary roles for prediction-only borrowing: recovering a pooled-data reference in a small breast-cancer cohort and updating predictions across transplant populations. On METABRIC, PSIS-selected borrowing improves prediction and reference agreement relative to independently selected internal Bayes, while narrowing survival-probability intervals. In kidney–pancreas recipients, positive borrowing weights improve deviance and reveal heterogeneous changes in patient-level uncertainty. Frequentist coverage under teacher discrepancy and adaptive selection remains a separate question.

AI USE STATEMENT

Generative AI tools assisted literature review, manuscript revision, mathmatical derivation and figure preparation. The authors assume responsibility for the scientific claims, analyzes, manuscript drafting, and final content of this work. All AI-assisted text and code were reviewed and verified by the authors.

ETHICS STATEMENT

This study uses existing de-identified benchmark and transplant registry data under the applicable data-use and institutional oversight requirements. The analyses evaluate statistical methodology for survival prediction and uncertainty quantification; no models were deployed in clinical care.

REFERENCES

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158.

Victor Chernozhukov and Han Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115:293–346, 2003.

Feiyang Deng, Lingfeng Luo, Di Wang, Qinmengge Li, Lingxuan Kong, and Kevin He. Discrete survival knowledge distillation for competing risks analysis. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=wjsUR2vppX.

A. Eleuteri, M. S. H. Aung, A. F. G. Taktak, B. Damato, and P. J. G. Lisboa. Continuous and discrete time survival analysis: Neural network approaches. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5420–5423, 2007. doi: 10.1109/IEMBS.2007.4353568.

Luyang Fang, Yongkai Chen, Wenxuan Zhong, and Ping Ma. Bayesian knowledge distillation: A bayesian perspective of distillation with uncertainty quantification. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12935–12956. PMLR, 2024. URL https://proceedings.mlr.press/v235/fang24a.html.

Dai Feng and Lili Zhao. BDNNSurv: Bayesian deep neural networks for survival analysis using pseudo values. *Journal of Data Science*, 19(4):542–554, 2021. doi: 10.6339/21-JDS1018. URL https://jds-online.org/journal/JDS/article/1244/info.

Paul Gustafson. Local sensitivity of inferences to prior marginals. *Journal of the American Statistical Association*, 91(434):774–781, 1996. doi: 10.1080/01621459.1996.10476945.

Jimmy Hickey, Jonathan P. Williams, and Emily C. Hector. Transfer learning with uncertainty quantification: Random effect calibration of source to target (RECaST). *Journal of Machine Learning Research*, 25(338):1–40, 2024. URL https://www.jmlr.org/papers/v25/22-1369.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. arXiv:1503.02531.

Joseph G. Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15:46–60, 2000. doi: 10.1214/ss/1009212673.

Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. On optimality properties of the power prior. *Journal of the American Statistical Association*, 98:204–213, 2003. doi: 10.1198/016214503388619229.

Joseph G. Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: Theory and applications. *Statistics in Medicine*, 34:3724–3749, 2015. doi: 10.1002/sim.6728.

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

B. J. K. Kleijn and A. W. van der Vaart. The bernstein–von mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4:4, 2021. doi: 10.1038/s41746-020-00367-3.

Yuying Lu, Tian Gu, and Rui Duan. Adaptive transfer learning for time-to-event modeling with applications in disease risk assessment. *Biostatistics*, 27(1):kxag011, 2026. doi: 10.1093/biostatistics/kxag011.

Aditya K. Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 7632–7642. PMLR, 2021. URL https://proceedings.mlr.press/v139/menon21a.html.

Jeffrey W. Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021. URL https://jmlr.org/papers/v22/20-469.html.

R cchs package. Data from a case-cohort study with stratified subcohort selection. R data documentation, 2026. URL https://search.r-project.org/CRAN/refmans/cchs/html/cchsData.html. Documents the study-label and time-unit mapping from the original nwtco cohort. Accessed September 23, 2026.

R survival package. Data from the national wilm's tumor study. R data documentation, 2026. URL https://stat.ethz.ch/R-manual/R-devel/library/survival/html/nwtco.html. Accessed September 23, 2026.

Richard D. Riley, Gary S. Collins, Laura Kirton, Kym I. E. Snell, Joie Ensor, Rebecca Whittle, Paula Dhiman, Maarten van Smeden, Xiaoxuan Liu, Joseph Alderman, Krishnarajah Nirantharakumar, Joshua Manson-Whitton, Alice J. Westwood, Jean-Baptiste Cazier, Karel G. M. Moons, Glen P. Martin, Matthew Sperrin, Alastair K. Denniston, Frank E. Harrell, and Louise Archer. Uncertainty of risk estimates from clinical prediction models: Rationale, challenges, and approaches. *BMJ*, 388:e080749, 2025. doi: 10.1136/bmj-2024-080749.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1): 255–268, 1998. doi: 10.1111/1467-9868.00123.

Scientific Registry of Transplant Recipients. About SRTR standard analysis files. https://srtr.hrsa.gov/requesting-data/about-srtr-standard-analysis-files-safs/, n.d. Accessed September 16, 2026.

Benjamin A. Shaby. The open-faced sandwich adjustment for MCMC using estimating functions. *Journal of Computational and Graphical Statistics*, 23:853–876, 2014.

Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106 (2):479–486, 2019. doi: 10.1093/biomet/asy054.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27:1413–1432, 2017. doi: 10.1007/s11222-016-9696-4.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024. URL https://www.jmlr.org/papers/v25/19-556.html.

A Posterior Computation and Analysis Specifications

This appendix gives the prior construction, generalized Bayesian interpretation, MALA transition, and calculation of posterior survival summaries used in Section 3.

A.1 Structural and temporal prior

The prior is independent of teacher predictions. Write $P_{\sigma,\infty}$ for the penalty of a proper Gaussian base, with σ controlling the declared parameter-block shrinkage (Appendices D.1.3 and E). Optional temporal smoothing gives $\pi_{\sigma,\tau} \propto \exp(-P_{\sigma,\tau})$, where

$$\begin{aligned} P_{\sigma,\tau}(\theta) &= P_{\sigma,\infty}(\theta) + \frac{r_T(\theta)}{\tau^2}, \\ r_T(\theta) &= \frac{1}{2M}\sum_{j=1}^{M}\sum_{k=2}^{K}\{f_{\theta,k}(x_j^{\text{prior}}) - f_{\theta,k-1}(x_j^{\text{prior}})\}^2. \end{aligned} \tag{9}$$

The reference covariates x_j^{prior} are fixed: the simulations use an independent panel, while the real-data temporal priors condition on the target-training covariate set (Appendix E). Smaller τ favors smoother adjacent logit trajectories; $\tau = \infty$ removes this term. The temporal factor is bounded above by one, so multiplying it by the proper Gaussian base yields a proper prior. Posterior propriety then follows from the bounded likelihood and guidance factors, as detailed below.

For block-specific Gaussian scales, write $P_{\sigma,\tau} = P_{\text{fixed}} + \rho_\sigma G + \rho_\tau r_T$, with $\rho_\sigma = \sigma^{-2}$ and $\rho_\tau = \tau^{-2}$. In the simulations, G penalizes output weights; representation and intercept scales are fixed separately. For a shared-risk student $f_{\theta,k}(x) = a_k + b^\top z_\psi(x)$, $r_T = \|Da\|^2/2$, where D takes adjacent differences. Since $D\mathbf{1} = 0$, temporal smoothing directly constrains time contrasts, not a common calibration offset. Coupling through the fitted likelihood can still change other parameter directions.

Prior-predictive survival scale. For interval-specific heads $f_{\theta,k}(x) = a_k + b_k^\top z_\psi(x)$ with independent zero-centered Gaussian (a_k, b_k), conditioning on $z_\psi(x)$ leaves independent, symmetric logits. Hence $\mathbb{E}_\pi\{1 - \lambda_{\theta,k}(x) \mid \psi\} = 1/2$ and

$$\mathbb{E}_\pi S_\theta(t_K \mid x) = \mathbb{E}_\psi \prod_{k=1}^{K} \mathbb{E}_\pi\{1 - \lambda_{\theta,k}(x) \mid \psi\} = 2^{-K}. \tag{10}$$

This applies to the transplant model's independent Gaussian heads, including intercepts: at $K = 20$ the mean is approximately 9.54×10^{-7}. Changing their zero-centered Gaussian scales alone does not change this mean. Temporal coupling or a shared risk coefficient generally removes the required conditional independence.

An existing prior-predictive pilot makes the scale effect explicit. It uses ten independent $N(0, \sigma^2)$ intercepts at $x = 0$, where covariate effects vanish, and 20,000 joint prior draws per scale (Table 1). Separate two-million-draw checks give mean $S(t_{10})$ of 0.00097669 (MCSE 0.00000138) for $\sigma = 1$ and 0.00097205 (MCSE 0.00001461) for $\sigma = 10$, consistent with $2^{-10} = 0.00097656$. These are checks of prior implications, not fitted posterior performance.

Table 1: **Existing Gaussian prior-predictive check.** Central 95% prior intervals for the first hazard and final survival probability, with $K = 10$ independent intercepts and $x = 0$. The 20,000 draws per row concern a simple prior-only pilot, not fitted transplant or neural posteriors.

σ	Prior interval for λ_1	Prior interval for $S(t_{10})$
0.25	$[0.378, 0.620]$	$[4.04 \times 10^{-4}, 1.91 \times 10^{-3}]$
1	$[0.119, 0.876]$	$[9.25 \times 10^{-6}, 5.88 \times 10^{-3}]$
3	$[0.002, 0.997]$	$[6.98 \times 10^{-12}, 6.62 \times 10^{-3}]$
10	$[2.10 \times 10^{-9},\ 1 - 3.32 \times 10^{-9}]$	$[5.66 \times 10^{-36}, 6.47 \times 10^{-5}]$

The survival-scale implications motivate checking the prior on the prediction scale, but do not establish posterior error or attribute wide fitted intervals entirely to the prior. The paired Simulation II

sensitivity in Appendix D.9 varies Gaussian shrinkage and temporal smoothing with the same prior on both borrowing arms. It does not recenter the prior to an event rate; a paired full-network comparison using such priors in transplantation is not part of the completed evidence.

A.2 Generalized Bayesian update and soft targets

Let $\mathcal{L}_\eta(\theta) = \sum_i \{r_i(\theta) + \eta q_i(\theta)\}$ and $Z = \int e^{-\mathcal{L}_\eta(\theta)} \pi_{\sigma,\tau}(\theta)\, d\theta$. For a probability distribution ν absolutely continuous with respect to the proper prior, whenever the terms are finite,

$$\mathbb{E}_\nu \mathcal{L}_\eta + D_{\mathrm{KL}}(\nu \| \pi_{\sigma,\tau}) = D_{\mathrm{KL}}(\nu \| \Pi_{\eta,\sigma,\tau}) - \log Z. \tag{11}$$

The update in Equation (4) therefore minimizes the left side, giving a loss-based Bayesian interpretation (Bissiri et al., 2016). This construction conditions on the supplied teacher predictions and the declared prior reference covariates.

To verify Equation (3), write $h(u) = -u \log u - (1-u)\log(1-u)$. Bernoulli KL equals $c(u,p) - h(u)$, so the teacher-dependent constant is $C_i = -\eta \sum_k A_{ik} h(\widetilde{\lambda}_{ik})$. The teacher entropy does not affect the posterior over θ. The factor $1+\eta$ multiplies the cross-entropy and remains part of the posterior energy; dividing it out would change both concentration and the prior's relative influence. The equivalent fractional event targets explain the contribution of teacher predictions without treating them as independent observations. Convexity in the logits does not imply convexity in all neural parameters; Appendix A.7 gives the affine-head special case.

A.3 Sampled state and exact energy

The risk-set likelihood corresponding to Equation (2) is

$$L_{\mathrm{tar}}(\theta; D) = \prod_i L_i(\theta), \qquad L_i(\theta) = \prod_k \lambda_{\theta,k}(X_i)^{A_{ik}\delta_i(t_k)} \{1 - \lambda_{\theta,k}(X_i)\}^{A_{ik}\{1-\delta_i(t_k)\}}. \tag{12}$$

The negative log target is

$$U_{\eta,\sigma,\tau}(\theta) = \sum_i \{r_i(\theta) + \eta q_i(\theta)\} + P_{\sigma,\tau}(\theta). \tag{13}$$

Write $R_n = n^{-1}\sum_i r_i$, $Q_n = n^{-1}\sum_i q_i$, and $M_{n,\eta} = R_n + \eta Q_n$. The posterior mode $\widehat{\theta}$ minimizes $M_{n,\eta}(\theta) + P_{\sigma,\tau}(\theta)/n$. The isotropic Gaussian special case is $P_{\sigma,\infty} = \|\theta\|^2/(2\sigma^2)$, giving $\lambda_n = (n\sigma^2)^{-1}$. Thus the MAP fit is a penalized distillation estimator, whereas posterior averaging defines the generalized Bayesian predictor.

The joint neural state includes representation and prediction-head parameters. The main simulations and real-data analyses use this joint state. The alternative in Appendix A.7 freezes a fitted representation; the retained conditional-head diagnostics explicitly identify that procedure.

Teacher outputs, risk-set masks, the time grid, and (η, σ, τ) remain fixed throughout each run.

The Gaussian prior contributes once to the summed energy in Equation (13). Equivalently, if an implementation evaluates mean losses, it must multiply their sum by n before combining it with the prior energy. With a proper Gaussian prior and finite logits, the Bernoulli likelihood is positive and at most one, and $\exp(-n\eta Q_n) \in (0, 1]$. The normalizing constant of Equation (4) is therefore finite and positive. A formal flat prior establishes the MAP objective connection but does not by itself establish posterior propriety.

For $a_{ik} = f_{\theta,k}(X_i)$ and $p_{ik} = \mathrm{expit}(a_{ik})$, differentiation of the untempered losses gives

$$\nabla U_{\eta,\sigma,\tau}(\theta) = \sum_{i,k} A_{ik}\big[(p_{ik} - \delta_i(t_k)) + \eta(p_{ik} - \widetilde{\lambda}_{ik})\big]\nabla a_{ik} + \nabla P_{\sigma,\tau}(\theta). \tag{14}$$

This identity applies to the student parameterization and is useful for checking automatic differentiation and loss scaling. The teacher is not differentiated during student inference.

A.4 Proposal and acceptance probability

Let B be symmetric positive definite and fixed during retained sampling. Define

$$\mu_\epsilon(\theta) = \theta - \frac{\epsilon^2}{2} B \nabla U_{\eta,\sigma,\tau}(\theta), \qquad q_\epsilon(\theta' \mid \theta) = \mathcal{N}\{\theta'; \mu_\epsilon(\theta), \epsilon^2 B\}. \tag{15}$$

For a proposed θ', accept with probability

$$a(\theta, \theta') = 1 \wedge \exp[-U_{\eta,\sigma,\tau}(\theta') + U_{\eta,\sigma,\tau}(\theta) + \log q_\epsilon(\theta \mid \theta') - \log q_\epsilon(\theta' \mid \theta)]\,. \tag{16}$$

Otherwise retain θ as the next state. Rejected transitions are part of the Markov chain and are not deleted from its posterior summary. Any adaptation of ϵ or B is restricted to warm-up and frozen before retained transitions. Identity preconditioning is obtained with $B = I$.

For $p(\theta) \propto \exp\{-U_{\eta,\sigma,\tau}(\theta)\}$, the accepted flow satisfies

$$p(\theta) q_\epsilon(\theta' \mid \theta) a(\theta, \theta') = \min\{p(\theta) q_\epsilon(\theta' \mid \theta), p(\theta') q_\epsilon(\theta \mid \theta')\}.$$

The symmetry of this expression gives detailed balance with the designated generalized posterior; the rejection probability supplies the remaining diagonal part of the transition. This establishes invariance of the target distribution by the standard Metropolis–Hastings argument (Roberts & Rosenthal, 1998).

Full-data energies and proposal gradients are evaluated consistently at both states. Exact accumulation of subject batches is compatible with this construction; substituting an uncorrected stochastic minibatch energy into Equation (16) is not the same transition. State-dependent preconditioning would also require its own correctly specified proposal, rather than the fixed-B density above.

A.5 Prediction and numerical assessment

For a survival functional ϕ, posterior draws estimate its mean and pointwise quantiles by

$$\widehat{\phi}_B = \mathbb{E}_\Pi \phi, \qquad \widehat{\phi}_B^{(M)} = M^{-1} \sum_m \phi(\theta^{(m)}), \qquad \widehat{I}_{1-\gamma}(\phi) = [\widehat{q}_{\gamma/2}, \widehat{q}_{1-\gamma/2}], \tag{17}$$

where $\widehat{q}_a$ is the empirical a-quantile of the functional draws. These intervals concern a survival probability, not an individual event time or a simultaneous survival band.

For each retained state, compute the complete vector of interval hazards and then the survival curve in Equation (1). Average those curves for point prediction and use their marginal quantiles for pointwise credible intervals. In general,

$$\frac{1}{M} \sum_m \prod_{\ell \le k} (1 - \lambda_{\theta^{(m)},\ell}) \neq \prod_{\ell \le k} \left(1 - \frac{1}{M} \sum_m \lambda_{\theta^{(m)},\ell}\right).$$

Consequently, separately averaging parameters or hazards before constructing the survival curve is not the posterior-mean predictor. Likewise, pointwise intervals across horizons are not automatically a simultaneous band.

In the discrete simulations, functional diagnostics evaluate all ten horizons for 12 fixed reference subjects. We retain functional $\widehat{R}$, functional effective sample sizes, posterior-mean Monte Carlo errors, and parameter diagnostics, separately from PSIS importance diagnostics. All fits remain in the experimental accounting. Interval-endpoint precision and stability of selection contrasts require assessment separately from posterior-mean MCSE.

Numerical convergence assesses approximation of the posterior for fixed data. Repeated-cohort evaluation assesses the statistical behavior of the fitted estimator and intervals. Sandwich-adjusted intervals use the latter covariance target.

A.6 Subject-level sandwich intervals

Raw generalized-posterior spread need not match the repeated-cohort variability of its fitted center. We use a subject-level sandwich estimate to obtain an alternative, sampling-based uncertainty

summary (Shaby, 2014; Miller, 2021). At fixed (η, σ, τ), let θ_* minimize $R + \eta Q + P_{\sigma,\tau}/n$. The sampling target is $\phi(\theta_*)$, which need not equal the true survival probability. At the empirical MAP $\widehat{\theta}$, define

$$\widehat{A} = \nabla^2\{R_n + \eta Q_n + P_{\sigma,\tau}/n\}(\widehat{\theta}),$$
$$s_i = \nabla r_i(\widehat{\theta}) + \eta \nabla q_i(\widehat{\theta}), \qquad \widehat{J} = \frac{1}{n-1}\sum_i (s_i - \overline{s})(s_i - \overline{s})^\top. \tag{18}$$

Here $\overline{s} = n^{-1}\sum_i s_i$, and derivatives use the sampled coordinates (w for a fixed head). Each s_i includes all intervals for one patient, preserving both within-patient dependence and the covariance between likelihood and teacher scores. The prior enters $\widehat{A}$ as deterministic curvature, not as additional random observations. When the local Hessian is positive definite, the curvature and sandwich covariance estimates are

$$\widehat{\Sigma}_{\mathrm{curv}} = \frac{\widehat{A}^{-1}}{n}, \qquad \widehat{\Sigma}_{\mathrm{sand}} = \frac{\widehat{A}^{-1}\widehat{J}\widehat{A}^{-1}}{n}. \tag{19}$$

For $\phi(\theta) = S_\theta(t_k \mid x)$ and $\widehat{g} = \nabla\phi(\widehat{\theta})$, set $\widehat{V}_{\mathrm{sand},\phi} = \widehat{g}^\top \widehat{\Sigma}_{\mathrm{sand}} \widehat{g}$. The adjusted interval is

$$I_{\mathrm{sand},1-\gamma}(\phi) = \left[\phi(\widehat{\theta}) - z_{1-\gamma/2}\sqrt{\widehat{V}_{\mathrm{sand},\phi}},\ \phi(\widehat{\theta}) + z_{1-\gamma/2}\sqrt{\widehat{V}_{\mathrm{sand},\phi}}\right] \cap [0,1], \tag{20}$$

where $z_{1-\gamma/2}$ is a standard-normal quantile. We compare this with the same MAP-centered Wald construction using $\widehat{\Sigma}_{\mathrm{curv}}$, and keep raw posterior quantile intervals separate. The adjustment changes the spread estimate, not the generalized posterior or its point predictor.

The local regularity conditions and oracle specialization are given in Section 4. For head-only use, the repeated-cohort experiment must hold an independently obtained representation fixed. Relearning it from target outcomes requires analysis of the complete fitting procedure rather than the head score alone.

A.7 Conditional-head inference

Write $\theta = (\psi, w)$ for representation and prediction-head parameters, including intercepts. If a representation $\widehat{\psi}$ is supplied and held fixed, a lower-dimensional alternative is

$$\Pi_{\eta,\sigma,\tau}(dw \mid D, \widetilde{P}, \widehat{\psi}) \propto \exp\left[-\sum_i \{r_i(\widehat{\psi}, w) + \eta q_i(\widehat{\psi}, w)\} - P_{\sigma,\tau}(\widehat{\psi}, w)\right] dw. \tag{21}$$

It quantifies uncertainty conditional on the supplied representation and is not the marginal of the joint posterior. The retained conditional-head diagnostics concern this alternative, with the fitted representation held fixed during sampling. The primary simulations and real-data analyses sample both blocks.

For an affine head $f_{\widehat{\psi},w,k}(X_i) = z_{ik}^\top w$, a proper Gaussian base and the quadratic temporal penalty give positive-definite precision P. The conditional negative log posterior has Hessian

$$\nabla_w^2 U = (1+\eta)\sum_{i,k} A_{ik}\lambda_{w,ik}(1-\lambda_{w,ik}) z_{ik} z_{ik}^\top + P \succ 0.$$

Hence the conditional MAP is unique. Teacher probabilities enter the score but not the Bernoulli cell curvature at a common evaluation point. This geometry facilitates computation; it does not by itself establish a Monte Carlo accuracy or truth-coverage guarantee.

The functional sensitivity and covariance results apply in w coordinates under the same local regularity conditions, holding $\widehat{\psi}$ fixed. In particular, with A_w the conditional average-loss Hessian including prior curvature and J_w the complete patient-score covariance,

$$V_{\mathrm{post},\phi} \approx n^{-1} g_w^\top A_w^{-1} g_w, \qquad V_{\mathrm{samp},\phi} \approx n^{-1} g_w^\top A_w^{-1} J_w A_w^{-1} g_w.$$

The second expression matches a repeated-cohort procedure only when the same independently obtained representation is held fixed. Learning it anew from target outcomes adds variation outside this conditional calculation. Similarly, conditional case deletion holds $\widehat{\psi}$ fixed; it does not validate representation learning on the deleted outcome. The posterior-mean sensitivity identity remains valid with integration over w, provided the representation is constant during differentiation.

B THEORETICAL RESULTS AND PROOFS

This appendix collects the conditions, statements, and proofs supporting Section 4. All results concern the declared inference state and fixed hyperparameters.

B.1 CONDITIONS FOR THE LOCAL ESTIMATOR AND POSTERIOR STATEMENTS

The local results concern a fixed teacher mapping, fixed time grid, fixed η, and a fixed-dimensional identifiable coordinate system; they do not assert global Gaussian behavior of an unrestricted neural network. Define $R = \mathbb{E}r_i$, $Q = \mathbb{E}q_i$, $m_{i,\eta} = r_i + \eta q_i$, $M_\eta = R + \eta Q$, $\theta_\eta \in \arg\min_\theta M_\eta(\theta)$, and $b_\eta = \theta_\eta - \theta_0$, where θ_0 minimizes the target risk R. Let $P_n = P_{\sigma_n,\tau_n}$ be a deterministic, twice continuously differentiable prior penalty, conditional on its fixed reference panel, and set

$$\theta_n = \arg\min_\theta\{M_\eta(\theta) + P_n(\theta)/n\}, \qquad \widehat{\theta}_n \in \arg\min_\theta\{n^{-1}\textstyle\sum_i m_{i,\eta}(\theta) + P_n(\theta)/n\}.$$

Here and below a locally unique minimizing branch is used. The isotropic special case has $P_n/n = \lambda_n\|\theta\|^2/2$, with $\lambda_n = (n\sigma_n^2)^{-1}$. The following are sufficient conditions for the stated local conclusions.

1. The penalized population criterion is smooth near its interior minimizer, $\theta_n \to \bar{\theta}$, and the empirical minimizing branch is consistent. The empirical Hessian converges uniformly on a neighborhood to its population counterpart. When P_n/n and its first two derivatives vanish locally, $\bar{\theta} = \theta_\eta$.
2. The centered subject scores obey a central limit theorem at θ_n, the required stochastic equicontinuity holds, and
$$H_n = \nabla^2 M_\eta(\theta_n) + n^{-1}\nabla^2 P_n(\theta_n) \longrightarrow H \succ 0, \qquad J_n = \mathrm{Var}\{\nabla m_{i,\eta}(\theta_n)\} \longrightarrow J. \tag{22}$$
Subject-level covariance estimates and curvature estimates are consistent for these limits.
3. The generalized posterior concentrates on shrinking neighborhoods of $\widehat{\theta}_n$, admits a uniform local quadratic expansion there, and its rescaled law for $\sqrt{n}(\theta - \widehat{\theta}_n)$ converges to $N(0, H^{-1})$ with sufficient tail control. Convergence of posterior means and covariance additionally requires uniform integrability of the corresponding first and second rescaled moments. These moment conditions are not inferred from weak convergence alone.
4. A reported functional is twice continuously differentiable near the limiting target, with controlled Taylor remainder under that posterior concentration. Its projected sampling variance is positive when a standardized interval or variance ratio is used. The pointwise coverage corollary uses asymptotically Gaussian equal-tail quantiles.

These conditions state the domain of the approximation (van der Vaart, 1998; Chernozhukov & Hong, 2003; Miller, 2021). They must be checked for the estimating problem being analyzed. They need not hold uniformly when event information vanishes, the dimension grows, or neural parameter symmetries remain. Fixed-hyperparameter theory also does not automatically cover selection on the same target cohort.

Scope of the uncertainty statements. A representation trained independently and frozen across cohorts defines a sampling experiment consistent with conditional-head inference. Relearning it in each cohort changes that procedure; freezing a representation learned from the same cohort does not itself justify a head-only subject-score sandwich. Likewise, the covariance results hold at fixed hyperparameters, whereas repeated selection changes the sampling distribution. Covariance describes spread around the population center; its displacement from truth can separately affect coverage.

B.2 POSTERIOR-MEAN FUNCTIONALS: THE LOCAL-MODE BRIDGE

The point predictor is $\bar{\phi}_n = \int \phi(\theta)\Pi_n(d\theta)$, rather than $\phi(\int \theta\Pi_n(d\theta))$. A local mode supplies an estimating equation; the following result transfers its first-order sampling law to the posterior-mean functional. It is a specialization of the usual moment-controlled local posterior argument (Chernozhukov & Hong, 2003; Miller, 2021).

Proposition 1 (First-order equivalence for a posterior-mean functional)**.** *Let $\widehat{\theta}_n \xrightarrow{p} \bar{\theta}$ be the local mode and $U_n = \sqrt{n}(\theta - \widehat{\theta}_n)$. Conditional on the data, suppose the posterior law of U_n converges weakly in probability to $N(0, A^{-1})$, $A \succ 0$, and its first absolute moments are uniformly integrable in probability. For $g_n = \nabla\phi(\widehat{\theta}_n) \xrightarrow{p} g$, assume*

$$\sqrt{n} \int \left| \phi(\theta) - \phi(\widehat{\theta}_n) - g_n^\top (\theta - \widehat{\theta}_n) \right| \Pi_n(d\theta) \xrightarrow{p} 0.$$

Then

$$\sqrt{n}\{\bar{\phi}_n - \phi(\widehat{\theta}_n)\} \xrightarrow{p} 0. \tag{23}$$

Consequently, every stated root-n sampling limit for $\phi(\widehat{\theta}_n)$ also holds for $\bar{\phi}_n$. If, for a deterministic target $\phi_{0,n}$, the squared errors $n(\bar{\phi}_n - \phi_{0,n})^2$ are uniformly integrable under repeated sampling, their expectations converge to the second moment of that limit. The analogous condition for the mode gives the same leading MSE for both estimators.

The remainder condition follows, for example, from a bounded local functional Hessian, posterior second moments $\mathbb{E}_\Pi \|U_n\|^2 = O_p(1)$, and negligible posterior contribution to the remainder outside that neighborhood. Posterior moment control and repeated-sampling moment control concern different probability distributions. For variance convergence within the posterior, the corresponding second-moment control and an L^2 version of the functional remainder are used. These are additional conditions to weak posterior convergence.

Proof. Weak convergence and uniform integrability of first posterior moments give $\mathbb{E}_{\Pi_n} U_n \xrightarrow{p} 0$. Integrating the Taylor expansion yields

$$\sqrt{n}\{\bar{\phi}_n - \phi(\widehat{\theta}_n)\} = g_n^\top \mathbb{E}_{\Pi_n} U_n + o_p(1) = o_p(1).$$

Slutsky's theorem transfers the distributional limit. Repeated-sampling uniform integrability of the squared errors then transfers their second moments. A sufficient alternative for equality of leading mean squared errors is $n\mathbb{E}_{D_n}\{\bar{\phi}_n - \phi(\widehat{\theta}_n)\}^2 \to 0$ together with bounded scaled mode MSE, by Cauchy–Schwarz. No $O_p(n^{-1})$ rate or finite-sample equality is asserted. □

All conclusions concern the exact posterior mean. A Monte Carlo estimate has the same first-order law only when its computational error is $o_p(n^{-1/2})$. Fixed-n comparisons in Appendix D.5 assess the size of the observed differences, rather than verifying an asymptotic rate.

B.3 Teacher-Induced Curvature and Conditional Sampling Variability

At the local population center of $R + \eta Q + P_n/n$, define $H_R = \mathbb{E}\nabla^2 r_i$, $H_Q = \mathbb{E}\nabla^2 q_i$, $J_R = \mathrm{Var}(\nabla r_i)$, $J_Q = \mathrm{Var}(\nabla q_i)$, and $C_{RQ} = \mathrm{Cov}(\nabla r_i, \nabla q_i)$. Then

$$\begin{aligned} A_{\eta,n} &= H_R + \eta H_Q + n^{-1}\nabla^2 P_n, \\ J_\eta &= J_R + \eta(C_{RQ} + C_{RQ}^\top) + \eta^2 J_Q. \end{aligned} \tag{24}$$

The prior contributes deterministic curvature and affects score covariance through the evaluation point. For interval-specific neural logits, the temporal penalty depends on the representation; its full Hessian, rather than a proposal-metric approximation or frozen-head block, enters $A_{\eta,n}$.

Proposition 2 (Local posterior and sampling variability)**.** *Fix the teacher and a finite-dimensional identifiable sampled state. Assume consistency of the local empirical center, a centered subject-score CLT, a uniform Hessian law with nonsingular limiting A, and posterior concentration and quadratic approximation with the moment control needed for posterior means and covariances. With $P_n = P_{\sigma_n,\tau_n}$ and centers and matrices evaluated along the corresponding deterministic sequence, the empirical center, centered at its population counterpart, has a root-n Gaussian limit with covariance $A^{-1}JA^{-1}$, whereas the local posterior covariance is $n^{-1}A^{-1}$. For a smooth functional with gradient g, the posterior and leading asymptotic sampling variances are*

$$\begin{aligned} V_{\mathrm{post},\phi} &\simeq n^{-1} g^\top A_{\eta,n}^{-1} g, \\ V_{\mathrm{samp},\phi} &\simeq n^{-1} g^\top A_{\eta,n}^{-1} J_\eta A_{\eta,n}^{-1} g. \end{aligned} \tag{25}$$

Under Proposition 1, this is also the sampling limit of the posterior-mean functional $\bar{\phi}_n$, centered at $\phi(\theta_n)$.

Generalized-posterior asymptotics and sandwich inference supply these approximations (Miller, 2021; Shaby, 2014). Estimate J by the centered covariance of complete patient scores $s_i = \nabla r_i + \eta \nabla q_i$, preserving within-patient dependence and the symmetric cross term. The sandwich estimate is $\widehat{A}^{-1}\widehat{J}\widehat{A}^{-1}/n$. Actual variance convergence additionally requires repeated-sampling moment control.

Proposition 3 (Well-specified oracle guidance)**.** *In addition to the preceding conditions, suppose the student is correctly specified, the fixed teacher equals the true conditional hazards, the observed-data likelihood and risk sets are correct under conditionally independent censoring, η is fixed and finite, and the prior satisfies $n^{-1/2}\nabla P_n(\theta_0) \to 0$ and $n^{-1}\nabla^2 P_n \to 0$ locally uniformly. At the true parameter θ_0, $\nabla q_i(\theta_0) = 0$, $H_Q \succeq 0$, and $J_\eta = J_R = H_R$. Writing $A_\eta = H_R + \eta H_Q$, the leading covariance terms satisfy*

$$\Sigma_{\text{post}} - \Sigma_{\text{samp}} \simeq \frac{\eta}{n} A_\eta^{-1} H_Q A_\eta^{-1} \succeq 0. \tag{26}$$

For the same Bernoulli risk-set contributions, $H_Q = H_R$ and $V_{\text{post},\phi}/V_{\text{samp},\phi} \simeq 1 + \eta$ for a functional with nonzero leading variance.

Both spreads decrease in this benchmark, with sampling spread decreasing faster; the associated SD ratio is $\sqrt{1+\eta}$. It supplies a conditional reference, rather than a universal neural interval correction.

Under the conditions of the following corollary, the posterior-mean interval satisfies

$$I_{\text{adj},\eta} = \bar{\phi}_{\eta,n} \pm z_{1-\gamma/2}\sqrt{\frac{\widehat{V}_{\text{post},\phi}}{1+\eta}}, \qquad \Pr_{D_n}\{\phi_0 \in I_{\text{adj},\eta}\} \longrightarrow 1-\gamma. \tag{27}$$

Corollary 4 (Pointwise oracle interval calibration)**.** *Assume Proposition 3 with identical Bernoulli risk-set contributions, a fixed finite $\eta \geq 0$, and a fixed functional with $v_\phi = g^\top H^{-1} g > 0$. Assume the functional Bernstein–von Mises approximation and the posterior-mean bridge of Proposition 1, and let $\widehat{V}_{\text{post},\phi}$ satisfy $n\widehat{V}_{\text{post},\phi} \xrightarrow{p} v_\phi/(1+\eta)$; using posterior variance requires the corresponding posterior moment control. Then the raw equal-tail posterior interval has limiting coverage*

$$C_{\text{raw},\eta} = 2\Phi\{z_{1-\gamma/2}\sqrt{1+\eta}\} - 1,$$

whereas $I_{\text{adj},\eta}$ in Equation (27) *has limiting coverage $1-\gamma$. The same conclusion holds with any variance estimate $\widehat{V}_{\text{samp},\phi}$ satisfying $n\widehat{V}_{\text{samp},\phi} \xrightarrow{p} v_\phi/(1+\eta)^2$, including a consistent subject-level sandwich. The ratio of adjusted interval widths at weights η and zero converges in probability to $1/(1+\eta)$. Equation* (27) *uses the posterior-mean functional as its center; centering at the local-mode functional has the same first-order limit.*

This is pointwise asymptotic coverage for a survival probability at a fixed subject and horizon. Simultaneous survival bands, growing borrowing weights, and data-selected configurations require separate analysis. The variance convention $V_{\text{adj}} = V_{\text{post}}/\omega$ gives $\omega = 1+\eta$ in this oracle setting; it does not supply a calibration rule for unknown teacher bias.

Corollary 5 (Local teacher discrepancy, prediction risk, and spread)**.** *Assume the correctly specified, identical-cell oracle setting of Proposition 3, and its regularity conditions uniformly for deterministic teachers $\widetilde{p}_{n,k}(x) = \operatorname{expit}\{\operatorname{logit} p_{0,k}(x) + h d_k(x)/\sqrt{n}\}$, where h and η are fixed and finite. Assume differentiability of the population score in this perturbation, the centered subject-score CLT along the sequence, and convergence of the perturbed teacher score to zero in second-moment norm at the convergent population centers. Write $H = H_R(\theta_0)$, $B = \partial_\varepsilon \nabla Q(\theta_0; 0)$, $g = \nabla\phi(\theta_0)$, $c_\phi = -g^\top H^{-1} B$, and $v_\phi = g^\top H^{-1} g > 0$. Under Proposition 1 along this sequence, the posterior-mean functional $\bar{\phi}_{\eta,n} = \mathbb{E}_{\Pi_{\eta,n}}\phi(\theta)$ satisfies*

$$\sqrt{n}(\bar{\phi}_{\eta,n} - \phi_0) \Rightarrow N\left(\frac{h\eta c_\phi}{1+\eta}, \frac{v_\phi}{(1+\eta)^2}\right), \qquad nV_{\text{post},\phi} \longrightarrow \frac{v_\phi}{1+\eta}. \tag{28}$$

If $n(\bar{\phi}_{\eta,n} - \phi_0)^2$ is uniformly integrable under repeated sampling, Equation (7) *follows. The local-mode functional has the same distributional limit and, under its corresponding sampling moment condition, the same leading risk.*

For $hc_\phi \neq 0$, the limiting risk expression over nonnegative finite weights is minimized at $\eta^* = v_\phi/(h^2 c_\phi^2)$; if $hc_\phi = 0$, it decreases with η. Subtracting the target-only risk yields

$$\mathcal{R}_\phi(\eta) - \mathcal{R}_\phi(0) = \frac{v_\phi \eta\{\eta(\kappa_\phi - 1) - 2\}}{(1+\eta)^2}, \qquad \kappa_\phi = \frac{h^2 c_\phi^2}{v_\phi},$$

which proves Equation (8). For multiple functionals with prespecified nonnegative weights w_j, the same comparison uses $\sum_j w_j v_j$ and $\sum_j w_j h^2 c_j^2$. Cauchy–Schwarz in the information metric also gives

$$c_\phi^2 \le (g^\top H^{-1} g)(B^\top H^{-1} B), \qquad \kappa_\phi \le h^2 B^\top H^{-1} B.$$

A specified bound on this curvature-weighted discrepancy therefore supplies a sufficient local risk bound across the reported functionals. These comparisons require the declared discrepancy direction and magnitude; they do not identify them from target data. The perturbation is a local asymptotic sequence. Fixed nonzero teacher bias can persist as n grows, and a general fixed-discrepancy path with finite prior or model misspecification need not obey the closed form. The finite-prior approximation in Appendix D.4 addresses the fitted linear benchmark separately.

B.4 Teacher Guidance and Student-Functional Sensitivity

Let $R = \mathbb{E} r_i$ and $Q(\theta;\varepsilon) = \mathbb{E} q_i(\theta;\varepsilon)$ over target subjects. Perturb baseline teacher hazards by $\widetilde{p}_{\varepsilon,k}(x) = \operatorname{expit}\{\operatorname{logit} \widetilde{p}_{0,k}(x) + \varepsilon d_k(x)\}$. A constant direction shifts teacher hazard logits; a covariate-dependent direction changes risk effects. The population center θ_* is a local minimizer of the expected penalized objective.

Proposition 6 (Local functional response at fixed borrowing)**.** *Fix finite $\eta \ge 0$, n, and a smooth prior penalty $P_{\sigma,\tau}$ independent of ε. Suppose $\mathcal{M}(\theta,\varepsilon) = R(\theta) + \eta Q(\theta;\varepsilon) + P_{\sigma,\tau}(\theta)/n$ has a locally unique interior stationary branch $\theta_*(\varepsilon)$ near zero. Assume its gradient is continuously differentiable in (θ,ε), differentiation commutes with expectation, and $A = \nabla_\theta^2 \mathcal{M}(\theta_*(0),0) \succ 0$. For a continuously differentiable functional ϕ, set $g = \nabla\phi(\theta_*(0))$ and $B = \partial_\varepsilon \nabla_\theta Q(\theta_*(0);0)$. Then*

$$\phi\{\theta_*(\varepsilon)\} - \phi\{\theta_*(0)\} = -\varepsilon\eta\, g^\top A^{-1} B + o(\varepsilon). \tag{29}$$

The same differentiation gives local responses to borrowing and prior precision (Appendix B.8).

For the empirical MAP the same implicit differentiation applies with R_n, Q_n and the unchanged penalty $P_{\sigma,\tau}/n$. It covers block Gaussian and temporal priors, using their actual gradients and Hessians. For $\phi_{x,k}(\theta) = S_\theta(t_k \mid x)$, patient/horizon gradients g yield different responses to the same teacher change. These are local derivatives, not monotonicity statements along an entire borrowing path.

Corollary 7 (Posterior-mean response to teacher perturbation)**.** *For fixed data and prior, let $\Pi_\varepsilon(d\theta) \propto e^{-U_\varepsilon(\theta)} d\theta$ have common support and finite normalizing constant near the point of differentiation. Suppose the integrands defining the normalizer and $\mathbb{E}_{\Pi_\varepsilon}\phi$ and their derivatives admit integrable local envelopes, and the covariance below is finite. For a functional with no explicit ε dependence,*

$$\begin{aligned} \frac{d}{d\varepsilon}\mathbb{E}_{\Pi_\varepsilon}\phi &= -\operatorname{Cov}_{\Pi_\varepsilon}(\phi, \partial_\varepsilon U_\varepsilon) \\ &= -n\eta \operatorname{Cov}_{\Pi_\varepsilon}\{\phi, \partial_\varepsilon Q_n(\theta;\varepsilon)\}. \end{aligned} \tag{30}$$

This specializes the posterior-sensitivity identity of Gustafson (1996); it requires neither an oracle teacher nor a Gaussian approximation. Conditional-head differentiation and perturbed refitting hold $\widehat{\psi}$ fixed. Appendix B.7 gives the complementary vanishing-borrowing expansion.

For a scalar hyperparameter $a \in \{\eta, \sigma, \tau\}$, the same identity applies with $U_a = \sum_i (r_i + \eta q_i) + P_{\sigma,\tau}$. Under the analogous differentiability and moment conditions for ϕ^2, writing $\mu_a = \mathbb{E}_{\Pi_a}\phi$ gives

$$\partial_a \mathbb{E}_{\Pi_a}\phi = -\operatorname{Cov}_{\Pi_a}(\phi, \partial_a U_a), \qquad \partial_a \operatorname{Var}_{\Pi_a}(\phi) = -\operatorname{Cov}_{\Pi_a}\{(\phi-\mu_a)^2, \partial_a U_a\}. \tag{31}$$

The variance identity follows by differentiating $\mathbb{E}\phi^2 - \mu_a^2$. These responses quantify how an existing posterior changes. Their signs do not identify movement toward truth: that comparison additionally requires the discrepancy and sampling-noise terms in Section 4.2.

B.5 Two paths through the same fitted center

Write $\theta_*(\eta,\varepsilon)$ for the local minimum of

$$R(\theta) + \eta Q(\theta;\varepsilon) + P_{\sigma,\tau}(\theta)/n.$$

Under the smooth-branch and positive-curvature conditions of Proposition 6, implicit differentiation gives

$$\frac{\partial\phi(\theta_*)}{\partial\eta} = -g^\top A^{-1}\nabla Q, \qquad \frac{\partial\phi(\theta_*)}{\partial\varepsilon} = -\eta g^\top A^{-1}B, \tag{32}$$

where $A = \nabla^2\{R + \eta Q + P_{\sigma,\tau}/n\}$, $g = \nabla\phi$, and $B = \partial_\varepsilon \nabla Q$ are evaluated at the current center. Both Hessian and loss derivatives use average-loss scale; the implementation solves linear systems. Along a smooth path $(\eta(s), \varepsilon(s))$, the derivative is

$$\frac{d\phi(\theta_*)}{ds} = -g^\top A^{-1}\{\nabla Q\, \eta'(s) + \eta B\, \varepsilon'(s)\}.$$

Thus borrowing strength and teacher discrepancy move the same survival functional through different score directions. Their signs need not agree for every patient and horizon. These identities retain finite prior effects and do not require the oracle relation $A = (1+\eta)H$.

At fixed η, integrating the derivative with respect to ε along an admissible smooth branch gives the finite teacher-induced change. If that derivative is bounded between L and U throughout $[0, \varepsilon]$, then the change lies in $[\varepsilon L, \varepsilon U]$ for $\varepsilon \geq 0$. This is a conditional range statement: derivatives at a few evaluated points do not certify such uniform bounds. Endpoint and intermediate-point calculations can assess direction and nonlinearity, but do not by themselves establish a continuous bias bound.

The change is an increment relative to the unperturbed teacher, not the total truth bias:

$$\begin{aligned}\phi\{\theta_*(\eta,\varepsilon)\} - S_0(t_k \mid x) = &\big[\phi\{\theta_*(\eta,0)\} - S_0(t_k \mid x)\big] \\ &+ \big[\phi\{\theta_*(\eta,\varepsilon)\} - \phi\{\theta_*(\eta,0)\}\big].\end{aligned} \tag{33}$$

The first term retains baseline approximation, prior, and estimated-teacher effects. Repeated-cohort estimation adds sampling fluctuation around the corresponding population center. A systematic change in the second term is not an additional independent random variance.

B.6 Local Gaussian interpretation of interval coverage

Let $g = \nabla\phi(\bar\theta)$ at the limiting penalized center in Appendix B.1, and define

$$v_{\text{post}} = g^\top H^{-1} g, \qquad v_{\text{samp}} = g^\top H^{-1} J H^{-1} g > 0. \tag{34}$$

Proposition 8 (Local Gaussian interval coverage)**.** *Under Proposition 2 and the functional and moment conditions in Appendix B.1, suppose*

$$\sqrt{n}\{\phi(\theta_n) - \phi_0\} \longrightarrow \delta. \tag{35}$$

Then the limiting coverage of a nominal $100(1-\gamma)\%$ *equal-tail raw posterior interval for* ϕ_0 *is*

$$\Phi\left(\frac{z_{1-\gamma/2}\sqrt{v_{\text{post}}} - \delta}{\sqrt{v_{\text{samp}}}}\right) - \Phi\left(\frac{-z_{1-\gamma/2}\sqrt{v_{\text{post}}} - \delta}{\sqrt{v_{\text{samp}}}}\right). \tag{36}$$

If the limiting center is instead separated from ϕ_0 *by a nonzero amount and interval radius is* $O_p(n^{-1/2})$*, true-target coverage converges to zero.*

To interpret Equation (36), let b_ϕ denote center bias divided by the matching sampling SD and let r_ϕ denote reported SD divided by that sampling SD. For an asymptotically symmetric interval with stable standardized radius,

$$C_\phi \simeq \Phi(z_{1-\gamma/2} r_\phi - b_\phi) - \Phi(-z_{1-\gamma/2} r_\phi - b_\phi). \tag{37}$$

A consistent sandwich gives $r_\phi \to 1$ but leaves b_ϕ unchanged. In Corollary 5, $b_\phi = h\eta c_\phi/\sqrt{v_\phi}$ and the raw posterior has $r_\phi = \sqrt{1+\eta}$. This quantifies why removing excess posterior spread need not improve truth coverage. Coverage of the penalized population target instead removes its displacement from truth, although finite-sample estimator bias may remain.

For empirical checks, bias and sampling SD are computed separately for each fixed reference subject and horizon, using the same estimator as the interval center. Equation (37) is applied before averaging over functionals. Ratios of aggregate RMS bias and SD cannot replace these pointwise terms. Estimated bias and width variation introduce finite-cohort error; agreement is a check of the Gaussian approximation, not an independent validation of its assumptions. Asymmetric posterior quantiles are retained as a separate interval family.

Prediction risk and coverage answer different questions. In the local-discrepancy model, $\mathcal{R}_\phi(\eta)/v_\phi = (1+\kappa_\phi\eta^2)/(1+\eta)^2$, whereas $b_\phi^2 = \kappa_\phi\eta^2$ and $r_\phi^2 = 1+\eta$ for raw intervals. Risk can therefore fall while truth coverage falls. For example, $\kappa_\phi = 1$ and $\eta = 3$ give risk ratio $10/16 = 0.625$, but raw 95% coverage is approximately $\Phi(0.92) - \Phi(-6.92) = 0.821$. A spread adjustment changes r_ϕ without removing the standardized center error. The empirical checks in Appendix D.6.2 distinguish these mechanisms using matched centers and repeated-cohort SDs; they do not estimate unknown truth bias in an application.

B.7 Vanishing-borrowing expansion around the target law

We next characterize how teacher borrowing and structural regularization can move the fitted center relative to the target law. Let

$$R(\theta) = \mathbb{E}\{r_i(\theta)\}, \qquad \theta_0 \in \arg\min_\theta R(\theta), \qquad Q(\theta) = \mathbb{E}\{q_i(\theta)\}, \tag{38}$$

and define

$$d_Q = \nabla Q(\theta_0), \qquad H_0 = \nabla^2 R(\theta_0). \tag{39}$$

Here θ_0 is the target-only population optimum within the student model. If the student is correctly specified, it indexes the target survival law; under student misspecification, it is the corresponding target-model pseudo-true value. We call the teacher *first-order compatible at* θ_0 when $d_Q = 0$. This condition differs from training an estimated teacher under the same source and target event law, which still permits finite-sample estimation error. Calibration shifts, covariate-effect shifts, and other simulation perturbations are examples of mechanisms that may make $d_Q \neq 0$; their effect enters through this score displacement.

Proposition 9 (Target-centered local decomposition)**.** *Suppose R and Q are twice continuously differentiable in a neighborhood of an identifiable interior minimizer θ_0, $H_0 \succ 0$, the empirical gradients obey the usual uniform law and stochastic equicontinuity conditions, the centered target and teacher subject scores have finite second moments and root-n sampling fluctuations, and $\eta_n \to 0$ and $\lambda_n = (n\sigma_n^2)^{-1} \to 0$. Let $\widehat{\theta}_n$ be a local minimizer of*

$$R_n(\theta) + \eta_n Q_n(\theta) + \frac{\lambda_n}{2}\|\theta\|_2^2$$

that converges in probability to θ_0. Then

$$\boxed{\widehat{\theta}_n - \theta_0 = -H_0^{-1}\left[\frac{1}{n}\sum_{i=1}^n \nabla r_i(\theta_0) + \eta_n d_Q + \lambda_n\theta_0\right] + o_p\Big(n^{-1/2} + \eta_n + \lambda_n\Big).} \tag{40}$$

Equation (40) separates three local contributions: target-cohort sampling variation, displacement induced by the teacher loss, and displacement induced by the structural prior. If $d_Q = 0$, the teacher creates no first-order deterministic shift at the target-only optimum. If $d_Q \neq 0$, the root-n relevance of teacher mismatch is governed by $\sqrt{n}\,\eta_n d_Q$. The prior contribution is controlled by $\lambda_n = (n\sigma_n^2)^{-1}$; this expansion does not prescribe a prior-tuning rule.

Corollary 10 (Target-centered posterior-mean survival functional)**.** *Let $\phi(\theta)$ be a smooth identifiable survival functional with gradient $g_0 = \nabla\phi(\theta_0)$. Under the conditions of Proposition 9 and a local posterior approximation for which posterior mean and local center differ by $o_p(n^{-1/2}+\eta_n+\lambda_n)$,*

$$\widehat{\phi}_B - \phi(\theta_0) = -g_0^\top H_0^{-1}\left[\frac{1}{n}\sum_{i=1}^n \nabla r_i(\theta_0) + \eta_n d_Q + \lambda_n\theta_0\right] + o_p\Big(n^{-1/2} + \eta_n + \lambda_n\Big). \tag{41}$$

The expansion separates target-sample fluctuation, teacher-induced displacement, and Gaussian-prior shrinkage when $\eta_n, \lambda_n \to 0$. The projection $g_0^\top H_0^{-1} d_Q$ determines the first-order effect of teacher discrepancy on a particular survival functional. For finite fixed borrowing weights, Proposition 6 instead gives local derivatives around the current center.

B.8 Proofs

The deterministic perturbation used throughout is

$$\widetilde{p}_{\varepsilon,k}(x) = \operatorname{expit}\{\operatorname{logit}\widetilde{p}_{0,k}(x) + \varepsilon d_k(x)\}. \tag{42}$$

Proof of Proposition 6. Let $F(\theta, \varepsilon) = \nabla_\theta \mathcal{M}(\theta, \varepsilon)$. The implicit function theorem applies because $F(\theta_*(0), 0) = 0$ and $\partial_\theta F = A$ is nonsingular. Differentiating $F(\theta_*(\varepsilon), \varepsilon) = 0$ at zero gives $A\theta'_*(0) + \eta B = 0$. Applying the chain rule and the first-order expansion of ϕ yields Equation (29). For any prior hyperparameter a, differentiation gives $A\partial_a\theta_* + n^{-1}\nabla_\theta\partial_a P = 0$; for borrowing it gives $A\partial_\eta\theta_* + \nabla Q = 0$. For Bernoulli KL, its gradient with respect to the student is $(p_\theta - \widetilde{p}_\varepsilon)\nabla f_\theta$ per cell. Since $\partial_\varepsilon \widetilde{p}_\varepsilon = \widetilde{p}_\varepsilon(1 - \widetilde{p}_\varepsilon)d$, summing over fixed risk-set contributions and taking expectations gives Equation (43). All derivatives are local to the specified branch and sampled coordinates. □

The Bernoulli derivative used above is

$$B = -\mathbb{E}\sum_k A_{ik}\widetilde{p}_{0,ik}(1 - \widetilde{p}_{0,ik})d_k(X_i)\nabla f_{\theta_*(0),k}(X_i). \tag{43}$$

Thus a common perturbation can move different patient/horizon functionals differently through student geometry. These derivatives describe local changes at a fixed borrowing weight.

On a corresponding smooth branch in borrowing and prior hyperparameters, evaluated at its current center,

$$\frac{\partial\phi(\theta_*)}{\partial\eta} = -g^\top A^{-1}\nabla Q(\theta_*), \qquad \frac{\partial\phi(\theta_*)}{\partial a} = -\frac{1}{n}g^\top A^{-1}\nabla_\theta\partial_a P_{\sigma,\tau}. \tag{44}$$

Using the prior decomposition in Appendix A.1, for $a = \rho_\sigma$ or ρ_τ, $\partial_a P$ is respectively G or r_T. The isotropic parameterization $P/n = \lambda\|\theta\|^2/2$ recovers $\partial_\lambda\phi = -g^\top A^{-1}\theta_*$. These derivatives describe the specified local branch; they do not guarantee that increasing shrinkage or smoothing reduces error.

Proof of Corollary 7. Write $Z_\varepsilon = \int e^{-U_\varepsilon(\theta)}d\theta$ and $N_\varepsilon = \int \phi(\theta)e^{-U_\varepsilon(\theta)}d\theta$. The envelope conditions permit differentiation under each integral. The quotient rule gives

$$\left(\frac{N_\varepsilon}{Z_\varepsilon}\right)' = -\mathbb{E}_{\Pi_\varepsilon}(\phi\,\partial_\varepsilon U_\varepsilon) + \mathbb{E}_{\Pi_\varepsilon}\phi\,\mathbb{E}_{\Pi_\varepsilon}(\partial_\varepsilon U_\varepsilon).$$

Only Q_n depends on the teacher perturbation, so $\partial_\varepsilon U_\varepsilon = n\eta\partial_\varepsilon Q_n$. A teacher-entropy term independent of student parameters drops out of the covariance. In particular, one can implement the derivative using

$$\partial_\varepsilon Q_n(\theta;\varepsilon) = \frac{1}{n}\sum_{i,k} A_{ik}\widetilde{p}_{\varepsilon,ik}(1 - \widetilde{p}_{\varepsilon,ik})d_k(X_i)\{\operatorname{logit}\widetilde{p}_{\varepsilon,ik} - f_{\theta,ik}\}.$$

For the conditional head posterior the integration variable is w and the representation is constant. Differentiating a data- or perturbation-dependent representation instead would require additional chain-rule terms and defines a different procedure. □

Proof of Proposition 2. The empirical estimating equation and its population counterpart, centered at $\theta_{*,n}$, yield

$$\sqrt{n}(\widehat{\theta}_n - \theta_{*,n}) = -A_n^{-1}\frac{1}{\sqrt{n}}\sum_i\{s_i(\theta_{*,n}) - \mathbb{E}s_i(\theta_{*,n})\} + o_p(1).$$

The prior gradient is deterministic and cancels in the centered score fluctuation; its derivative remains in A_n. The assumed Hessian law and subject-level CLT give the sandwich limit. The assumed posterior localization and quadratic expansion give local precision nA_n, with moment control giving the covariance statement. The delta method gives the local-mode functional projections, and Proposition 1 transfers their sampling limit to the posterior-mean functional. The covariance of $\nabla r_i + \eta\nabla q_i$ expands into $J_R + \eta(C_{RQ} + C_{RQ}^\top) + \eta^2 J_Q$; replacing a full patient score by independent cell scores would generally change that covariance. □

Proof of Proposition 3. At the true parameter the student probability equals the oracle teacher probability in every observed risk-set cell. The teacher score $(p_\theta - p_0)\nabla f_\theta$ is therefore zero for each subject, including its random risk-set mask. Its Hessian at truth is the sum of $A_{ik}p_{0,ik}(1 - p_{0,ik})\nabla f_{0,ik}\nabla f_{0,ik}^\top$, hence positive semidefinite. The correctly specified observed-data likelihood information identity (with noninformative censoring and correct risk sets) gives $J_R = H_R$; J_Q and

both cross terms vanish. The implicit expansion of the penalized population score gives $\theta_{*,n} - \theta_0 = -A_\eta^{-1}\nabla P_n(\theta_0)/n + o(n^{-1/2})$ under the stated local prior conditions. Its displacement is therefore negligible at root-n order, justifying evaluation at truth. Consequently,

$$A_\eta^{-1} - A_\eta^{-1}H_R A_\eta^{-1} = A_\eta^{-1}(A_\eta - H_R)A_\eta^{-1} = \eta A_\eta^{-1}H_Q A_\eta^{-1}.$$

With identical Bernoulli risk-set contributions, the expected likelihood Hessian equals the teacher Hessian at truth, since likelihood residual terms have conditional expectation zero. Thus $A_\eta = (1+\eta)H_R$. The two functional variances are respectively $g^\top H_R^{-1} g/[n(1+\eta)]$ and $g^\top H_R^{-1} g/[n(1+\eta)^2]$. Their ratio is $1+\eta$ when the leading variance is positive. For affine logits, the second derivatives of logits vanish identically, so the equal-Hessian identity in Equation (45) holds at every common evaluation point, even for a non-oracle teacher. □

Proof of Corollary 5. At the oracle center the population Hessian is $(1+\eta)H$. Proposition 6 and the negligible prior score imply

$$\sqrt{n}(\theta_{*,n} - \theta_0) = -\frac{h\eta}{1+\eta}H^{-1}B + o(1).$$

The local perturbation changes neither the limiting likelihood information nor the leading centered subject-score covariance: the latter remains H, since the oracle teacher score is zero and its perturbation vanishes in the required second-moment norm. The centered estimating-equation expansion therefore has covariance $H^{-1}/(1+\eta)^2$. The delta method first gives the local-mode functional limit. Proposition 1 and Slutsky's theorem give Equation (28) for the posterior mean, while the posterior quadratic approximation with second-moment and remainder control gives its posterior variance. Uniform integrability converts the distributional limit into the stated MSE limit. Differentiating that limiting risk with respect to η yields $2(h^2 c_\phi^2 \eta - v_\phi)/(1+\eta)^3$. Finally, dividing the limiting mean shift by the sampling SD gives $h\eta c_\phi/\sqrt{v_\phi}$. □

For an affine-logit student or fixed head, $f_{\theta,ik} = z_{ik}^\top \theta$, an implementation identity at a common evaluation point is

$$\nabla^2 r_i(\theta) = \nabla^2 q_i(\theta) = \sum_k A_{ik} p_{\theta,ik}(1 - p_{\theta,ik}) z_{ik} z_{ik}^\top. \tag{45}$$

Here $p_{\theta,ik} = \lambda_{\theta,k}(X_i)$. Teacher labels change the score and fitted center. As that center moves, the evaluated curvature changes; Equation (45) therefore supplies no universal width monotonicity result for incompatible teachers.

Proof of Proposition 9 and Corollary 10. Expand the empirical penalized score at θ_0. Its deterministic terms are $\eta_n d_Q + \lambda_n \theta_0$, and the local-likelihood score is $n^{-1}\sum_i \nabla r_i(\theta_0)$. The teacher empirical-score fluctuation is negligible on the stated $n^{-1/2} + \eta_n + \lambda_n$ scale under the uniform-law assumptions. The Hessian converges to H_0 and the Taylor remainder is negligible on that scale, giving Equation (40). Taylor expansion of ϕ and the assumed posterior-mean/center equivalence then give Equation (41). These conclusions require vanishing η_n, λ_n and do not extend the expansion to a fixed large borrowing weight. □

Proof of Corollary 4. The oracle score has no teacher contribution at truth, so the subject-score CLT, local Hessian law, delta method, and Proposition 1 give

$$\sqrt{n}(\bar{\phi}_{\eta,n} - \phi_0) \Rightarrow N\left(0, \frac{v_\phi}{(1+\eta)^2}\right).$$

The functional Bernstein–von Mises approximation places the two raw posterior endpoints at $\bar{\phi}_{\eta,n} \pm z_{1-\gamma/2}\sqrt{v_\phi/\{n(1+\eta)\}} + o_p(n^{-1/2})$. Standardizing their coverage event by the sampling SD gives $C_{\mathrm{raw},\eta}$. For the adjusted interval, variance consistency and Slutsky's theorem yield

$$\frac{\bar{\phi}_{\eta,n} - \phi_0}{\sqrt{\widehat{V}_{\mathrm{post},\phi}/(1+\eta)}} \Rightarrow N(0,1),$$

which proves nominal coverage. A consistent sandwich has the same denominator limit. Dividing the adjusted radii at η and zero gives the width ratio; an $o_p(n^{-1/2})$ change of center leaves these limits unchanged. The centered CLT and variance consistency establish the adjusted Wald result; the posterior-quantile result additionally uses the Bernstein–von Mises approximation (Miller, 2021). □

Proof of Proposition 8. Proposition 2, the functional delta method, and Equation (35) yield

$$\sqrt{n}\{\phi(\widehat{\theta}_n) - \phi_0\} \Rightarrow N(\delta, v_{\text{samp}}).$$

The rescaled equal-tail posterior endpoints approach $\pm z_{1-\gamma/2}\sqrt{v_{\text{post}}}$ about $\phi(\widehat{\theta}_n)$. Taking the probability that a variable with law $N(\delta, v_{\text{samp}})$ falls between those limits gives Equation (36). With a persistent nonzero limiting center discrepancy, consistency and a shrinking $O_p(n^{-1/2})$ radius instead imply coverage tending to zero. □

C Target-Predictive Selection and Case Deletion

Selection evaluates target-outcome prediction under a declared fitting procedure. Both main simulations fix the reference prior. Simulation I reports PSIS choices from 0 to 30. Simulation II selects among six weights including zero; its main prediction figure and supplementary tables report the resulting cohort-wise predictions. Its full-network PSIS, WAIC, and CV-weight comparison uses the same reference-prior menu. METABRIC profiles three priors; the NWTCO target-size sensitivity holds its reference prior fixed, and the transplant application uses an isotropic Gaussian prior. Every deletion identity below holds a candidate's complete prior fixed; π_σ can be replaced by $\pi_{\sigma,\tau}$ without changing the algebra.

C.1 Generalized case-deletion CPO

For $\ell_i(\theta) = \log L_i(\theta)$ and $a_i = \ell_i - \eta q_i$, write

$$Z = \int \pi_\sigma(\theta) \exp\{\textstyle\sum_j a_j(\theta)\}\, d\theta, \qquad Z_{-i} = \int \pi_\sigma(\theta) \exp\{\textstyle\sum_{j\neq i} a_j(\theta)\}\, d\theta.$$

Hold the teacher, prior, sampled coordinates, preprocessing, imputation, time grid, and all remaining subjects' loss contributions fixed. Assume common support, proper full and deleted posteriors, and finite displayed expectations. Normalizing the deleted density gives

$$\frac{d\Pi_{-i}}{d\Pi}(\theta) = \frac{e^{-a_i(\theta)}}{\mathbb{E}_\Pi e^{-a_i(\theta)}}, \tag{46a}$$

$$\mathrm{CPO}_i = \frac{\mathbb{E}_\Pi\{L_i e^{-a_i}\}}{\mathbb{E}_\Pi e^{-a_i}} = \frac{\mathbb{E}_\Pi e^{\eta q_i}}{\mathbb{E}_\Pi e^{\eta q_i - \ell_i}}. \tag{46b}$$

The proper Gaussian prior, nonnegative KL losses, and finite logistic predictions make the remaining generalized factors positive and bounded by one. Their normalizers are therefore finite and positive. This does not establish finite importance-weight variance. At $\eta = 0$, Equation (46) reduces to ordinary likelihood deletion, $\mathrm{CPO}_i = \{\mathbb{E}_\Pi(1/L_i)\}^{-1}$. The aggregate criterion is the log pseudo-marginal likelihood, $\mathrm{LPML}(\eta) = \sum_i \log \mathrm{CPO}_i(\eta)$; its grid maximizer selects a borrowing weight.

C.2 Importance estimation and PSIS

The algebraic identity is exact for the specified posterior; finite-draw estimates are approximate. For posterior draws, let $b_{im} = \eta q_i(\theta^{(m)}) - \ell_i(\theta^{(m)})$. The self-normalized estimate is

$$\widehat{\mathrm{CPO}}_i = \frac{\sum_m e^{b_{im}} L_i(\theta^{(m)})}{\sum_m e^{b_{im}}}, \qquad \log \widehat{\mathrm{CPO}}_i = \mathrm{LSE}_m(b_{im} + \ell_{im}) - \mathrm{LSE}_m(b_{im}), \tag{47}$$

where LSE denotes log-sum-exp. Teacher-only constants in q_i shift all of a subject's log weights equally and cancel, so they may be removed for stability. There is no additional factor of n on a subject's q_i.

PSIS smooths the raw ratios $e^{b_{im}}$ (Vehtari et al., 2024). With normalized smoothed weights v_{im}, the predictive estimate remains $\widehat{\mathrm{CPO}}_i^{\mathrm{PSIS}} = \sum_m v_{im} L_i(\theta^{(m)})$. Weighting uses a_i but scoring uses L_i: passing a_i to a standard LOO routine and reporting its default score would instead evaluate the generalized factor e^{a_i}.

Assess underlying posterior means and predictive scores before importance estimation. Record subject/candidate Pareto $\widehat{k}$, importance ESS, autocorrelation adjustment, chain and sampling-budget

sensitivity, and explicit-refit status under a documented, prespecified diagnostic rule. Larger borrowing can produce difficult ratios, but $\widehat{k}$ need not increase monotonically because the posterior also changes. PSIS cannot repair poor MCMC or recover regions absent from the draws. Explicit deletion fits can assess unreliable conditional importance estimates; complete refitted CV instead evaluates the full training procedure. A finite importance score alone is insufficient evidence of numerical accuracy. The reported simulations retain all prespecified PSIS selections and identify their warnings rather than applying an unplanned fallback.

C.3 Conditioning and complete refitting

Full-state case deletion reweights every sampled parameter under the fixed specification. Conditional-head deletion removes both subject contributions while keeping $\widehat{\psi}$ fixed. When that representation was learned from the full target cohort, the identity remains valid for its conditional posterior but does not give leakage-free validation of the complete training algorithm. A representation trained on independent data and then frozen permits direct evaluation of that conditional procedure. Deletion-dependent changes in preprocessing, imputation, or grid also require refitting or an explicitly modified density ratio.

For complete K-fold validation, relearn all declared data-dependent transformations and the representation using training subjects only, excluding held-out subjects from both likelihood and teacher loss. At each fixed η, fit the full-state or conditional-head posterior specified by the algorithm and score

$$S_v(\eta) = \sum_{i \in v} \log \mathbb{E}_{\Pi_\eta^{(-v)}} L_i(\theta), \qquad \widehat{\eta}_{\mathrm{CV}} \in \arg\max_{\eta \in \mathcal{G}} \sum_v S_v(\eta).$$

The deployed teacher stays fixed. A fold with m training subjects keeps the same σ, giving average-loss penalty $1/(m\sigma^2)$; retaining the full-data penalty coefficient would change the prior. Refit the final model on all target-training subjects at the selected weight. Folds with overlapping training sets are not independent replications.

Exact LOO refits with the *same conditioning* can assess the accuracy of importance estimates. This numerical validation is distinct from descriptively comparing their implied weights and predictions with end-to-end K-fold CV, as in the simulation selection comparison. Selector comparisons concern the resulting predictions: independent-test deviance, simulation survival RMSE, paired excess risk relative to $\eta = 0$, and differences between selected predictors, together with runtime and numerical diagnostics. Retrospective candidate-path regret is

$$R_{\mathrm{test}}(\widehat{\eta}_{\mathrm{selector}}) - \min_{\eta \in \mathcal{G}} R_{\mathrm{test}}(\eta).$$

This finite-test-sample hindsight comparison is specific to $\mathcal{G}$ and never selects the operational model. Exact agreement of selected weights and boundary frequencies are secondary; a claim that disagreement has little predictive consequence requires measured differences and uncertainty.

C.4 Likelihood-only WAIC comparator

The simulation WAIC criterion is

$$\mathrm{WAIC} = -2 \sum_i \{\mathrm{lppd}_i - \mathrm{Var}_\Pi(\ell_i)\}, \qquad \mathrm{lppd}_i = \log \mathbb{E}_\Pi e^{\ell_i}. \tag{48}$$

It scores the target likelihood but does not explicitly represent deletion of the subject's teacher contribution. A WAIC–refitting discrepancy can also reflect representation or preprocessing refitting, smaller training samples in K-fold CV, Monte Carlo error, and finite-sample score variability. A single comparison does not isolate these causes (Vehtari et al., 2017).

C.5 Deletion approximation and selection sensitivity

The deletion identity, the covariance expansion, and finite-draw PSIS have different approximation roles. The identity in Equation (46) is exact under its stated conditioning. To display its relation to WAIC, fix a candidate posterior and write $F_X(t) = \log \mathbb{E}_\Pi e^{tX}$. Suppose F_X is three times

continuously differentiable on $[0,1]$ for $X \in \{\eta q_i, \eta q_i - \ell_i, \ell_i\}$. Taylor's formula gives

$$F_X(1) = \mathbb{E}_\Pi X + \tfrac{1}{2}\mathrm{Var}_\Pi X + R_X, \qquad R_X = \tfrac{1}{2}\int_0^1 (1-t)^2 F_X'''(t)\,dt. \tag{49}$$

Applying this formula to the two exact CPO normalizers and to lppd_i yields

$$\begin{aligned} \log \mathrm{CPO}_i &= \mathrm{lppd}_i - \mathrm{Var}_\Pi(\ell_i) + \eta\,\mathrm{Cov}_\Pi(\ell_i, q_i) + \mathcal{E}_i, \\ \mathcal{E}_i &= R_{\eta q_i} - R_{\eta q_i - \ell_i} - R_{\ell_i}. \end{aligned} \tag{50}$$

If $|F_X'''(t)| \le M_X$ along each segment, then $|\mathcal{E}_i| \le (M_{\eta q_i} + M_{\eta q_i - \ell_i} + M_{\ell_i})/6$. The covariance term accounts for the teacher contribution omitted by likelihood-only WAIC; its sign and higher-order remainder depend on the candidate. This specializes the predictive-score expansion underlying LOO/WAIC comparisons (Vehtari et al., 2017). It is an explanation of the existing criterion, not an additional selector, and the displayed bound supplies no numerical error guarantee without bounds on the tilted third cumulants. PSIS estimates the full density-ratio expression, not its covariance truncation; finite-draw error additionally depends on posterior sampling and importance tails (Vehtari et al., 2024).

Selection can change even when candidate scores are close. Let $s_\eta = n^{-1}\sum_i \log \mathrm{CPO}_i(\eta)$, $\eta_* \in \arg\max_{\eta\in\mathcal{G}} s_\eta$, and $\widehat{\eta} \in \arg\max_{\eta\in\mathcal{G}} \widehat{s}_\eta$. If $\max_\eta |\widehat{s}_\eta - s_\eta| \le \epsilon$, then

$$0 \le s_{\eta_*} - s_{\widehat{\eta}} \le 2\epsilon. \tag{51}$$

Indeed, insert $\widehat{s}_{\eta_*}$ and $\widehat{s}_{\widehat{\eta}}$ between the two exact scores and use their maximizing property. A unique exact winner is preserved if its gap to every other candidate exceeds 2ϵ; nearly tied weights can change while losing little score. This deterministic observation concerns the same exact deletion target, not truth-functional MSE. Half-draw or chain comparisons diagnose sensitivity but are not certified values of ϵ.

The existing five-fold benchmark relearns a MAP representation and samples its head, using smaller training cohorts. Appendix C.6.2 evaluates full-network score stability and its consequences for prediction and interval width, alongside practical CV-weight comparisons. Matched full-network deletion refits remain a separate validation target.

C.6 Empirical selection checks in the simulations

C.6.1 Selection scope and supplementary refitted validation

For ten prespecified cohorts under compatible, larger-calibration, and late-temporal guidance, the saved five-fold benchmark refits the MAP representation in each fold . Its held-out outcome scores are restricted here to the same six-weight reference-prior menu. Table 2 compares weights selected by full-network PSIS and WAIC with this benchmark, evaluating every weight through the corresponding full-network posterior. Native CV-head prediction is a separate quantity: at CV-selected weights, mean RMSE is 0.0178, 0.0660, and 0.0210 for the three regimes; full-network predictions at those same weights have RMSE 0.0174, 0.0522, and 0.0200.

PSIS selects the same weight as CV in 9/10 compatible, 9/10 larger-calibration, and 7/10 late-temporal cohorts. Mean CV-score losses per subject are 0.00021, 0.00040, and 0.00039. Prediction differences are small on the compatible and late-temporal paths, where different large weights can perform similarly. These ten-cohort comparisons assess practical agreement, subject to the posterior and importance precision qualifications below.

Table 2: Full-network selectors compared at CV-selected weights on ten paired cohorts. Score loss is per subject. Prediction differences are selector minus CV prediction; negative values favor the selector. Parentheses give paired-cohort MCSEs.

Teacher	Selector	Same η	CV loss	Δ NLL	Δ IBS	Δ RMSE
Compatible	PSIS	9/10	0.00021	-0.00071 (0.00046)	-0.00010 (0.00007)	-0.00242 (0.00153)
Compatible	WAIC	8/10	0.00021	-0.00073 (0.00045)	-0.00010 (0.00007)	-0.00260 (0.00149)
Calibration $+0.50$	PSIS	9/10	0.00040	-0.00154 (0.00154)	-0.00029 (0.00029)	-0.00213 (0.00213)
Calibration $+0.50$	WAIC	8/10	0.00208	+0.00005 (0.00232)	-0.00008 (0.00038)	-0.00068 (0.00271)
Late temporal	PSIS	7/10	0.00039	-0.00046 (0.00046)	-0.00010 (0.00006)	-0.00133 (0.00136)
Late temporal	WAIC	6/10	0.00029	-0.00052 (0.00045)	-0.00009 (0.00006)	-0.00120 (0.00137)

C.6.2 Full-network selection stability and its predictive consequences

Full-network score sensitivity. Table 3 uses all 100 cohorts per teacher and the saved full-network scores. The gap is the difference between the largest and second-largest pooled LPML. Split-half selections use the first and second production halves separately; chain sensitivity records whether any individual chain selects a weight different from the pooled choice. Smaller median gaps accompany more frequent half-draw disagreements in these data. This pattern is consistent with sensitivity of near-tied rankings, but does not separate MCMC error from importance-estimation error or constitute an exact-LOO error bound. No matched full-network deletion refits or numerical validation of the covariance-expansion remainder are available in this study.

Table 3: Full-network selection sensitivity, with 100 paired cohorts per teacher. LPML gaps are unnormalized sums over 500 subjects. Half and chain columns count cohorts, not independent replications.

Teacher	Median top-two gap	Halves disagree	Any chain differs
Compatible	.388	37/100	86/100
Calibration $+0.10$	.353	45/100	88/100
Calibration $+0.50$	1.874	5/100	39/100
Risk magnitude	1.428	14/100	48/100
Late temporal	.352	33/100	90/100

Table 2 compares predictive scores and independent-test performance at PSIS- and CV-selected weights. This practical benchmark and the within-fit score-sensitivity analysis answer different questions.

Prediction and interval consequences of selection changes. We apply each saved half-production and single-chain choice to that candidate's full-production prediction and posterior interval summaries, holding the evaluation subjects and outcomes fixed. This comparison isolates the effect of choosing a different weight while holding candidate summaries fixed; it does not assess within-fit Monte Carlo accuracy of interval endpoints. The check includes all 600 Simulation I and 500 Simulation II decisions, retaining their original candidate menus; half-production scores are unavailable for the linear student. Table 4 summarizes all available alternatives.

Mean survival-RMSE gains over target-only inference remain positive in every law–student or teacher group under every available half or chain choice. Changes in cohort-average RMSE are generally smaller than changes in interval width. Individual cohorts can respond more strongly: the largest absolute RMSE change reaches 0.0438, and under risk-magnitude discrepancy a single-chain choice changes whether RMSE improves over target-only inference in up to 16 of 100 cohorts. The table separates these individual consequences from the average borrowing benefit.

Numerical scope. All candidates are retained. The maximum functional $\widehat{R}$ is 1.0058, minimum tail ESS is $2{,}883$, and maximum posterior-mean MCSE is 0.00384. For the interval-specific MLP across both generation laws, the corresponding values are 1.0124, 691, and 0.00966. Table 3 reports the sensitivity of selection to production halves and individual chains. The full-network results therefore describe exploratory prediction and interval-concentration patterns; close score rankings

Table 4: **Consequences of full-state PSIS selection variability.** Each cell reports half-production / single-chain results, with 100 paired cohorts per row. For each family, columns give the largest mean absolute change from the pooled-choice RMSE or mean interval width, the largest absolute cohort RMSE change, and the largest count of cohorts whose positive RMSE gain versus target-only changes status. Maxima are over the two half choices or four chain choices, respectively; counts are not pooled across choices. Widths are raw 95% survival-probability interval widths. All outcomes use the same full-production candidate summaries. Dashes denote unavailable half scores.

Setting	Mean $\lvert\Delta\mathrm{RMSE}\rvert$	Mean $\lvert\Delta\mathrm{width}\rvert$	Max $\lvert\Delta\mathrm{RMSE}\rvert$	Gain-status changes
I: Linear/Linear	– / 0.0000	– / 0.0008	– / 0.0017	– / 0
I: Linear/MLP	0.0002 / 0.0004	0.0023 / 0.0044	0.0029 / 0.0033	0 / 0
I: Linear/Shared NN	0.0002 / 0.0006	0.0028 / 0.0077	0.0029 / 0.0038	0 / 0
I: Nonlinear/Linear	– / 0.0000	– / 0.0012	– / 0.0005	– / 1
I: Nonlinear/MLP	0.0009 / 0.0013	0.0029 / 0.0046	0.0113 / 0.0113	0 / 0
I: Nonlinear/Shared NN	0.0003 / 0.0006	0.0037 / 0.0089	0.0044 / 0.0045	0 / 0
II: Calibration $+0.50$	0.0001 / 0.0015	0.0008 / 0.0034	0.0098 / 0.0267	2 / 9
II: Calibration $+0.10$	0.0008 / 0.0016	0.0113 / 0.0178	0.0147 / 0.0326	0 / 1
II: Compatible	0.0010 / 0.0027	0.0065 / 0.0173	0.0176 / 0.0201	0 / 0
II: Risk magnitude	0.0006 / 0.0016	0.0030 / 0.0071	0.0129 / 0.0165	6 / 16
II: Late temporal	0.0010 / 0.0026	0.0063 / 0.0160	0.0178 / 0.0438	0 / 1

and precise interval-efficiency differences remain unresolved. The Monte Carlo condition in Appendix B.2 is separate from the statistical approximation.

Completed MLP paths and production-budget sensitivity. The six-weight Simulation I grid contains 3,600 configurations, with 100 cohorts per cell. Of the 524 added MLP fits, 522 retain sampling-review flags and 48 importance warnings. Substituting each production half for these added fits preserves both decreasing mean RMSE paths, with maximum point change 0.00088; reused fits are held fixed. This is a path-sensitivity check, not convergence certification or a controlled comparison of production lengths.

D Simulation Design and Supplementary Results

D.1 Experimental design

D.1.1 Native-discrete event and censoring laws

Let $X \sim N(0, I_3)$, $\beta = (0.8, -0.6, 0.5)^\top$, and $s_L = \|\beta\|$. With $s_G^2 = (1 - e^{-2})/2 + 1/4$, define

$$f_L(x) = x^\top \beta, \qquad f_N(x) = c\left[x^\top \beta + \frac{0.75 s_L}{s_G}\{\sin(x_1) + 0.5 x_2 x_3\}\right],$$

where

$$c = \frac{s_L}{\{s_L^2 + (0.75 s_L)^2 + 2(0.75 s_L)(0.8)e^{-1/2}/s_G\}^{1/2}}.$$

The two risk functions therefore have the same marginal variance. For $r \in \{L, N\}$, the true discrete hazards are

$$\lambda_{0k}(x) = \mathrm{expit}\{a_r + 0.9[-1 + 2(k-1)/9] + f_r(x)\}, \qquad k = 1, \ldots, 10.$$

Generate the first event from these hazards, setting $T = 11$ if no event occurs by interval ten. Independently draw $C \in \{6, 7, 10\}$ with probabilities $(0.25, 0.25, 0.50)$. An event at $T = C$ is observed. Each a_r is calibrated by independent population integration to $\Pr(T \le C) = 0.628$. The observed zero-based index is $d = \min(T, C) - 1$, with event indicator $e = \mathbb{1}\{T \le C\}$. Thus an occupied censoring interval contributes an observed non-event. All target cohorts have $n = 500$.

D.1.2 Teachers and students

The linear teacher is a fitted logistic-hazard model and the nonlinear teacher a shared-risk neural model with SiLU hidden layers of widths 64 and 32. Both use 50,000 independent source subjects,

with a separate validation cohort for source-model fitting. An independent panel of 20,000 covariates measures

$$R_F^2 = 1 - \frac{\sum_{j,k}\{\widetilde{F}_k(x_j) - F_{0k}(x_j)\}^2}{\sum_{j,k}\{F_{0k}(x_j) - \overline{F}_{0k}\}^2}.$$

The realized values are 0.9997515 and 0.9970913. Teacher training is not repeated across target cohorts. "Compatible" denotes an estimated teacher trained under the same event law; it does not assert the exact population score condition $d_Q = 0$. Direct-teacher results in Table 6 use these fixed mappings.

The neural students have 16 SiLU representation units, divided by $\sqrt{16}$. The shared-risk model has ten intercepts and 16 shared output weights; the interval-specific MLP has a separate intercept and 16 weights at every interval. Full-network sampling includes all 90 and 234 parameters, respectively; the linear student samples all 13 parameters. Each candidate is initialized from the penalized fitting procedure, then samples its complete state. The conditional-head alternative has 26 and 170 sampled parameters and is used only in separately labelled supplementary analyses.

For Simulation II, let $\widetilde{f}_k(x) = a_{Tk} + f_T(x)$ be the compatible teacher logit, with shared covariate score f_T. Calibration teachers add 0.10 or 0.50 at every interval. The risk-magnitude teacher adds $0.50\{f_T(x) - \overline{f}_T\}/s_T$, with location and RMS scale fixed using the independent prior covariate panel. The temporal teacher adds $0.15 d_{Tk}$, where

$$d_T = (0, 0, 0, 0, 0, 0, 0, 1, -1, 1)\sqrt{10/3}.$$

This direction has RMS one across intervals and has nonzero mean. These are fixed prediction mappings; no teacher is retrained across target cohorts.

D.1.3 Prior, fitting, and selection

Writing the head as (a, b), the implemented negative log prior, up to a constant, is

$$P_{\sigma,\tau}(\psi, a, b) = \frac{\|\psi\|^2}{2} + \frac{\|a\|^2}{2(10)^2} + \frac{\|b\|^2}{2\sigma^2} + \frac{1}{2\tau^2 M}\sum_{j=1}^{M}\sum_{k=2}^{10}\{f_{\psi,a,b,k}(x_j^{\text{prior}}) - f_{\psi,a,b,k-1}(x_j^{\text{prior}})\}^2. \tag{52}$$

The $M = 2,000$ prior-reference covariates are independent of fitting data, test data, and the distinct 2,000-subject evaluation reference panel, and remain fixed under subject deletion and CV. The absent representation block is omitted for the linear student. The intercept and representation scales are fixed separately from the output-weight and temporal scales. Temporal regularization acts on full logits; for a shared-risk student, the covariate effect cancels in adjacent differences, so it smooths interval intercepts. For the interval-specific MLP it also smooths time-varying covariate effects. This is a difference between parameterizations, not an additional source of teacher information.

Losses are summed over subjects, with the prior added once. Neither n nor $1 + \eta$ rescales the posterior or prior during fitting. Both main simulations fix $(\sigma, \tau) = (10, 0.25)$. Simulation I evaluates fixed paths at $\eta \in \{0, 0.25, 1, 3, 10, 30\}$ and reports PSIS predictions chosen among it. Simulation II maximizes full-network PSIS LPML over $\{0, 0.25, 1, 3, 10, 30\}$ separately in every cohort. Exact ties favor the smaller weight. Test metrics, truth, and coverage never enter selection; warnings do not remove candidates. Historical conditional-head diagnostics retain their original declared priors and menus.

The original main comparisons use four neural chains with 32,000 warm-up and 256,000 production transitions per chain, retaining every 32nd transition. They comprise 3,500 distinct neural configurations and 600 full-state linear configurations, with shared configurations reused between experiments. The complete six-weight Simulation I paths additionally reuse historical candidates and add 524 MLP configurations at $\eta = 0.25, 1, 3$. These added configurations use the same initialization, target, proposal, seeds, warm-up and thinning, with 128,000 production transitions per chain. The path grid contains 3,600 configurations (36 cells of 100 cohorts), including 3,076 reused configurations; this overlapping count is separate from the original main-study union. The linear candidates retain their original four chains, 2,000 warm-up and 4,000 retained transitions per chain. Proposal adaptation ends before production. The earlier conditional-head supplementary fits use the original linear-sized sampling budget and are not substituted for full-network candidates.

D.1.4 PREDICTIVE AND UNCERTAINTY METRICS

Let s_{ik} be the posterior-mean survival probability at interval k, and let $\widehat{G}_k$ estimate $\Pr(C \geq k)$ from the target training cohort, using the event-before-censor ordering. Define

$$E_{ik} = \mathbb{1}\{e_i = 1, d_i + 1 \leq k\}, \qquad A_{ik}^{\text{surv}} = \mathbb{1}\{d_i + 1 > k\} + \mathbb{1}\{e_i = 0, d_i + 1 = k\}.$$

For m test subjects,

$$\text{IBS} = \frac{1}{10m} \sum_{i,k} \left\{ \frac{E_{ik} s_{ik}^2}{\widehat{G}_{d_i+1}} + \frac{A_{ik}^{\text{surv}} (1 - s_{ik})^2}{\widehat{G}_k} \right\},$$

$$\text{IBLL} = -\frac{1}{10m} \sum_{i,k} \left\{ \frac{E_{ik} \log(1 - s_{ik})}{\widehat{G}_{d_i+1}} + \frac{A_{ik}^{\text{surv}} \log s_{ik}}{\widehat{G}_k} \right\}.$$

These are survival-status scores, distinct from the observed-history negative log predictive density, $\text{NLL} = -m^{-1} \sum_i \log \mathbb{E}_\Pi L_i$. Predictive deviance is $2\,\text{NLL}$; the two columns differ only in scale. Concordance compares an event at d_i only with subjects having $d_j > d_i$, assigning half credit to prediction ties. For cohort r, reference-panel RMSE is $r_r = \{(MK)^{-1} \sum_{j,k} (\bar{S}_{rjk} - S_{0jk})^2\}^{1/2}$. Unless explicitly stated otherwise, prediction tables and the aggregate panels of the main simulation figures report $\bar{r}$ and its sample SD across cohorts; supplementary center–spread decompositions use pooled RMSE $\{R^{-1} \sum_r r_r^2\}^{1/2}$. These summaries differ because the square root is nonlinear. For repeated-cohort prediction scores m_r, tables give $\bar{m}$ (s_m) with $s_m^2 = \sum_r (m_r - \bar{m})^2/(R - 1)$. The Monte Carlo standard error of the reported mean is $s_m/\sqrt{R} = s_m/10$ for $R = 100$. The supplementary selector comparison instead reports paired differences and their MCSEs, using pooled RMSE as stated in its caption. The real-data analysis retains its grouped-time evaluation convention; these native-discrete formulas describe the simulations. Figure 2(b) uses cohort index 0 and reference-patient index 0, fixed before inspecting the displayed curves. In this single-event setting, $F = 1 - S$; posterior mean CIF is $1 - \mathbb{E}_\Pi S$ and its pointwise 95% interval is $[1 - S_{0.975}, 1 - S_{0.025}]$. The curves and the stars in panel (a) use each teacher's actual PSIS selection for that cohort; circle sizes in (a) summarize selection frequencies over 100 cohorts. For the paired RMSE-gain summaries, let d_r be target-only minus PSIS-selected reference-panel RMSE in cohort r. The pointwise interval for the mean gain is $\bar{d} \pm t_{0.975,99} s_d/\sqrt{100}$. It measures repeated-cohort uncertainty conditional on the fixed teacher and reference panel, is not adjusted for simultaneous comparisons, and does not include error from posterior computation.

Survival, cumulative risk, and hazard summaries. The primary estimate is $\bar{S}_k(x) = \mathbb{E}_\Pi \prod_{\ell \leq k} (1 - \lambda_{\theta,\ell}(x))$, not a product of posterior-mean hazards. For the single event, $\bar{F}_k = 1 - \bar{S}_k$ and $F_{0k} = 1 - S_{0k}$, so $\text{RMSE}(\bar{F}, F_0) = \text{RMSE}(\bar{S}, S_0)$ on the same reference cells.

Two distinct hazard summaries are

$$\bar{\lambda}_k(x) = \mathbb{E}_\Pi \lambda_{\theta,k}(x), \qquad \lambda_k^{\text{pred}}(x) = 1 - \frac{\bar{S}_k(x)}{\bar{S}_{k-1}(x)}, \quad \bar{S}_0(x) = 1.$$

The first is the posterior mean of the conditional-hazard functional; the second is the hazard of the posterior predictive event distribution. Except at the first interval or in special cases, they need not agree. Supplementary hazard RMSE uses the same fixed reference panel and equal interval weighting as survival RMSE, separately for these two summaries. The at-risk test Brier score is

$$\text{BS}_\lambda = \frac{\sum_{i,k} Y_{ik} (D_{ik} - \lambda_k^{\text{pred}}(x_i))^2}{\sum_{i,k} Y_{ik}}, \qquad Y_{ik} = \mathbb{1}\{d_i + 1 \geq k\}, \quad D_{ik} = \mathbb{1}\{d_i + 1 = k, e_i = 1\}.$$

It assesses observed at-risk cells and is distinct from the IPCW survival-status IBS.

Bias relative to the true survival probability is first estimated across cohorts separately for each fixed subject and horizon, then summarized by its RMS. EmpSD is likewise formed across cohorts before RMS aggregation. In the center–spread tables, reported SD is the RMS posterior, curvature, or sandwich SD across the same functionals and cohorts; it is not an average SD ratio. Raw intervals use posterior quantiles. Both Wald alternatives are centered at the MAP and clipped to $[0, 1]$. Coverage

and width are averaged within each cohort and then across cohorts. Monte Carlo standard errors use cohorts as the independent units; pooled RMSE differences use 2,000 paired cohort-bootstrap resamples. Patient width deciles are formed within each cohort using width alone.

For the posterior-mean decomposition, each cohort's estimate is its posterior-mean survival curve. At each fixed subject and horizon, the empirical squared bias is the squared difference between its across-cohort mean and true survival. With $R = 100$ cohorts and EmpSD computed using divisor $R - 1$, the exact pooled decomposition is $\mathrm{MSE} = \mathrm{RMSBias}^2 + (R-1)\mathrm{EmpSD}^2_{\mathrm{RMS}}/R$. The sampling-variance component includes this finite-R factor. Posterior SD is aggregated in RMS over cohorts, subjects, and horizons; empirical SD is aggregated in RMS over the same subjects and horizons.

D.2 Complete Simulation I outcomes and profiles

For the complete 100-cohort configurations in Figure 1, error bars are cross-cohort SDs of individual-cohort RMSE, not confidence intervals for a mean. Cohort-paired PSIS-minus-target-only contrasts have pointwise 95% t intervals excluding zero in the favorable direction for concordance, deviance, IBS, and survival RMSE in all six law–student combinations. These intervals quantify variation across cohorts; they do not incorporate posterior Monte Carlo error or simultaneous-comparison adjustment. Numerical precision is assessed separately in Appendix C.6.2.

Table 5 retains absolute prediction means and cross-cohort SDs for the procedures compared in the main text. Figure 5 complements the main borrowing paths with horizon-specific profiles. For SharedNN under nonlinear generation, mean interval width falls from 0.1953 to 0.0488; for MLP it falls from 0.1664 to 0.0454, alongside the corresponding survival-RMSE gains. The nonlinear linear student's width also contracts, despite persistent representation error. A matched linear-student analysis in Appendix D.6.1 checks the center–spread approximation and explains the role of residual teacher error in truth coverage.

Table 5: **Simulation I: complete absolute predictive outcomes.** PSIS-selected borrowing versus target-only inference with the same reference prior. Prediction entries are mean (SD) over 100 paired cohorts. Higher C^{td} and lower deviance and IBS are better. Width is the mean raw pointwise 95% interval width; its magnitude is distinct from the cross-cohort SD of a prediction metric.

Student	Procedure	C^{td}	Deviance	IBS	Width
Linear event law					
Linear	Target only	0.7520 (0.0047)	3.4863 (0.0279)	0.1495 (0.0034)	0.1028
Linear	PSIS	0.7528 (0.0046)	3.4716 (0.0270)	0.1487 (0.0032)	0.0246
Shared-risk NN	Target only	0.7490 (0.0051)	3.5042 (0.0280)	0.1511 (0.0036)	0.1896
Shared-risk NN	PSIS	0.7528 (0.0046)	3.4718 (0.0269)	0.1487 (0.0032)	0.0417
MLP	Target only	0.7507 (0.0047)	3.5032 (0.0276)	0.1502 (0.0035)	0.1500
MLP	PSIS	0.7528 (0.0046)	3.4719 (0.0268)	0.1487 (0.0032)	0.0385
Nonlinear event law					
Linear	Target only	0.7424 (0.0054)	3.5342 (0.0318)	0.1536 (0.0032)	0.1063
Linear	PSIS	0.7433 (0.0053)	3.5227 (0.0297)	0.1528 (0.0032)	0.0246
Shared-risk NN	Target only	0.7503 (0.0054)	3.4980 (0.0338)	0.1501 (0.0034)	0.1953
Shared-risk NN	PSIS	0.7553 (0.0049)	3.4624 (0.0293)	0.1473 (0.0032)	0.0488
MLP	Target only	0.7485 (0.0054)	3.5193 (0.0325)	0.1511 (0.0034)	0.1664
MLP	PSIS	0.7553 (0.0048)	3.4629 (0.0292)	0.1473 (0.0032)	0.0454

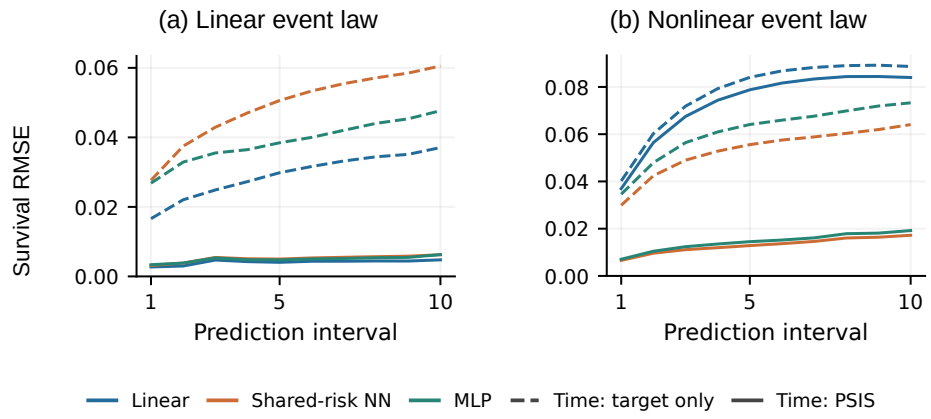


Figure 5: **Simulation I: horizon-specific prediction error.** Linear (left) and nonlinear (right) event laws, with student indicated by color and target-only/PSIS prediction by dashed/solid lines. Each curve uses all 100 paired cohorts and the reference prior. RMSE pools squared errors over cohorts and reference subjects at each horizon.

D.3 Prediction and borrowing selection

Selection frequencies and predictive outcomes are reported below. Appendix C.6.2 gives the corresponding sampling and importance qualifications, score stability, and consequences of selection changes.

Table 6: Direct fixed-teacher prediction on the same independent test cohorts and reference panel. Test prediction entries are mean (SD) over 100 cohorts, with SD below the mean. Reference-panel RMSE is constant because both teacher and reference subjects are fixed. These references have no student posterior intervals.

Event law	Teacher	C^{td}	Deviance	IBS	IBLL	RMSE
		0.7528	3.4715	0.1487	0.4537	
Linear	Compatible	(0.0046)	(0.0270)	(0.0032)	(0.0092)	0.0037
		0.7554	3.6174	0.1610	0.4912	
Nonlinear	Calibration $+0.50$	(0.0048)	(0.0281)	(0.0040)	(0.0118)	0.1180
		0.7554	3.4696	0.1480	0.4525	
Nonlinear	Calibration $+0.10$	(0.0048)	(0.0286)	(0.0033)	(0.0095)	0.0302
		0.7554	3.4618	0.1472	0.4503	
Nonlinear	Compatible	(0.0048)	(0.0292)	(0.0032)	(0.0092)	0.0119
		0.7554	3.5753	0.1535	0.4747	
Nonlinear	Risk magnitude	(0.0048)	(0.0364)	(0.0034)	(0.0104)	0.0802
		0.7554	3.4684	0.1473	0.4506	
Nonlinear	Late temporal	(0.0048)	(0.0295)	(0.0033)	(0.0093)	0.0157

D.3.1 Detailed fixed-weight prediction results

All rows use the same 100 cohorts, reference prior, and posterior-mean predictions as the main figures. Tables 7 and 8 retain the three full-state weights 0, 10, and 30. IBLL complements the deviance and IBS reported in the main text. Survival RMSE equals cumulative-event-risk RMSE in the single-event setting.

D.3.2 Reference-prior PSIS-selected prediction

The full-network procedure in Table 9 chooses among all six weights, independently within each of the 100 cohorts. Under compatible guidance, mean cohort survival RMSE is 0.0151, versus 0.0535 at zero borrowing and 0.0124 at fixed $\eta = 30$. Actual choices favor strong borrowing for

Table 7: Simulation I: linear event law, fixed borrowing weights $\eta = 0, 10, 30$. Prediction entries are mean (SD), with SD below the mean; RMSE is averaged over cohort RMSEs. Width and truth coverage are cohort averages. Every row uses $R = 100$.

Student	η	C^{td}	Deviance	IBS	IBLL	RMSE	Width	Cov. (%)
		0.7520	3.4863	0.1495	0.4563	0.0285		
Linear	0	(0.0047)	(0.0279)	(0.0034)	(0.0096)	(0.0090)	0.1028	95.1
		0.7528	3.4716	0.1487	0.4538	0.0043		
Linear	10	(0.0046)	(0.0270)	(0.0032)	(0.0092)	(0.0010)	0.0319	100.0
		0.7528	3.4715	0.1487	0.4537	0.0037		
Linear	30	(0.0046)	(0.0270)	(0.0032)	(0.0092)	(0.0004)	0.0191	100.0
		0.7490	3.5042	0.1511	0.4609	0.0492		
Shared-risk NN	0	(0.0051)	(0.0280)	(0.0036)	(0.0103)	(0.0091)	0.1896	97.5
		0.7528	3.4720	0.1487	0.4538	0.0060		
Shared-risk NN	10	(0.0046)	(0.0269)	(0.0032)	(0.0092)	(0.0011)	0.0546	100.0
		0.7528	3.4716	0.1487	0.4537	0.0041		
Shared-risk NN	30	(0.0046)	(0.0269)	(0.0032)	(0.0092)	(0.0004)	0.0326	100.0
		0.7507	3.5032	0.1502	0.4585	0.0384		
MLP	0	(0.0047)	(0.0276)	(0.0035)	(0.0097)	(0.0089)	0.1500	96.7
		0.7528	3.4722	0.1487	0.4538	0.0057		
MLP	10	(0.0046)	(0.0268)	(0.0032)	(0.0092)	(0.0010)	0.0490	100.0
		0.7528	3.4717	0.1487	0.4538	0.0041		
MLP	30	(0.0046)	(0.0269)	(0.0032)	(0.0092)	(0.0004)	0.0299	100.0

Table 8: Simulation I: nonlinear event law, fixed borrowing weights $\eta = 0, 10, 30$. Prediction entries are mean (SD), with SD below the mean; RMSE is averaged over cohort RMSEs. Width and truth coverage are cohort averages. Every row uses $R = 100$.

Student	η	C^{td}	Deviance	IBS	IBLL	RMSE	Width	Cov. (%)
		0.7424	3.5342	0.1536	0.4685	0.0792		
Linear	0	(0.0054)	(0.0318)	(0.0032)	(0.0091)	(0.0042)	0.1063	59.5
		0.7433	3.5227	0.1528	0.4665	0.0747		
Linear	10	(0.0053)	(0.0297)	(0.0032)	(0.0092)	(0.0011)	0.0328	22.3
		0.7433	3.5226	0.1528	0.4665	0.0747		
Linear	30	(0.0053)	(0.0297)	(0.0032)	(0.0092)	(0.0011)	0.0196	13.5
		0.7503	3.4980	0.1501	0.4587	0.0535		
Shared-risk NN	0	(0.0054)	(0.0338)	(0.0034)	(0.0096)	(0.0083)	0.1953	96.8
		0.7553	3.4626	0.1473	0.4505	0.0140		
Shared-risk NN	10	(0.0049)	(0.0293)	(0.0032)	(0.0091)	(0.0014)	0.0607	99.8
		0.7554	3.4620	0.1473	0.4503	0.0124		
Shared-risk NN	30	(0.0049)	(0.0292)	(0.0032)	(0.0092)	(0.0006)	0.0370	96.8
		0.7485	3.5193	0.1511	0.4617	0.0618		
MLP	0	(0.0054)	(0.0325)	(0.0034)	(0.0096)	(0.0081)	0.1664	87.8
		0.7550	3.4652	0.1475	0.4510	0.0195		
MLP	10	(0.0048)	(0.0291)	(0.0032)	(0.0091)	(0.0019)	0.0649	97.4
		0.7553	3.4623	0.1473	0.4504	0.0131		
MLP	30	(0.0048)	(0.0291)	(0.0032)	(0.0091)	(0.0009)	0.0409	96.7

compatible and mildly shifted guidance, and smaller weights under the larger calibration and risk-magnitude changes (Table 10). These are the same selected procedures reported in the main text. Numerical qualifications below accompany the recovered gains.

Table 9: **Simulation II: actual full-network PSIS-selected prediction.** Same 100 paired cohorts and reference prior. Each borrowing row selects among six weights including zero. Prediction entries are mean (SD), with SD below the mean; width averages raw 95% survival-probability interval widths. All candidates are retained.

Teacher	C^{td}	Deviance	IBS	IBLL	RMSE	Width
	0.7503	3.4980	0.1501	0.4587	0.0535	
Target only	(0.0054)	(0.0338)	(0.0034)	(0.0096)	(0.0083)	0.1953
	0.7552	3.4634	0.1474	0.4506	0.0151	
Compatible	(0.0049)	(0.0297)	(0.0032)	(0.0091)	(0.0046)	0.0584
	0.7550	3.4707	0.1481	0.4526	0.0305	
Calibration $+0.10$	(0.0050)	(0.0297)	(0.0033)	(0.0095)	(0.0036)	0.0715
	0.7524	3.4912	0.1497	0.4574	0.0505	
Calibration $+0.50$	(0.0051)	(0.0318)	(0.0035)	(0.0101)	(0.0087)	0.1702
	0.7527	3.4915	0.1494	0.4568	0.0469	
Risk magnitude	(0.0051)	(0.0344)	(0.0034)	(0.0098)	(0.0071)	0.1647
	0.7551	3.4668	0.1475	0.4510	0.0192	
Late temporal	(0.0049)	(0.0296)	(0.0032)	(0.0092)	(0.0049)	0.0717

Table 10: Full-network PSIS choices with the reference prior. Each row totals 100 cohorts.

Teacher	0	0.25	1	3	10	30
Compatible	0	0	2	19	33	46
Calibration $+0.10$	0	0	16	22	27	35
Calibration $+0.50$	4	95	1	0	0	0
Risk magnitude	2	87	11	0	0	0
Late temporal	0	0	5	38	29	28

Coverage and interval width after PSIS selection. Table 11 evaluates the saved PSIS choice separately within each cohort, using simulation truth only for evaluation. The common $\eta = 0$ comparator uses the same 100 paired cohorts. All selected fits are retained, including those with the numerical warnings documented in Appendix C.6.2. Coverage and width are unchanged when survival intervals are transformed to cumulative event-risk intervals by $1 - S$. The final-horizon column separates late-time behavior from averages over the full time grid.

Table 11: **Simulation II: empirical coverage and width at the actual PSIS choices.** Raw 95% equal-tailed survival-probability intervals, reference prior, and 100 paired cohorts. Mean coverage and width average over the same 2,000 reference patients and ten horizons within each cohort, then over cohorts. Final-horizon coverage uses interval ten. Coverage entries are percentages, with cohort Monte Carlo standard errors in parentheses; these quantify variation across cohorts, not MCMC error. RMSE is the mean of cohort RMSEs. Width and RMSE are in probability units.

Teacher / procedure	Mean coverage (%)	Mean width	RMSE	Final-horizon coverage (%)
Target only ($\eta = 0$)	96.79 (0.32)	0.1953	0.0535	96.97 (0.29)
Compatible	98.46 (0.16)	0.0584	0.0151	98.38 (0.17)
Calibration $+0.10$	79.62 (3.50)	0.0715	0.0305	61.18 (3.62)
Calibration $+0.50$	94.17 (0.78)	0.1702	0.0505	96.02 (0.72)
Risk magnitude	94.84 (0.70)	0.1647	0.0469	95.71 (0.64)
Late temporal	95.48 (0.63)	0.0717	0.0192	83.18 (2.45)

D.4 Independent verification of center, spread, and risk

This comparison uses the correctly specified linear event law and 13-parameter linear student from Simulation I, with its fixed teacher estimated under the same event law, reference prior, six borrowing weights, and 100 paired cohorts. All 600 fits contribute. We predict four quantities before comparing them with fitted-cohort errors: the survival-functional displacement, the MAP's sampling variance, the posterior variance, and MAP mean squared error. The predictions use the known simulation law, deployed teacher, and prior; the fitted-cohort errors do not determine their inputs or any multiplicative calibration.

Approximation retaining the finite prior. Write the linear student's quadratic prior as $P_{\sigma,\tau}(\theta) = \theta^\top \Lambda\theta/2$, with the precision Λ induced by Equation (52). At the true parameter θ_0, let $H_0 = \mathbb{E}\nabla^2 r_i(\theta_0)$ and $d_Q = \mathbb{E}\nabla q_i(\theta_0)$. With affine logits and the same risk-set cells, the likelihood and teacher Hessians coincide at any common evaluation point, even when their labels differ. A first-order solution of the penalized population score around truth is therefore

$$\delta_\eta = -\{(1+\eta)H_0 + \Lambda/n\}^{-1}\{\eta d_Q + \Lambda\theta_0/n\}, \qquad \theta_\eta^\dagger = \theta_0 + \delta_\eta. \tag{53}$$

This is one linearization of the score, not an optimized population target. We use the nonlinear functional displacement $b_{\eta,\phi} = \phi(\theta_\eta^\dagger) - \phi(\theta_0)$; the linearized displacement $g_0^\top \delta_\eta$ is retained separately. At $\theta_\eta^\dagger$, independently integrate

$$A_\eta = \mathbb{E}\nabla^2(r_i + \eta q_i) + \Lambda/n, \qquad J_\eta = \mathrm{Var}\{\nabla r_i + \eta\nabla q_i\}.$$

With $g_\eta = \nabla\phi(\theta_\eta^\dagger)$, the variance predictions are $g_\eta^\top A_\eta^{-1} g_\eta/n$ and $g_\eta^\top A_\eta^{-1} J_\eta A_\eta^{-1} g_\eta/n$. Predicted MAP MSE is the latter plus $b_{\eta,\phi}^2$. The score covariance includes complete patient contributions, preserving dependence across risk-set intervals and between likelihood and teacher scores. The deterministic prior contributes curvature but no random score covariance.

Local reference and numerical integration. For comparison, define the teacher's realized logit discrepancy $d_k(x) = \mathrm{logit}\,\widetilde{\lambda}_k(x) - \mathrm{logit}\,\lambda_{0k}(x)$. The teacher-direction score derivative at the oracle is

$$B = -\mathbb{E}\sum_k A_{ik}\lambda_{0k}(1-\lambda_{0k})d_k(X_i)z_{ik},$$

where $z_{ik} = \nabla f_{\theta,k}(X_i)$ is fixed for the linear student. Let $a_\phi = -g_0^\top H_0^{-1} B$ and $v_\phi = g_0^\top H_0^{-1} g_0$. The local reference neglecting the finite prior uses center shift $\eta a_\phi/(1+\eta)$, sampling variance $v_\phi/\{n(1+\eta)^2\}$, posterior variance $v_\phi/\{n(1+\eta)\}$, and their implied MSE. This applies a local approximation to a realized finite teacher discrepancy. A single sample size and fixed teacher do not verify the shrinking-discrepancy sequence of Corollary 5.

Two independently scrambled Sobol panels, with 2^{16} and 2^{18} covariate values, integrate these inputs. Conditional on each covariate, we enumerate every possible observed event and censoring outcome under the declared event-before-censor convention. The oracle information identity and the original patient-score implementation agree to numerical precision. Between integration panels, the RMS change in the predicted bias vector is below 9×10^{-8}; RMS changes in functional variance and risk predictions are below 3.64×10^{-8}. These numerical differences are smaller than the observed finite-sample approximation errors. The residual population score at $\theta_\eta^\dagger$ is retained; no exact population-target coverage is claimed from this calculation.

Empirical quantities and agreement. At each fixed reference subject and horizon, MAP bias, empirical variance, and MSE are computed across the same 100 cohorts. Posterior variance is measured from saved MALA functional draws, not substituted by a second curvature approximation. Cohort-based curvature and sandwich estimates remain separate comparisons. Summaries use RMS bias and averages of functional variances or MSE over the 2,000 reference subjects and all ten horizons. MSE and posterior-variance Monte Carlo errors use cohorts as independent units; bias and empirical-variance summaries use leave-one-cohort jackknife errors. The noise-subtracted squared-bias summary is retained without clipping.

Figure 6 separates all four comparisons. Sampling variance is underestimated by approximately 4–5%, compared with empirical-variance Monte Carlo error of about 6%. Posterior variance agrees within approximately 0.1–1.1%, although discrepancies can exceed its smaller Monte Carlo error. MSE is underestimated by 4.3–10.3%; at $\eta = 10$ the difference is about 2.3 cohort Monte Carlo standard errors. The independently computed curves thus reproduce the main changes without being exact finite-sample predictions.

The finite prior matters most for center error at weaker borrowing: the local reference neglecting it predicts zero displacement at $\eta = 0$. At stronger borrowing, offsetting approximation errors can instead make that simpler reference closer in aggregate MSE; at $\eta = 30$ its predicted MSE is 1.374×10^{-5} versus observed 1.393×10^{-5}. Both approximations are shown across all weights. Both predict useful borrowing across the evaluated positive weights and continuing MSE reduction through 30. This experiment establishes neither an interior optimum beyond the grid nor a universal neural-network result.

Table 12: **Independent approximation versus observed estimation behavior.** Predicted quantities retain the specified prior. Bias and MSE columns give prediction / observation; variance columns give their ratio. MSE is scaled by 10^{-5}.

η	RMS bias (pred./obs.)	Sampling ratio	Posterior ratio	MSE (pred./obs.)
0	.00528 / .00633	.960	.999	85.14 / 88.97
.25	.00424 / .00538	.959	.998	54.93 / 57.77
1	.00301 / .00415	.957	.994	22.02 / 23.57
3	.00280 / .00352	.955	.989	6.19 / 6.84
10	.00323 / .00350	.954	.990	1.78 / 1.99
30	.00351 / .00360	.957	.989	1.32 / 1.39

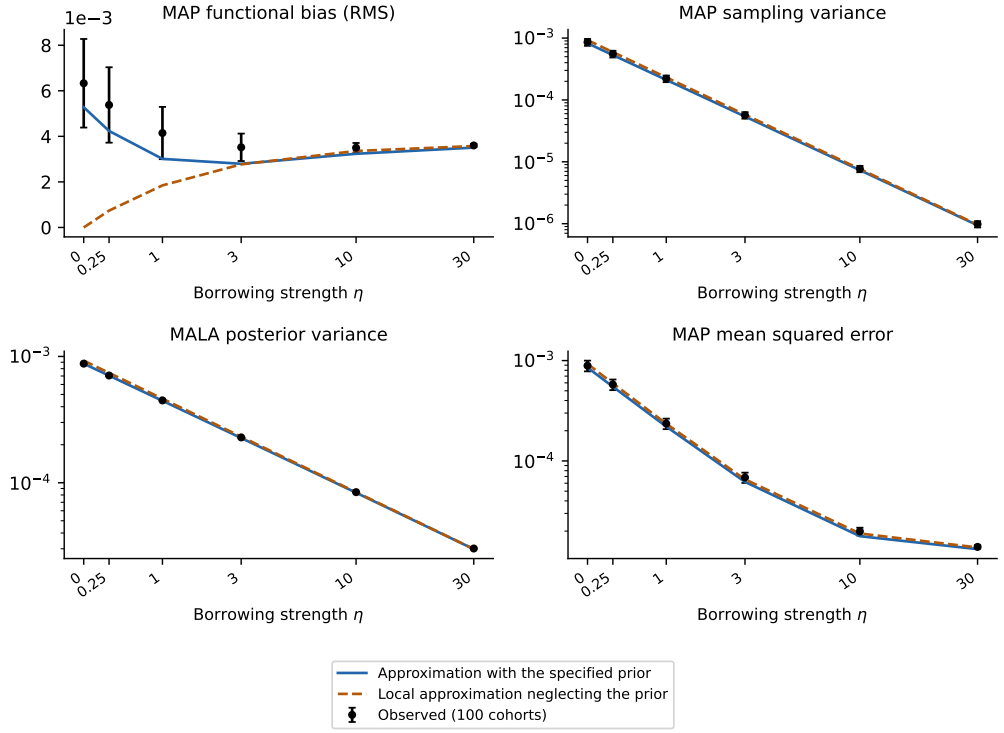


Figure 6: **Four quantities predicted independently of fitted-cohort errors.** The same estimated teacher, prior, and 100 cohorts are used at all six weights. Black points and bars show empirical summaries and 1.96 cohort Monte Carlo standard errors. Solid curves retain the specified prior; dashed curves show the local reference neglecting it. The horizontal position is $\log(1+\eta)$, labelled by η; connecting lines summarize evaluated configurations. Posterior variance is measured from MALA draws.

D.5 Finite-sample posterior-mean and local-mode comparison

What the Bayesian extension adds to DiSKD. The discrete-hazard representation and teacher-guidance loss are inherited from DiSKD. The Bayesian extension supplies a posterior over patient-specific survival probabilities, posterior-mean predictions and pointwise intervals, together with a subject-deletion predictive score for choosing borrowing strength. The existing comparisons distinguish what these additions contribute.

Comparisons of $\eta = 0$ with selected borrowing under the same prior assess the effect of adding teacher guidance within the Bayesian model. Comparisons with point DiSKD also change the prior, objective scale and fitting procedure, as specified in Appendix E; they do not isolate a prior effect or an averaging effect. At fixed teacher, prior and weight, the matched MAP and posterior mean instead summarize the same generalized posterior. MAP uses a local minimum of the penalized

distillation objective; the posterior mean averages the survival function over parameter draws. This is the squared-error summary under the fitted posterior, rather than survival evaluated at the posterior-mean parameter, and need not uniformly improve repeated-sampling RMSE.

Table 13 compares the reported posterior-mean survival function with the survival function at a matched local mode of the same objective, using identical cohorts, teacher, prior, and weight. Simulation I retains fixed weights 0 and 30; Simulation II retains its actual PSIS choices and the target-only comparator. There are 1,800 labelled cohort comparisons and 1,648 distinct configurations after reuse. This comparison isolates the predictive effect of posterior averaging within the same fitted configuration; it does not validate deletion or interval calibration. The mode is neither independently tuned nor certified to be a global optimum. Existing stationary solutions are reused, with local refinement where needed; no posterior is resampled for this comparison.

Alongside the two RMSEs, the direct functional gap is

$$D = \left\{ \frac{1}{100 \cdot 2000 \cdot 10} \sum_{r,j,k} \left[\mathbb{E}_{\Pi_r} S_\theta(t_k \mid x_j) - S_{\widehat{\theta}_r}(t_k \mid x_j) \right]^2 \right\}^{1/2}.$$

Unlike a difference of RMSEs, D measures disagreement of the predictions themselves. Similar aggregate risks do not imply identical predictions. Linear gaps are about 0.0014 without borrowing and 0.0001 at 30. Neural gaps at zero borrowing range from 0.0380 to 0.0696, falling to 0.0015–0.0079 at fixed 30. Under PSIS selection, the larger calibration and risk-magnitude regimes retain gaps of 0.0286 and 0.0292. These fixed-$n = 500$ results do not establish a vanishing asymptotic rate or uniformly negligible neural differences.

All comparisons are retained. Of the 1,648 distinct local-mode records, 47 fail the stationarity screen and 45 have an invalid local Hessian; the latter are within the former group. Their predictions remain displayed and the screen failures are counted by row. These local-optimization qualifications are separate from the posterior sampling warnings in Appendix C.6.2. The table therefore documents the finite-sample connection and its limits, while Proposition 1 states sufficient conditions for the asymptotic connection.

Table 13: Matched posterior-mean and local-mode survival predictions. The two RMSE columns are means (SDs) over 100 cohorts; D is the pooled direct prediction gap. "NS" counts configurations failing the local-mode stationarity screen, retained in every summary. Sim I settings specify event law/student/weight; Sim II uses the nonlinear shared-risk student and its actual PSIS choice.

Study	Setting	Posterior-mean RMSE	Local-mode RMSE	D	NS
Sim I	Linear/Linear/0	0.0285 (0.0090)	0.0284 (0.0090)	0.0014	0
Sim I	Linear/Linear/30	0.0037 (0.0004)	0.0037 (0.0004)	0.0001	0
Sim I	Linear/Shared-risk NN/0	0.0492 (0.0091)	0.0680 (0.0155)	0.0380	0
Sim I	Linear/Shared-risk NN/30	0.0041 (0.0004)	0.0039 (0.0004)	0.0015	8
Sim I	Linear/MLP/0	0.0384 (0.0089)	0.0691 (0.0087)	0.0537	1
Sim I	Linear/MLP/30	0.0041 (0.0004)	0.0041 (0.0004)	0.0016	29
Sim I	Nonlinear/Linear/0	0.0792 (0.0042)	0.0792 (0.0042)	0.0014	0
Sim I	Nonlinear/Linear/30	0.0747 (0.0011)	0.0747 (0.0011)	0.0001	0
Sim I	Nonlinear/Shared-risk NN/0	0.0535 (0.0083)	0.0742 (0.0126)	0.0400	0
Sim I	Nonlinear/Shared-risk NN/30	0.0124 (0.0006)	0.0120 (0.0007)	0.0036	1
Sim I	Nonlinear/MLP/0	0.0618 (0.0081)	0.0738 (0.0091)	0.0696	1
Sim I	Nonlinear/MLP/30	0.0131 (0.0009)	0.0131 (0.0007)	0.0079	6
Sim II	Target only	0.0535 (0.0083)	0.0742 (0.0126)	0.0400	0
Sim II	Compatible	0.0151 (0.0046)	0.0149 (0.0054)	0.0060	0
Sim II	Calibration $+0.10$	0.0305 (0.0036)	0.0313 (0.0055)	0.0087	0
Sim II	Calibration $+0.50$	0.0505 (0.0087)	0.0632 (0.0110)	0.0286	0
Sim II	Risk magnitude	0.0469 (0.0071)	0.0607 (0.0097)	0.0292	0
Sim II	Late temporal	0.0192 (0.0049)	0.0195 (0.0058)	0.0072	1

D.6 Spread and coverage under fitted guidance

D.6.1 Linear-student borrowing paths

Under the linear event law, the 13-parameter linear student samples every parameter. We examine the fixed reference prior $(\sigma, \tau) = (10, 0.25)$ and all six borrowing weights over the same 100 cohorts, using the same-law estimated teacher from Simulation I. These fixed-configuration paths provide a matched comparison of posterior and sampling spread.

Posterior-mean RMSE decreases from 0.02985 at $\eta = 0$ to 0.00828 at $\eta = 3$ and 0.00374 at $\eta = 30$. At $\eta = 0, 0.25, 1, 3, 10, 30$, the ratio of average posterior variance to the posterior mean's empirical variance is 1.024, 1.282, 2.036, 4.044, 10.950, and 30.579, respectively. These values are close to $1 + \eta$, which is an analytical reference for exact oracle guidance; the fitted teacher here is estimated. Sandwich variance divided by the MAP's empirical variance remains between 0.957 and 0.959. All three covariance summaries concern the same fully sampled linear student.

Table 14 expresses the same saved variance summaries relative to $\eta = 0$, directly comparing posterior contraction with $(1 + \eta)^{-1}$ and local-MAP sampling contraction with $(1 + \eta)^{-2}$. The posterior variance is measured from functional draws; the sampling variance is measured across the matched MAP estimates. Their different rates agree closely with the benchmark despite the estimated teacher and finite prior. The posterior-mean/MAP connection is assessed separately in Appendix D.5.

At stronger borrowing, a small residual center error can exceed the reduced sampling spread. At $\eta = 30$, MAP RMS bias is 0.00360, compared with empirical SD 0.000992 and sandwich SD 0.000971. The ratio of these RMS SD summaries is 0.979, so the loss of truth coverage does not arise from a comparably large error in the aggregate spread estimate. Sandwich truth coverage is 0.383, although posterior-quantile truth coverage is 1.000. At $\eta = 3$, sandwich truth coverage

Table 14: **Variance contraction relative to target-only inference.** Same six-weight, 100-cohort Linear control. Entries are ratios of average functional variances to their $\eta = 0$ values; sampling variance concerns the MAP. Benchmarks are the exact-guidance leading rates, not a refit or a calibrated correction.

	Posterior variance		MAP sampling variance	
η	Observed ratio	$(1+\eta)^{-1}$	Observed ratio	$(1+\eta)^{-2}$
0	1.00000	1.00000	1.00000	1.00000
0.25	0.80539	0.80000	0.64592	0.64000
1	0.51144	0.50000	0.25721	0.25000
3	0.26063	0.25000	0.06596	0.06250
10	0.09605	0.09091	0.00896	0.00826
30	0.03437	0.03226	0.00115	0.00104

is 0.925 (cohort-level MCSE 0.0087). Corollary 4 assumes exact guidance; the present estimated teacher retains a center discrepancy as sampling variability contracts. The finite-prior calculation in Appendix D.4 independently predicts RMS bias 0.00351 at $\eta = 30$, close to the observed 0.00360. The pointwise bias–spread approximation predicts coverage 0.385 versus observed 0.383 (Table 15). These checks support the separate roles of center error and spread. The aggregate SD ratio alone is not a pointwise coverage calculation.

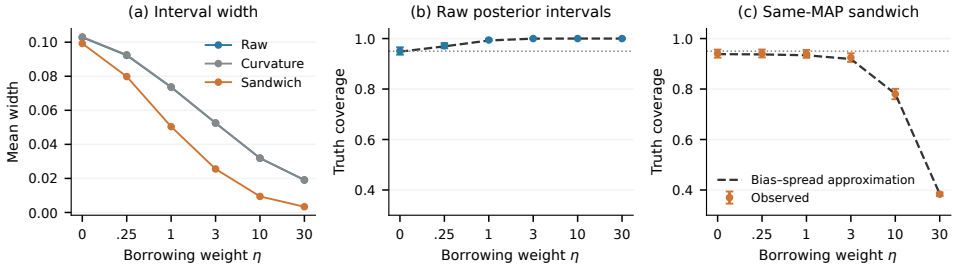


Figure 7: **Spread adjustment and predicted coverage under estimated guidance.** The fully sampled Linear student uses the same 100 cohorts and estimated teacher. (a) Mean width; curvature and sandwich share a MAP center, whereas raw intervals use posterior quantiles. (b,c) Observed truth coverage and the Gaussian approximation formed pointwise from empirical bias and sampling/reported SD, then averaged over subjects and horizons. Bars show 1.96 cohort-level Monte Carlo standard errors. At $\eta = 30$, sandwich coverage is 0.383 versus the approximation 0.385: residual center error dominates the contracted sampling spread. This descriptive approximation uses repeated-cohort information; the exact-oracle guarantee in Corollary 4 concerns exact guidance.

D.6.2 Quantitative assessment of the coverage approximation

We assess Equation (37) along the six-weight linear-student path over 100 cohorts. At each fixed reference subject and horizon, the Gaussian approximation combines estimated center bias, empirical SD, and RMS reported SD before averaging. A complementary split-cohort calculation estimates these terms on one half of the cohorts and compares its approximation with inclusion on the other, then reverses the halves. The split column averages the two predicted coverages; equally averaging inclusion over the two held-out halves gives the full-sample observed column. All 600 existing fits are retained. Table 15 shows that the approximation tracks the decline in sandwich truth inclusion as the estimated teacher's residual bias becomes large relative to sampling spread.

The neural selected procedure relearns its representation across cohorts and has no matched complete-procedure population-target calculation here. Its empirical results therefore do not extend this fixed-configuration coverage approximation into a post-selection calibration theorem.

Table 15: Gaussian coverage approximation versus observed sandwich truth inclusion for the linear student, with the reference prior and 100 cohorts. "Split" averages the two approximations, each estimating bias and both spreads from the same half of the cohorts.

η	Approximation	Observed	Split
0	.938	.940	.937
0.25	.937	.941	.936
1	.934	.939	.933
3	.919	.925	.918
10	.781	.780	.780
30	.385	.383	.385

D.7 Teacher-discrepancy sensitivity

At fixed data, prior, grid and borrowing weight, Proposition 6 describes the local MAP-functional response to a prescribed teacher perturbation; Corollary 7 gives the posterior-mean response. A descriptive scale for that change is

$$\frac{\left|\mathbb{E}_{\Pi_{\eta,\varepsilon}}\phi - \mathbb{E}_{\Pi_{\eta,0}}\phi\right|}{\mathrm{SD}_{\Pi_{\eta,0}}(\phi)}. \tag{54}$$

The direction and amplitude define a sensitivity scenario; they do not estimate unknown teacher bias. The calculation below checks population-MAP derivatives under a fixed representation. A matched full-network check of posterior-mean perturbations is not part of these results.

D.7.1 Population response to prescribed teacher discrepancies

We evaluate the penalized population MAP under a fixed shared-risk representation, with $n = 500$ and $(\sigma, \tau) = (10, 0.25)$. The representation is learned on an independent 500-subject development cohort under the nonlinear event law, using the compatible estimated teacher, $\eta = 3$, and the reference prior in the penalized likelihood-plus-guidance objective. It is then held identical across every population configuration. The grid crosses $\eta \in \{0, 0.25, 1, 2, 3, 10, 30\}$ with calibration offsets $\{0, 0.05, 0.10, 0.25, 0.50\}$ or late temporal amplitudes $\{0, 0.025, 0.05, 0.10, 0.15\}$. Shared baselines give 55 distinct configurations. Two independent integration panels check each solution: the largest survival difference is 1.01×10^{-4} and the largest remaining Newton-implied functional step is below 10^{-6}. The same 2,000 reference patients are used throughout.

Figure 8 displays the two paths in Equation (32) at interval ten. Positive calibration and late temporal perturbations lower mean interval-ten survival, and their effects increase with borrowing across the evaluated grid. Composite trapezoidal integration of the borrowing derivative $\partial_\eta \phi$ at $\eta = 1, 2, 3$ reconstructs the change from 1 to 3 with relative RMS errors of 7.3% for calibration $+0.10$, 8.0% for calibration $+0.50$, and 6.7% for temporal amplitude 0.15. Each percentage divides RMS reconstruction residual by RMS actual change over the fixed reference patients and all ten horizons. This checks the borrowing derivative along the evaluated population path. At $\eta = 3$, the baseline teacher derivative predicts the small calibration shift $+0.10$ with RMS residual 0.00042 against actual displacement 0.01693; for the larger shift $+0.50$, the residual grows to 0.01050 against displacement 0.08736. The local response thus captures small perturbations more closely than large ones. Early-horizon responses differ because the student shares parameters across time.

The response quantifies displacement from a declared reference teacher. Its baseline center can already differ from truth. These fixed-representation calculations therefore explain a component of the borrowing trade-off; the repeated-cohort prediction results assess its net estimation benefit.

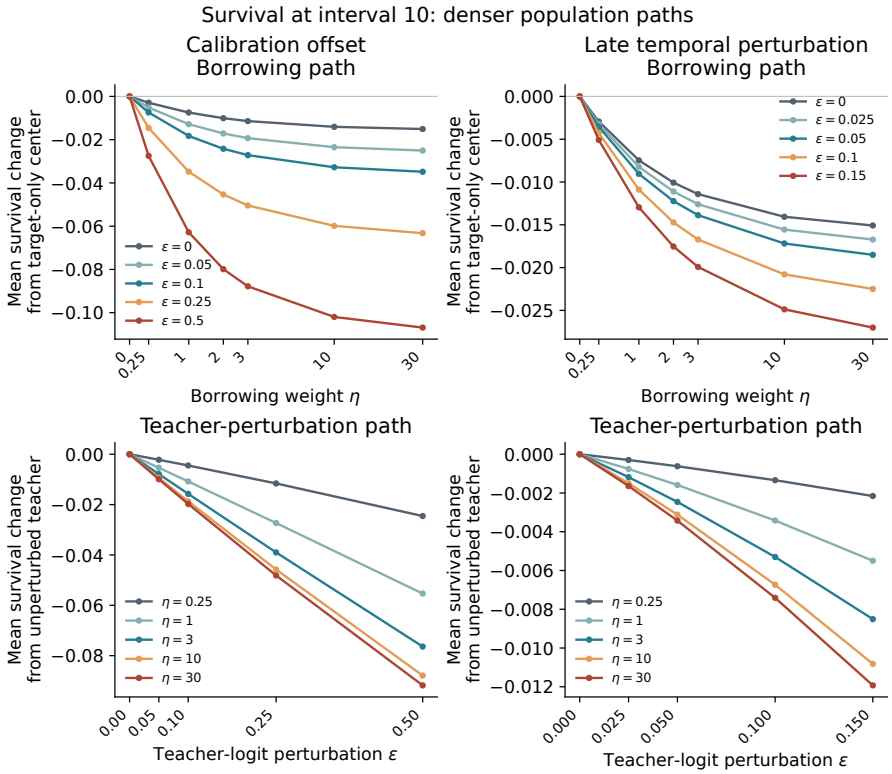


Figure 8: **Population survival response under fixed representation.** All panels show mean survival changes at interval ten. Top: change from the target-only population center as borrowing increases, at fixed teacher perturbation. Bottom: change from the unperturbed teacher's population center at the same borrowing weight. Each point is a penalized population MAP calculation; connecting lines join evaluated configurations. These are separate from the refitted-representation posterior means in Simulation II.

D.7.2 Finite-mismatch bias and variance in Simulation II

Table 16 separates the center and spread components of the current full-network calibration paths, using the same 100 paired cohorts, reference prior, and fixed reference grid as Figure 2. Every displayed weight is fixed across cohorts; the table does not substitute a modal choice for the selected procedure. Bias and sampling SD are computed separately at each patient and horizon before RMS aggregation, and RMSE pools squared errors over cohorts and reference cells.

With calibration $+0.10$, increasing borrowing reduces sampling variation enough to offset the increased center bias, leaving RMSE below target-only throughout the grid. With calibration $+0.50$, $\eta = 0.25$ still improves RMSE, but larger weights increase bias faster than the reduction in sampling variation can compensate. Posterior SD continues to decrease in both settings. This is the finite-sample bias–variance pattern motivating Section 4.2. The fixed teacher shifts, trained representation and finite prior differ from the local $n^{-1/2}$ sequence, so the table does not validate its numerical optimal weight or asymptotic boundary. The prescribed population-response calculation above separately checks how teacher perturbations move the center. Existing sampling qualifications in Appendix C.6.2 apply to these full-network summaries.

D.8 High-borrowing extension

Figure 9 uses the original first three cohorts, both event laws and all three students, comparing $\eta = 200$ with their matching full-state fits at 0, 10 and 30. The reference prior, teacher, sampler and target data are unchanged. All 18 high-weight fits and 54 comparator records are retained. This separate pilot does not enter the main PSIS candidate sets.

Table 16: **Full-network Simulation II: center and spread under finite calibration mismatch.** All six weights and both calibration amplitudes, with 100 paired cohorts and fixed reference prior. Bias is total posterior-mean error relative to truth, retaining baseline error. EmpSD is the posterior mean's repeated-cohort SD; PostSD is posterior functional SD. Both SD columns use RMS aggregation. Pooled RMSE differs from the mean of cohort RMSEs in the main prediction tables. All values are in probability units.

Calibration shift	η	RMS bias	EmpSD	PostSD	Pooled RMSE
+0.10	0	0.0182	0.0513	0.0605	0.0542
+0.10	0.25	0.0182	0.0406	0.0537	0.0443
+0.10	1	0.0219	0.0258	0.0424	0.0337
+0.10	3	0.0258	0.0141	0.0309	0.0294
+0.10	10	0.0285	0.0068	0.0195	0.0293
+0.10	30	0.0294	0.0039	0.0122	0.0297
+0.50	0	0.0182	0.0513	0.0605	0.0542
+0.50	0.25	0.0329	0.0385	0.0527	0.0505
+0.50	1	0.0669	0.0232	0.0407	0.0708
+0.50	3	0.0937	0.0123	0.0290	0.0944
+0.50	10	0.1095	0.0060	0.0182	0.1097
+0.50	30	0.1150	0.0035	0.0114	0.1151

The linear student's error under the linear law approaches the fixed teacher's reference error, while posterior SD continues to contract. Under nonlinear generation, its error remains near 0.074, reflecting the representation limit. Neural cohort errors can improve or worsen from 30 to 200; neither neural architecture consistently dominates. Both event laws share covariate risk across intervals, so stronger guidance supplies no additional time-varying risk structure favoring the interval-specific MLP.

These curves illustrate diminishing prediction gains alongside continued posterior concentration. They remain a three-cohort finite-sample check, separate from the fixed-finite-weight local theorem and the main selector. All 12 neural high-weight fits retain sampling-review flags, and 8 of 18 high-weight fits have importance warnings; the pilot does not establish a universal saturation threshold or coverage.

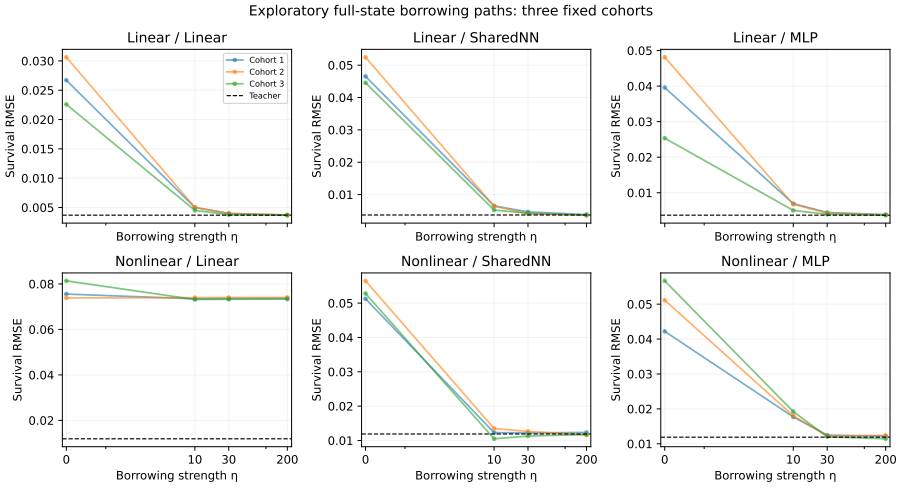


Figure 9: **High-borrowing behavior in three matched cohorts.** Columns identify students and rows identify event laws. Curves follow the same three cohorts across $\eta = 0, 10, 30, 200$; dashed lines give the corresponding fixed-teacher survival RMSE. These full-state posterior-mean comparisons are separate from the 100-cohort main study. The borrowing axis equally spaces evaluated weights.

D.9 Paired prior sensitivity in Simulation II

An earlier conditional-head experiment compares $(\sigma, \tau) = (10, 0.25)$, $(1, 0.25)$ and $(10, \infty)$ on the same 100 target cohorts, nonlinear event law, fixed teachers and 2,000-patient/ten-horizon reference grid. Representations are fitted separately for each configuration and held fixed during posterior sampling. The first alternative strengthens Gaussian shrinkage of output weights while retaining temporal smoothing; the second removes temporal smoothing while retaining the Gaussian scales. Intercept and representation prior scales remain fixed. These are conditional-head sensitivity results, separate from the full-network main comparisons.

Table 17 pairs each fixed-$\eta = 3$ borrowing configuration with its own same-prior target-only baseline and retains all five teachers. Stronger Gaussian shrinkage improves target-only RMSE from 0.0743 to 0.0512, but increases compatible-teacher RMSE at $\eta = 3$ from 0.0207 to 0.0247; borrowing still improves on that stronger baseline. Mild calibration and late-temporal guidance also improve on each same-prior baseline, whereas calibration $+0.50$ is harmful at this weight under all three priors. Thus the prior changes both baseline performance and the size of the borrowing benefit. Removing temporal smoothing has smaller effects in these displayed configurations. This comparison changes scale and smoothing, not prior centering, and does not establish robustness of the transplant results to an event-rate-informed prior.

Table 17: **Existing paired prior sensitivity: conditional-head Simulation II.** Pooled survival RMSE and mean raw 95% interval width, in probability units, across 100 paired cohorts. Every borrowing row uses fixed $\eta = 3$ and is compared with the target-only row in the same prior column; no prior or weight is selected using these outcomes. The three-prior sweep is not full-network inference.

		$(10, 0.25)$		$(1, 0.25)$		$(10, \infty)$	
Teacher	η	RMSE	Width	RMSE	Width	RMSE	Width
Target only	0	0.0743	0.1335	0.0512	0.1121	0.0767	0.1362
Compatible	3	0.0207	0.0653	0.0247	0.0615	0.0209	0.0659
Calibration $+0.10$	3	0.0299	0.0641	0.0331	0.0605	0.0298	0.0647
Calibration $+0.50$	3	0.0946	0.0597	0.0964	0.0566	0.0942	0.0603
Risk magnitude	3	0.0646	0.0593	0.0640	0.0556	0.0645	0.0600
Late temporal	3	0.0218	0.0650	0.0258	0.0613	0.0223	0.0655

E Real-Data Cohorts, Prediction and Uncertainty

E.1 Shared model and evaluation specifications

E.1.1 Cohorts and model specification

METABRIC uses nine released benchmark predictors. Each of three event-stratified partitions has 1,446 source-training, 77 target-training and 381 test participants; target-training deaths through 120 months number 29, 35 and 31. The NWTCO target-size sensitivity is specified in Appendix E.4.

NWTCO uses age, categorical stage and local-institution histology, giving five encoded predictors; central-laboratory histology is excluded. Its released outcome is recorded relapse with observation time. Death before relapse is not separately identified. Administrative horizons are 120 months for METABRIC and 1,826.25 days for NWTCO. These are single-event analyses.

For METABRIC and NWTCO, source teachers use a shared nonlinear risk score with hidden widths 64 and 32 and interval-specific intercepts. A source-only validation split selects Gaussian regularization before refitting the teacher. Source and target transformations are fitted to their respective training records. Students have a 16-unit SiLU hidden layer, scaled by $\sqrt{16}$, and $K = 10$ interval-specific outputs. Full-network inference samples all 330 METABRIC or 266 NWTCO parameters, including hidden weights and biases.

The METABRIC pooled reference fits the same student architecture to source plus target-training outcomes, with $\eta = 0$ and the reference prior. It retains the target-training transformation, time grid and temporal-prior covariate panel, so the reference comparison does not change the student

architecture or prior design. The common reference for all candidate priors within a partition is this reference-prior posterior mean. It is an estimated high-information fit, not a true survival law; test outcomes enter neither its fitting nor teacher training.

Both datasets use Gaussian body and interval-intercept scales 1 and 10. The reference temporal prior uses output-weight SD 10 and $\tau = 0.25$. The METABRIC prior comparison additionally considers output-weight SD 1 and, separately, no temporal smoothing ($\tau = \infty$) at output-weight SD 10; NWTCO retains the reference prior. Temporal smoothing averages adjacent-logit penalties over fixed target-training covariates, differentiating through body and head. This design-conditional prior panel stays fixed during predictive subject deletion.

For METABRIC and NWTCO, source-event quantiles define a $K = 20$ grid between zero and the administrative horizon; retaining every second boundary gives the student grid. Teacher hazards are coarsened by multiplying within-block survival probabilities. Censoring within an occupied interval is encoded as a non-event for that interval. This is a grouped approximation for continuous follow-up, not the native-discrete observation law used in simulation.

E.1.2 Selection and point comparators

For METABRIC, PSIS jointly selects a prior and $\eta \in \{0, 0.25, 1, 3, 10, 30\}$ from target-training outcomes. Deletion removes the subject's likelihood and observed-risk-set teacher loss, holding teacher, prior, transformations and grid fixed. The score uses the deleted outcome likelihood alone. Internal Bayes separately selects a prior at $\eta = 0$; matched-prior controls instead keep the borrowing-selected prior. Prior-specific sensitivity applies the same PSIS criterion within each prior.

Likelihood-only WAIC is evaluated on the same posterior candidate menu. A bounded threefold CV comparison on METABRIC partition 2 refits all 18 prior–weight candidates in each training fold, omitting validation contributions from both losses. Standardization and the temporal-prior covariate panel are fitted within each fold; the external teacher and source-derived time grid remain fixed. Validation outcome log predictive density chooses the candidate, which is evaluated after full target-training refitting. This refitting comparison is distinct from conditional subject deletion.

The METABRIC point DiSKD comparator minimizes the subject-average $(r_i + \eta q_i)/(1 + \eta)$, up to teacher-only constants, with AdamW and weight decay 10^{-4}. Fivefold target-training outcome NLL selects learning rate from $\{5 \times 10^{-4}, 10^{-3}\}$ and batch size from $\{32, 128\}$ at $\eta = 0$, then selects borrowing with those optimizer settings fixed. Fits use at most 128 epochs and patience 5; final fitting uses the ceiling of the median stopping epoch. These controls match student architecture, not the original paper's Transformer search or the Bayesian prior. The MAP under the selected Bayesian objective is retained separately.

E.1.3 Evaluation definitions

On METABRIC and NWTCO, constant-hazard interpolation converts interval probabilities to native-time survival curves before posterior averaging. RMST integrates each interpolated curve analytically. Concordance uses Antolini's event-time comparison through the administrative horizon. IBS and integrated binomial log loss (IBLL) use 201 equally spaced times through the common training-supported endpoints, 112.227 months on METABRIC and 963.6 days on NWTCO. Support requires training censoring survival at least 0.05. All methods within a partition share censoring weights estimated from target-training data; marginal IPCW requires marginal independent censoring. NWTCO uses the common full target-training censoring estimate across subset sizes, as detailed in Appendix E.4.

METABRIC reference survival RMSE equally weights test subjects and 201 times through 120 months. RMST MAE averages absolute error against the pooled posterior mean over subjects. Figure 3b pairs reference RMSE with mean 95% pointwise interval width on this same subject/time grid. Each arrow is a model comparison, not a patient or independent replicate. Panel (c) reports the proportion with both smaller absolute five-year reference error and a narrower 95% interval, using all 381 test patients in each partition as the denominator. Table 18 complements this proportion with $\Delta W_i = W_{i,0}(60) - W_{i,\eta}(60)$, the paired reduction in five-year interval width. Its means and sample SDs include all patients, including negative changes; SD describes variation in paired differences across patients, not uncertainty about the mean. NWTCO width ratios divide the selected mean

one-year survival-probability width by that of the same-prior $\eta = 0$ posterior. Neither quantity is event-time predictive coverage.

Mean and sample SD across the three overlapping METABRIC partitions describe variability, not independent-cohort standard errors. NWTCO subsets share the test cohort and are reported separately. Although candidate selection uses training outcomes, prior inspection of these datasets makes the applications exploratory rather than independent validation of the complete development process. No posterior temperature, conformal correction or interval rescaling is applied.

E.2 METABRIC: PREDICTION AND REFERENCE RECOVERY

All panels of the METABRIC main figure hold the PSIS-selected no-temporal prior fixed for the borrowing comparison; independently prior-selected target-only Bayes remains in Table 19. Figure 10 retains all three priors and separates partitions; Figure 11 gives complete risk curves for three prespecified METABRIC patients.

Table 18: **Paired five-year interval-width reductions on METABRIC.** Mean (patient SD) of $100(W_0 - W_\eta)$ for 95% survival-probability intervals, in percentage points. Each comparison uses PSIS-selected borrowing and its matched-prior target-only model, with all test patients included regardless of the sign of the change. Partitions are summarized separately because they overlap.

Partition	Test patients	Width reduction: mean (SD)
P1	381	60.7 (20.1)
P2	381	57.3 (23.8)
P3	381	61.4 (23.3)

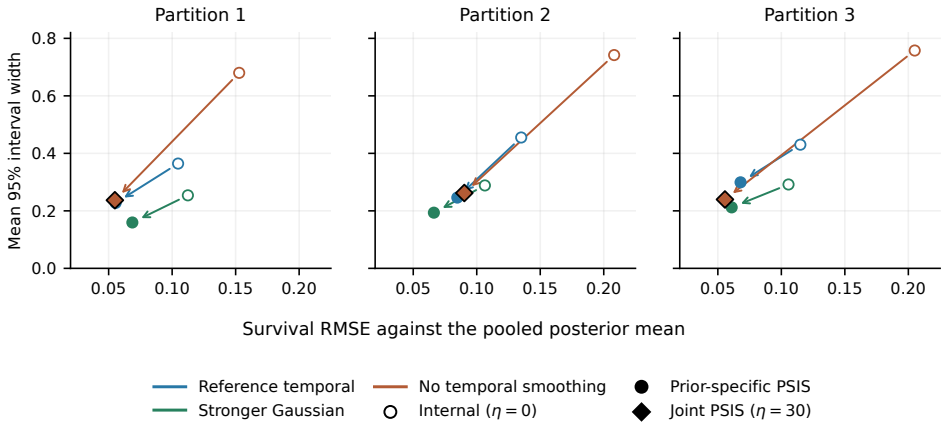


Figure 10: **METABRIC prior sensitivity.** Open circles denote same-prior target-only models, filled markers prior-specific PSIS choices, and diamonds joint PSIS choices. Each panel is one partition; all evaluations use 381 held-out subjects and 201 time points. RMSE is relative to the matched-student pooled posterior mean.

Table 19 retains per-partition outcomes for every primary METABRIC comparator. Table 20 separates joint selection from prior-specific choices. PSIS selects $\eta = 30$ without temporal smoothing in all partitions, whereas WAIC chooses no borrowing. The bounded CV comparison also selects $\eta = 30$, with stronger Gaussian regularization: its IBS is close to PSIS, with higher concordance and smaller reference error. These outcomes do not establish superiority on every metric. Selections at $\eta = 30$ reach the evaluated grid boundary, not an unrestricted optimum.

Point DiSKD, matched MAP, and posterior averaging. Bayesian DiSKD improves on the matched-architecture point DiSKD comparator in partitions 1 and 2: NLL changes from 1.4657 to 1.4284 and from 1.5699 to 1.5515, and IBS from 0.12610 to 0.11932 and from 0.13523 to 0.13023,

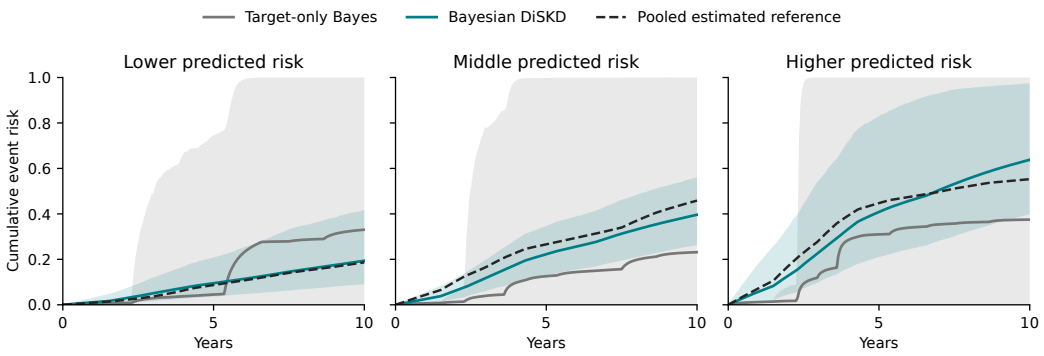


Figure 11: **Individual METABRIC risk trajectories.** Patients at the 10th, 50th and 90th percentile ranks of teacher five-year risk in partition 1, chosen without reference to their outcomes or borrowing benefit. Lines show posterior means and shading pointwise 95% intervals. The dashed pooled posterior mean is an estimated reference.

respectively. Both metrics worsen in partition 3. The matched MAP under the selected Bayesian objective already improves both metrics over point DiSKD in partitions 1 and 2, so these gains cannot be attributed entirely to posterior averaging. Differences between point DiSKD and this MAP combine regularization, objective scale, optimization and selection; the comparison does not separate their individual contributions.

Posterior averaging is evaluated directly against the matched MAP. Reference RMSE decreases from 0.06610 to 0.05492 in partition 1 and from 0.05704 to 0.05560 in partition 3, but increases from 0.05399 to 0.09023 in partition 2. The matched simulation comparison likewise shows averaging gains under larger calibration and risk-magnitude discrepancies, with similar MAP and posterior-mean errors under compatible guidance (Table 13). Thus the extension provides probability intervals and an averaged prediction, with an empirical point-prediction benefit that depends on the setting. METABRIC reference errors assess agreement with the pooled fit, not truth coverage.

Table 19: **Per-partition METABRIC outcomes.** NLL is mean grouped-time negative log predictive density. Reference errors use the partition-specific matched-student pooled posterior mean; RMST error is in months. Internal Bayes independently selects its prior, and Bayesian DiSKD uses joint PSIS.

Partition	Method	NLL	IBS	IBLL	C^{td}	Ref. RMSE	RMST MAE
1	Source teacher	1.4333	0.11925	0.3773	0.6682	0.06882	4.615
1	Pooled reference	1.4189	0.11869	0.3757	0.6609	0.00000	0.000
1	Internal point	1.4793	0.12765	0.4022	0.4793	0.13224	10.255
1	Point DiSKD	1.4657	0.12610	0.3959	0.6289	0.09809	7.564
1	Internal Bayes	1.7562	0.14002	0.4275	0.6213	0.15279	10.557
1	DiSKD MAP	1.4286	0.11856	0.3753	0.6643	0.06610	4.331
1	Bayesian DiSKD	1.4284	0.11932	0.3779	0.6702	0.05492	4.210
2	Source teacher	1.5427	0.12988	0.3977	0.6593	0.05076	3.382
2	Pooled reference	1.5240	0.12925	0.3958	0.6486	0.00000	0.000
2	Internal point	1.6052	0.14239	0.4332	0.6116	0.10973	9.217
2	Point DiSKD	1.5699	0.13523	0.4124	0.6360	0.08349	6.582
2	Internal Bayes	1.5847	0.13819	0.4217	0.6065	0.10640	8.917
2	DiSKD MAP	1.5379	0.12912	0.3960	0.6594	0.05399	3.811
2	Bayesian DiSKD	1.5515	0.13023	0.4000	0.6457	0.09023	6.194
3	Source teacher	1.4857	0.12677	0.3938	0.6234	0.05830	3.936
3	Pooled reference	1.4694	0.12505	0.3876	0.6294	0.00000	0.000
3	Internal point	1.5729	0.12856	0.4033	0.5776	0.10935	8.988
3	Point DiSKD	1.4781	0.12649	0.3946	0.6098	0.06594	5.170
3	Internal Bayes	1.9238	0.17347	0.5173	0.5728	0.20501	16.045
3	DiSKD MAP	1.4904	0.12854	0.3974	0.6164	0.05704	4.054
3	Bayesian DiSKD	1.4911	0.12939	0.4010	0.6074	0.05560	4.222

Table 20: **METABRIC selector and prior sensitivity.** All choices use training data. The fixed-prior PSIS rows supplement rather than replace joint selection. CV refits the same prior–weight menu in three folds for partition 2 only.

Partition	Selector	Prior	$\widehat{\eta}$	IBS	C^{td}	Ref. RMSE
1	PSIS	No temporal	30	0.11932	0.6702	0.05492
1	WAIC	No temporal	0	0.14002	0.6213	0.15279
1	PSIS (fixed prior)	Reference temporal	30	0.11825	0.6722	0.05552
1	PSIS (fixed prior)	Stronger Gaussian	30	0.11874	0.6675	0.06867
2	PSIS	No temporal	30	0.13023	0.6457	0.09023
2	WAIC	No temporal	0	0.16413	0.5886	0.20819
2	CV	Stronger Gaussian	30	0.13094	0.6543	0.06468
2	PSIS (fixed prior)	Reference temporal	30	0.12946	0.6521	0.08469
2	PSIS (fixed prior)	Stronger Gaussian	10	0.13083	0.6553	0.06624
3	PSIS	No temporal	30	0.12939	0.6074	0.05560
3	WAIC	No temporal	0	0.17347	0.5728	0.20501
3	PSIS (fixed prior)	Reference temporal	10	0.13263	0.5956	0.06787
3	PSIS (fixed prior)	Stronger Gaussian	10	0.12990	0.5965	0.06089

E.2.1 Reference inclusion

Equal-tailed intervals use credibility levels $0.50, 0.65, 0.80, 0.90, 0.95, 0.975, 0.99$. Survival reference inclusion and width average over matched subject/time pairs; RMST summaries average over subjects. Width comparisons interpolate within overlapping achieved inclusion ranges, without extrapolation or rescaling.

At 95% credibility, joint PSIS includes the pooled survival reference for approximately 90.2%, 85.5% and 91.3% of subject/time pairs across partitions. Despite lower reference RMSE and narrower intervals, inclusion does not uniformly improve over independently selected internal Bayes. This measures recovery of the fitted reference, rather than truth coverage.

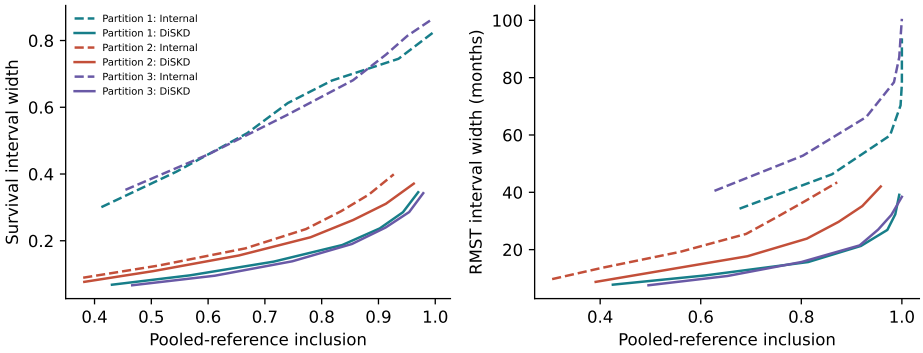


Figure 12: **METABRIC pooled-reference inclusion and width.** The left panel summarizes survival probabilities and the right panel summarizes 120-month RMST. Colors distinguish the three partitions; solid curves show joint PSIS-selected DiSKD and dashed curves independently selected internal Bayes. Curves trace achieved inclusion of the fitted pooled posterior mean across credible levels. Width comparisons are confined to overlapping achieved inclusion ranges. The reference is estimated, and the horizontal coordinate is not individual truth coverage.

E.3 Transplant prediction and uncertainty

The kidney-alone to kidney–pancreas analysis retains its source-trained teacher, training-only transformations and 20-interval time grid. The target split contains 1,935 training and 484 test recipients; the composite death-or-kidney-graft-failure endpoint has 83 and 21 events, respectively. The student uses 127 encoded predictors, one 16-unit SiLU hidden layer and 20 output logits, giving 2,388

parameters. All parameters, including intercepts, have independent $N(0, 1)$ priors; no temporal penalty is applied. Four full-network MALA chains per candidate use 16,000 warm-up and 128,000 production transitions, retaining 4,000 draws per chain. Survival is a right-continuous step function on the original grid. IPCW scores use target-training censoring estimates and 201 evaluation times through 329.8 days; concordance and interval widths retain the one-year administrative horizon.

Table 21 reports every evaluated posterior-mean and MAP configuration. The MAP uses the same generalized posterior objective, rather than a separately tuned point-estimation method. Table 22 retains the complete uncertainty path. Figure 4(b) compares the $\eta = 0$ and PSIS selected $\eta = 10$ posterior widths as a descriptive borrowing example; panel (c) reports deviance gains across all positive weights. Existing posterior-exploration and importance-tail limitations qualify fine comparisons; no numerical warning was used to remove a candidate. The displayed intervals summarize survival-probability uncertainty, not patient event-time intervals or verified truth coverage.

For Figure 4a, 2,000 paired nonparametric resamples of the 484 held-out subjects quantify conditional test-sample variability. Each resample uses the same subject multiplicities for all seven models. Trained predictions, the censoring estimator, time grid and selections remain fixed. Entries are the original test estimates with bootstrap SD, not across-fit SD or posterior Monte Carlo error. Likelihood losses and IPCW integrals are averaged over resampled subjects; concordance weights each comparable pair by the product of its subject multiplicities and retains the original event-time and prediction-tie rules. Table 21 gives all prediction metrics and their bootstrap SDs for the posterior-mean fits. The individual examples in Figure 13 use the 10th, 50th and 90th percentile ranks of teacher one-year risk, chosen before examining student intervals or outcomes. The paired width panel includes all 484 patients: 88.8% have narrower intervals with borrowing. Median one-year widths are 0.783 and 0.072 for target-only and borrowing, respectively.

Table 21: **Complete transplant prediction and conditional test-sample uncertainty.** All configurations use the same 484 held-out recipients. Teacher and posterior-mean entries are original test estimates (SD over 2,000 paired test-subject bootstrap replicates, conditional on fitted models). MAP entries are the original point estimates; corresponding bootstrap SDs were not computed. Deviance is twice mean negative log predictive density.

Method	η	IBS	C^{td}	Deviance	IBLL
Source teacher	–	0.03274 (0.0070)	0.6106 (0.067)	0.6145 (0.106)	0.1449 (0.024)
Posterior mean	0	0.03395 (0.0070)	0.6138 (0.055)	0.7449 (0.140)	0.1525 (0.027)
Posterior mean	0.3	0.03342 (0.0070)	0.5861 (0.057)	0.6467 (0.113)	0.1492 (0.025)
Posterior mean	1	0.03306 (0.0070)	0.5862 (0.059)	0.6274 (0.107)	0.1471 (0.024)
Posterior mean	3	0.03273 (0.0070)	0.6118 (0.063)	0.6184 (0.105)	0.1449 (0.024)
Posterior mean	10	0.03274 (0.0069)	0.6186 (0.057)	0.6218 (0.104)	0.1449 (0.023)
Posterior mean	30	0.03263 (0.0069)	0.6106 (0.068)	0.6128 (0.103)	0.1444 (0.023)
MAP	0	0.03854	0.5973	0.9886	0.2148
MAP	0.3	0.03724	0.5819	0.7549	0.1753
MAP	1	0.03330	0.5990	0.6588	0.1459
MAP	3	0.03319	0.6342	0.6327	0.1463
MAP	10	0.03279	0.6577	0.6382	0.1425
MAP	30	0.03215	0.6403	0.6133	0.1420

Table 22: **Mean 95% pointwise survival-probability interval widths in kidney–pancreas recipients.** Each entry averages across the same 484 test recipients.

η	30 days	90 days	180 days	One year
0	0.2131	0.3735	0.4556	0.6458
0.3	0.1504	0.2419	0.2899	0.3989
1	0.0974	0.1597	0.1944	0.2735
3	0.0557	0.0949	0.1176	0.1684
10	0.0485	0.0742	0.0869	0.1189
30	0.0327	0.0599	0.0707	0.0947

E.3.1 Individual posterior summaries

Figure 13 shows transplant recipients at the 10th, 50th and 90th percentile ranks of teacher one-year risk, fixed before inspecting outcomes or borrowing gains. A–C identify individuals. Borrowing

changes both the location and width of risk estimates; intervals need not contract for every patient. METABRIC individual trajectories are shown separately in Figure 11.

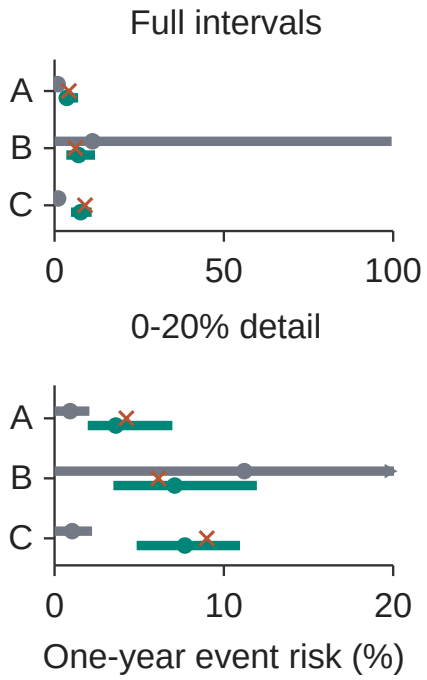


Figure 13: **Individual kidney–pancreas posterior risk intervals.** Gray denotes target-only and teal borrowing. Full one-year risk intervals are shown alongside an enlarged 0–20% view; crosses mark teacher predictions and the arrow denotes continuation beyond that view.

E.4 NWTCO: SENSITIVITY TO TARGET INFORMATION

Settings. The NWTCO cohort released with the R `survival` package records relapse and follow-up in National Wilms Tumor Studies 3 and 4 (R survival package, 2026; R cchs package, 2026). Study 3 supplies the fixed source teacher ($n = 1,857$). Study 4 supplies three nested target-training sequences with $n = 200, 400, 800$ and 27, 53, 106 events, respectively, sharing a full-training endpoint of 1,736 participants and 230 events. All comparisons use the same independent test cohort of 435 participants with 58 events within five years. The overlapping sequences are not independent population replicates.

Predictors are age, stage and local-institution histology; central-laboratory histology is excluded. The single-event outcome is recorded relapse; death before relapse is not separately identified. The $K = 10$ student has a 16-unit SiLU hidden layer and 266 parameters, all sampled. The reference temporal prior is fixed at $(\sigma, \tau) = (10, 0.25)$, with representation-weight scale one. PSIS selects among $\{0, 0.25, 1, 3, 10, 30\}$ using each subset's training outcomes. The teacher, architecture, prior and test cohort remain fixed as target-training size varies.

All sizes share IPCW scores using the full target-training censoring estimate; those additional outcomes enter scoring only, not small-subset fitting or selection. IBS uses 201 times through 963.6 days, concordance the 1,826.25-day administrative horizon, and interval widths 365.25 days. Teacher-risk terciles use full target-training covariates with cutpoints 0.0461092 and 0.0849017. The test groups contain 145, 152 and 138 patients, with 12, 7 and 13 one-year events. The 2,000 paired test-subject bootstrap resamples keep fitted models, selections, group cutpoints and training censoring estimates fixed; common resamples preserve shared-test dependence. The figure intervals describe conditional test-sample variability. These sensitivity analyses followed inspection of the test data.

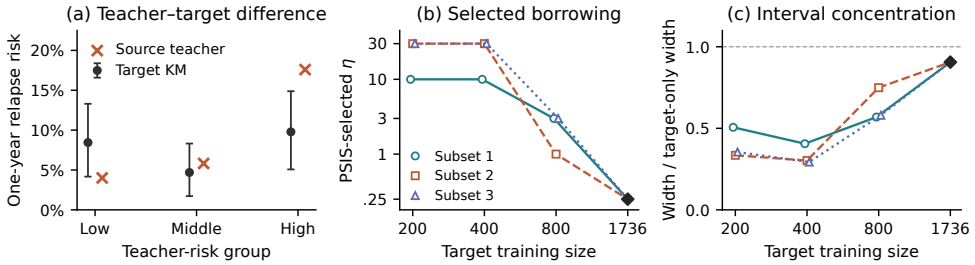


Figure 14: **Sensitivity to teacher discrepancy and target sample size.** (a) Teacher and target one-year risks ($1 - \mathrm{KM}$) in fixed teacher-risk groups; bars give pointwise 95% KM bootstrap intervals. (b,c) PSIS choices and one-year interval-width ratios versus same-prior target-only fits. Colors denote nested training sequences; black diamonds mark their shared full-size endpoint. Test patients and scoring are common; sample-size categories are equally spaced.

Teacher-risk structure and target information. Figure 14(a) shows greater separation between the teacher's low- and high-risk groups than between the corresponding target KM estimates. Across the nested training sequences, larger target samples select smaller borrowing weights (b), while the selected-to-target-only width ratio approaches one (c). Thus the posterior response to fixed external guidance depends on the amount of target information. Changing n also changes event information, cohort composition and the relative strength of the fixed prior.

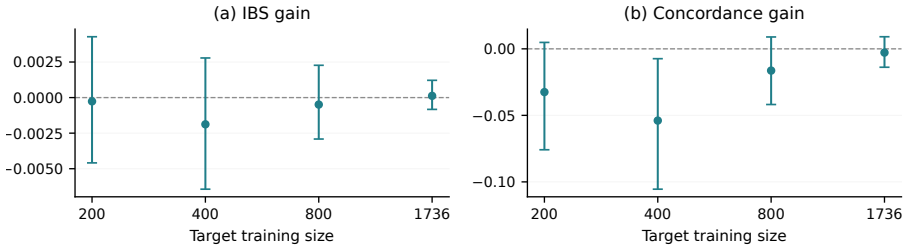


Figure 15: **Paired prediction differences across target sample sizes.** Positive favors PSIS over same-prior target-only inference; negative favors target-only inference. Points average the fixed overlapping training sequences, with one shared full-size endpoint. Bars are pointwise 95% paired test-subject bootstrap intervals conditional on fitted models and selections. All sizes share the test cohort and scoring scheme.

Figure 15 complements the interval-width view with outcome-based comparisons. Concordance most clearly separates selected and target-only predictions at $n = 400$, in favor of target-only inference; at the shared full-size endpoint their predictions are close. These paired comparisons retain the training-based PSIS choices without selecting weights from test performance.

E.5 Posterior computation and numerical limitations

Each METABRIC and NWTCO candidate uses four MALA chains with 8,000 warm-up and 256,000 sampling transitions per chain, retaining 4,000 draws per chain. Independent joint MAP starts and fixed chain-specific Gauss–Newton whitening define the proposals. The transplant computation instead uses the diagonal-preconditioned schedule described in Appendix E.3. Sampling includes the complete representation and output layers, rather than only the head.

The jointly selected fits retain functional sampling warnings, and importance-tail warnings qualify some predictive-deletion scores. We interpret interval tails and small predictive differences cautiously. Prior and borrowing-weight sensitivity describes dependence on model specification; it does not establish convergence or explain away adverse outcomes.