UNCERTAINTY-AWARE SURVIVAL DISTILLATION FOR FLEXIBLE CONTINUOUS-TIME RELATIVE RISK MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical decision-making requires accurate individualized survival predictions and clinically interpretable characterization of their uncertainty. Prominent survival-distillation approaches typically provide only point predictions. Moreover, existing continuous-time distillation methods either rely on classical regression formulations that may not extend naturally to flexible machine-learning models or transfer relative-risk information without directly incorporating external survival probabilities. To address these gaps, we introduce BayCSD (Bayesian Continuous-Time Survival Distillation), an uncertainty-aware survival-distillation framework for flexible continuous-time relative-risk models, including deep neural networks. The framework introduces a censor-aware interval KL distillation criterion that compares teacher and student models, allowing external survival predictions to inform both nonlinear relative-risk functions and cumulative baseline hazards while retaining the original event and censoring times in the target likelihood. A generalized Bayesian update combines target-data fit, a borrowing-weighted distillation discrepancy, and a structural prior, propagating joint posterior uncertainty in the risk function and baseline hazard to individualized survival probabilities and restricted mean survival time. Controlled synthetic experiments characterize the benefits and limitations of borrowing under teacher compatibility and mismatch. Real-data pooled-reference benchmarks and kidney-alone-to-kidney–pancreas transfer demonstrate improved survival prediction relative to matched target-only learning, while pooled-reference analyses yield narrower posterior intervals at matched reference inclusion. BayCSD also compares favorably with existing survival-transfer approaches in the transplant application.

1 INTRODUCTION

Reliable individualized survival prediction is difficult in target populations with few observed events, particularly for flexible models with nonlinear relationships between patient characteristics and prognosis. Patient-specific survival probabilities and time-based summaries such as restricted mean survival time, together with their uncertainty, can inform clinical decisions (Vickers & Elkin, 2006; Royston & Parmar, 2013; Van Calster et al., 2019; Begoli et al., 2019; Wey et al., 2018). Yet sparse failures limit reliable estimation of flexible models (Riley et al., 2019; Wiegrebe et al., 2024). External survival models trained on larger related populations therefore offer predictive information that may be difficult to learn from sparse target outcomes alone.

Direct reuse of such models is often limited. Source records may be unavailable due to privacy or governance (Wolfson et al., 2010); source and target populations may differ in baseline risk or covariate effects (Chen et al., 2021; Sheng et al., 2021; Cheng et al., 2023); and parameter-level transfer may be ill defined across different covariates, parameterizations, or architectures (Li et al., 2022; Tian & Feng, 2023). Prediction-level transfer instead uses external model outputs on target subjects, without requiring source records or parameter correspondence. Knowledge distillation provides a natural mechanism for such borrowing (Hinton et al., 2015; Gou et al., 2021).

For survival prediction, however, a central challenge is to borrow absolute survival-probability information without imposing the teacher's reporting grid on the target likelihood. Existing probability-

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

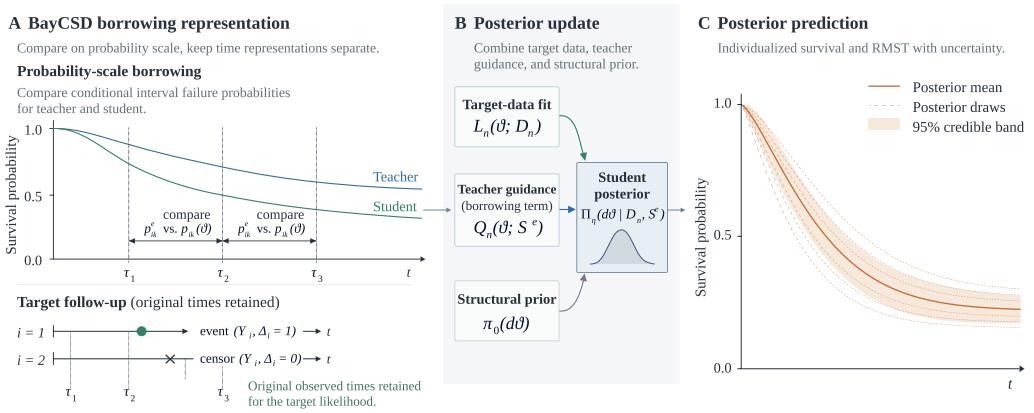


Figure 1: BayCSD uses probability-scale borrowing while keeping the comparison and target temporal representations separate. Teacher predictions contribute to posterior survival prediction together with the target-data criterion and a structural prior.

based distillation methods (Wang et al., 2025; Deng et al., 2026a) primarily couple teacher comparison to grouped-time target learning, whereas Cox risk-set formulations (Wang et al., 2026; Deng et al., 2026b) integrate teacher risk-score information without directly borrowing the teacher's baseline survival scale. A further challenge is to characterize posterior uncertainty after borrowing, beyond the point estimators targeted by these distillation objectives. Together, these considerations motivate probability-scale continuous-time survival distillation that preserves the original-time target likelihood, coupled with joint uncertainty for the risk function and baseline hazard.

We introduce **BayCSD** (**Bayesian Continuous-Time Survival Distillation**) for flexible continuous-time relative-risk models, including neural-network risk functions. Conditional interval failure probabilities define a censor-aware Kullback–Leibler divergence measuring the discrepancy between teacher and student models, while the target likelihood retains the original event and censoring times. The comparison partition can differ from the student's baseline-hazard partition: for example, annual teacher predictions need not force annual grouping of target follow-up. Exponentiating the negative, borrowing-weighted KL discrepancy yields a teacher-induced fractional pseudo-likelihood with a beta-kernel pseudo-count interpretation, thereby explaining how compatible teacher predictions can stabilize local estimation and enabling uncertainty-aware Bayesian survival distillation. Joint inference over the risk function and baseline hazard yields uncertainty in individualized survival probabilities and restricted mean survival time, conditional on the fitted teacher, comparison partition, and borrowing weight. Figure 1 summarizes how teacher information, target-data fitting, and the student's temporal representation enter the posterior.

Our contributions are fourfold. First, we propose probability-scale survival distillation that preserves the original-time target likelihood, allowing flexible nonlinear risk functions without grouping observed follow-up. Second, we give this distillation a teacher-induced fractional pseudo-likelihood interpretation and develop joint posterior inference for individualized survival functionals. Third, under stated regularity conditions, we characterize local information augmentation and posterior–sampling covariance differences under teacher compatibility, first-order displacement under mismatch, and the integrated-hazard limit under comparison-partition refinement. Fourth, we evaluate capacity-dependent borrowing, pooled-information recovery, and kidney-alone-to-kidney–pancreas transfer, including controlled variants that isolate the roles of the distillation criterion and the original-time target likelihood.

2 Related Work

External information transfer in survival analysis. Published survival summaries have been incorporated through estimating equations, constraints, and model-specific integration procedures (Chen et al., 2021; Sheng et al., 2021; Cheng et al., 2023). Prediction-level methods instead operate on individualized external outputs. DiscreteKL and DiSKD transfer conditional event probabilities using

discrete-time target likelihoods (Wang et al., 2025; Deng et al., 2026a), whereas CoxKL and the Cox risk-set formulation of component-wise distillation integrate teacher risk-score information but normalize away the common baseline term in their discrepancies (Wang et al., 2026; Deng et al., 2026b). BayCSD combines probability-scale borrowing with a continuous-time target likelihood that retains original follow-up. This separation addresses a substantive modeling choice. Continuous-time hazard models are widely established in statistical and neural survival analysis (Katzman et al., 2018; Kvamme et al., 2019; Kvamme & Borgan, 2021). For continuously recorded outcomes, coarse grouping removes within-interval event timing and, under common discretization conventions, partial exposure before censoring; finer discrete-time output grids can increase model dimension and estimation variability (Kvamme & Borgan, 2021). A continuous-time model also defines survival between teacher-reported horizons through its hazard specification, rather than a separately chosen interpolation rule. BayCSD preserves these features while allowing the probability-comparison resolution to be chosen separately from baseline-hazard complexity. Its contribution is therefore not a claim that discrete-time models are generally inadequate, but a way to avoid making their temporal representation a prerequisite for probability-scale distillation.

Bayesian borrowing and information sharing. Bayesian borrowing methods differ in the external information they require and how that information enters target inference. Conventional historical-data power priors control the influence of external data by weighting the source survival log-likelihoods, so they require source records to evaluate them (Ibrahim & Chen, 2000; Ibrahim et al., 2015). Prior predictions can also inform borrowing priors. However, survival probabilities reported at selected horizons cannot simply replace event times and censoring indicators in a continuous-time survival likelihood; an explicit construction must translate these predictions into prior information about the survival model. Alternatively, posterior-to-prior transfer reuses distributions over source parameters: Paolucci et al. (2023) update corresponding survival-model parameters, whereas Shwartz-Ziv et al. (2022) transfer an approximate posterior over shared neural-network feature-extractor weights. These approaches avoid retaining source records after training but require access to a source parameter distribution and correspondence for the transferred parameter block, rather than predictions alone. BayCSD instead directly converts teacher survival probabilities evaluated on target subjects at selected horizons into censor-aware fractional event and event-free contributions. Together with the structural prior, these contributions define a data-adaptive predictive pseudo-prior without reconstructing source records or transferring source parameter posteriors. The borrowing factor informs both nonlinear relative-risk functions and cumulative baseline hazards, while the target likelihood retains the original event and censoring times. Its contribution is therefore a survival-specific, prediction-only borrowing construction that separates the teacher-comparison grid from continuous-time target learning and supports joint posterior uncertainty for individualized survival predictions.

3 Bayesian Continuous-Time Survival Distillation

BayCSD separates two modeling roles: the temporal representation governing the target-data likelihood and the representation governing teacher–student comparison. The target model is formulated in continuous time, whereas external guidance is introduced through conditional failure probabilities on an auxiliary comparison partition.

3.1 Continuous-time target student model

Consider an event-sparse target cohort of n independent subjects. Let $T_i > 0$ and $C_i > 0$ denote the event and censoring times, $Y_i = \min(T_i, C_i)$ the observed follow-up, $\Delta_i = \mathbb{I}\{T_i \le C_i\}$ the event indicator, and $\mathbf{X}_i \in \mathbb{R}^p$ the covariates entering the student model. The observed target data are $\mathcal{D}_n = \{(Y_i, \Delta_i, \mathbf{X}_i)\}_{i=1}^n$. We consider the continuous-time relative-risk model

$$\begin{aligned} h_{\boldsymbol{\vartheta}}(t \mid \mathbf{x}) &= h_0(t;\boldsymbol{\gamma}) \exp\{f_{\boldsymbol{\theta}}(\mathbf{x})\}, \\ S_{\boldsymbol{\vartheta}}(t \mid \mathbf{x}) &= \exp\{-H_0(t;\boldsymbol{\gamma}) \exp\{f_{\boldsymbol{\theta}}(\mathbf{x})\}\}, \qquad \boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\gamma}), \end{aligned} \tag{1}$$

where $h_0(t;\boldsymbol{\gamma})$ is the baseline hazard, $H_0(t;\boldsymbol{\gamma}) = \int_0^t h_0(u;\boldsymbol{\gamma})\,\mathrm{d}u$ is the cumulative baseline hazard, and $f_{\boldsymbol{\theta}}(\mathbf{x})$ is a possibly nonlinear log-relative-risk function.

Under conditionally independent right censoring, the target likelihood is

$$L_n(\boldsymbol{\vartheta}; \mathcal{D}_n) = \prod_{i=1}^{n} h_{\boldsymbol{\vartheta}}(Y_i \mid \mathbf{X}_i)^{\Delta_i} S_{\boldsymbol{\vartheta}}(Y_i \mid \mathbf{X}_i). \tag{2}$$

3.2 Censor-aware probability-scale KL distillation

Suppose an external teacher provides individualized survival predictions $S_i^e(t)$ for subjects in the target cohort. Introduce a comparison partition

$$\mathcal{T} = \{0 = \tau_0 < \tau_1 < \cdots < \tau_K\}, \qquad \mathcal{J}_k = (\tau_{k-1}, \tau_k],$$

over the teacher's prediction support. On each comparison interval, the teacher and student induce the conditional failure probabilities

$$p_{ik}^e = 1 - \frac{S_i^e(\tau_k)}{S_i^e(\tau_{k-1})}, \qquad p_{ik}(\boldsymbol{\vartheta}) = 1 - \frac{S_{\boldsymbol{\vartheta}}(\tau_k \mid \mathbf{X}_i)}{S_{\boldsymbol{\vartheta}}(\tau_{k-1} \mid \mathbf{X}_i)}. \tag{3}$$

Conditional on survival through τ_{k-1}, both quantities refer to the Bernoulli outcome of failure in $\mathcal{J}_k$ versus survival beyond τ_k. Teacher predictions are held fixed during target learning and are assumed to refer to the same event definition and time origin as the student.

For right-censored observations, define the resolved-comparison indicator

$$W_{ik} = \mathbb{I}[Y_i \geq \tau_k \text{ or } \{\Delta_i = 1,\ \tau_{k-1} < Y_i < \tau_k\}]\,. \tag{4}$$

The empirical teacher–student discrepancy is

$$Q_n(\boldsymbol{\vartheta}; S^e) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} W_{ik} D_{\mathrm{KL}}\{\mathrm{Bern}(p_{ik}^e) \parallel \mathrm{Bern}[p_{ik}(\boldsymbol{\vartheta})]\}\,, \tag{5}$$

where $\mathrm{Bern}(p)$ denotes a Bernoulli distribution with event probability p. The partition $\mathcal{T}$ determines the temporal resolution of the probability comparison, and W_{ik} restricts the criterion to comparisons resolved by the observed follow-up. Additional support and censoring conventions are given in Appendix A.1. The resolved-comparison rule specifies where teacher information is borrowed; it is not required merely to evaluate the two predicted probabilities. It weights borrowing according to observed follow-up rather than removing the influence of the censoring distribution. Subjects censored within an omitted interval still contribute their complete observed exposure to the target likelihood in equation 2. Under equation 1, let

$$\Delta H_{0k}(\boldsymbol{\gamma}) = H_0(\tau_k; \boldsymbol{\gamma}) - H_0(\tau_{k-1}; \boldsymbol{\gamma})$$

denote the cumulative baseline-hazard increment over $\mathcal{J}_k$. Then

$$p_{ik}(\boldsymbol{\vartheta}) = 1 - \exp[-\Delta H_{0k}(\boldsymbol{\gamma}) \exp\{f_{\boldsymbol{\theta}}(\mathbf{X}_i)\}]\,. \tag{6}$$

Hence the probability-scale criterion depends jointly on the covariate-dependent relative-risk function $f_{\boldsymbol{\theta}}$ and the cumulative baseline-hazard increment ΔH_{0k}. The resulting distillation criterion carries information about both relative-risk structure and the absolute survival scale.

3.3 Bayesian borrowing and posterior inference

Let $\pi_0(d\boldsymbol{\vartheta})$ denote a structural prior specified independently of the teacher, and let $\eta \geq 0$ denote the borrowing weight. Combining the target likelihood with the teacher–student discrepancy gives the BayCSD generalized posterior

$$\Pi_\eta(d\boldsymbol{\vartheta} \mid \mathcal{D}_n, S^e) \propto L_n(\boldsymbol{\vartheta}; \mathcal{D}_n)\, \exp\{-n\eta Q_n(\boldsymbol{\vartheta}; S^e)\}\, \pi_0(d\boldsymbol{\vartheta}). \tag{7}$$

The likelihood, KL factor, and structural prior contribute target-outcome information, teacher guidance, and structural regularization, respectively. The parameter η controls the strength of borrowing; at $\eta = 0$, equation 7 reduces to the target-only posterior.

The teacher factor admits a fractional pseudo-likelihood representation. Expanding the Bernoulli KL in equation 5 yields

$$\exp\{-n\eta Q_n(\boldsymbol{\vartheta}; S^e)\} \propto \prod_{i=1}^{n} \prod_{k=1}^{K} p_{ik}(\boldsymbol{\vartheta})^{\eta W_{ik} p_{ik}^e} \{1 - p_{ik}(\boldsymbol{\vartheta})\}^{\eta W_{ik}(1 - p_{ik}^e)}. \tag{8}$$

For a resolved subject–interval comparison, the two exponents correspond to $\eta W_{ik} p^e_{ik}$ fractional pseudo-events and $\eta W_{ik}(1 - p^e_{ik})$ fractional pseudo-non-events. If p_{ik} were an unconstrained Bernoulli probability, this contribution would be proportional to the kernel of

$$\mathrm{Beta}(1 + \eta W_{ik} p^e_{ik},\, 1 + \eta W_{ik}(1 - p^e_{ik})) \tag{9}$$

as a distribution in p_{ik}. Thus ηW_{ik} has the interpretation of a teacher pseudo-count, partitioned between failure and event-free outcomes according to p^e_{ik}. In BayCSD, the probabilities $p_{ik}(\boldsymbol{\vartheta})$ are coupled through the common continuous-time survival model, so the collection of beta-kernel contributions defines a structured fractional pseudo-likelihood on $\boldsymbol{\vartheta}$.

With a flat structural prior, the posterior mode minimizes the target criterion plus a borrowing-weighted KL criterion, linking BayCSD to KL-based survival-distillation objectives. Appendix A.3 gives the corresponding connections to DiscreteKL, DiSKD, CoxKL, and component-wise distillation.

Posterior inference is joint over the relative-risk function and baseline hazard. The posterior-mean survival function is

$$\overline{S}_\eta(t \mid \mathbf{x}) = \int S_{\boldsymbol{\vartheta}}(t \mid \mathbf{x})\, \Pi_\eta(d\boldsymbol{\vartheta} \mid \mathcal{D}_n, S^e), \tag{10}$$

and posterior quantiles of $S_{\boldsymbol{\vartheta}}(t \mid \mathbf{x})$ define pointwise credible intervals. Restricted mean survival time (RMST) is obtained by integrating posterior survival curves over the specified restriction horizon.

Implementation. The empirical studies use a piecewise-exponential (PWE) baseline hazard together with linear or neural-network log-relative-risk functions. The corresponding likelihood, identification, structural priors, and posterior computation are given in Appendices A.2 and C. We select η from a prespecified grid containing zero using target-likelihood WAIC (Appendix D), and compare this selector with subject-level cross-validation in Appendix G.3.3.

3.4 Statistical properties of borrowing

We summarize three statistical properties of BayCSD borrowing: information augmentation under teacher compatibility, first-order displacement under teacher mismatch, and the continuous-time limit of the probability-scale criterion. Let

$$R_n(\boldsymbol{\vartheta}) = -n^{-1} \log L_n(\boldsymbol{\vartheta}),$$

and let R and Q denote the population counterparts of R_n and Q_n. Suppose $\boldsymbol{\vartheta}_0$ is a locally unique interior minimizer of R, with $H_R = \nabla^2 R(\boldsymbol{\vartheta}_0) \succ 0$. Regularity conditions and proofs are given in Appendix B.

Proposition 1 (Compatible borrowing adds local information)**.** *Suppose the target likelihood is correctly specified and the teacher is compatible with the target population, so that $p^e_{ik} = p_{ik}(\boldsymbol{\vartheta}_0)$ for resolved comparisons. Under the sampling conditions in Appendix B.1, $\boldsymbol{\vartheta}_0$ remains a local minimizer of $R + \eta Q$, and*

$$G := \nabla^2 Q(\boldsymbol{\vartheta}_0) \succeq 0, \qquad H_\eta := H_R + \eta G. \tag{11}$$

For a consistent local minimizer $\widetilde{\boldsymbol{\vartheta}}_{n,\eta}$ of $R_n + \eta Q_n$,

$$\sqrt{n}\big(\widetilde{\boldsymbol{\vartheta}}_{n,\eta} - \boldsymbol{\vartheta}_0\big) \Rightarrow N\big(0, H_\eta^{-1} H_R H_\eta^{-1}\big)\,,$$

whose covariance is no larger than the target-only covariance H_R^{-1} in the positive-semidefinite ordering.

At compatibility, the teacher criterion has zero first-order score at $\boldsymbol{\vartheta}_0$ and contributes positive-semidefinite curvature. Borrowing therefore preserves the local population center while weakly reducing the first-order sampling variability of the fitted center. The generalized posterior has local covariance $n^{-1} H_\eta^{-1}$, which differs from the sampling covariance above; the corresponding results for smooth survival functionals are given in Appendix B.3.

Under teacher mismatch, borrowing instead changes the local population center. If $\nabla Q(\boldsymbol{\vartheta}_0) \neq 0$, the local minimizer $\boldsymbol{\vartheta}^\star_\eta$ of $R + \eta Q$ satisfies

$$\boldsymbol{\vartheta}^\star_\eta - \boldsymbol{\vartheta}_0 = -\eta H_R^{-1} \nabla Q(\boldsymbol{\vartheta}_0) + o(\eta), \qquad \eta \downarrow 0. \tag{12}$$

Thus the local teacher–target discrepancy determines the first-order direction of the borrowing-induced displacement. Appendix B.4 gives the corresponding expansion for smooth survival functionals.

Proposition 2 (Continuous-time limit of the interval KL)**.** *Fix $\mathcal{D}_n$, a student state $\boldsymbol{\vartheta}$, and a horizon $\tau > 0$. Suppose the teacher and student survival curves are absolutely continuous on $[0, \tau]$, with piecewise-continuous hazards bounded above and away from zero. For comparison partitions with vanishing maximum interval width,*

$$Q_{n,m}(\boldsymbol{\vartheta}) \longrightarrow \frac{1}{n}\sum_{i=1}^{n}\int_0^{Y_i\wedge\tau} d_{\mathrm{I}}\{h_i^e(t), h_{\boldsymbol{\vartheta}}(t \mid \mathbf{X}_i)\}\,\mathrm{d}t, \qquad d_{\mathrm{I}}(a,b) = a\log(a/b) - a + b. \quad (13)$$

Proposition 2 establishes the fine-partition limit of the Bernoulli KL criterion. Appendix B.5 gives its proof and the posterior convergence result.

4 Experiments

We evaluate BayCSD in three increasingly realistic settings. Controlled synthetic experiments characterize how student flexibility and teacher compatibility shape borrowing against a known target law. Within-population benchmarks measure recovery of pooled-data predictive information from teacher predictions alone. Kidney-alone (KI) to kidney–pancreas (KP) transfer then evaluates the complete procedure under genuine source–target population shift.

4.1 Common protocol

Within each matched comparison, target-trained methods use the same target observations. *Target-only Bayesian PWE* denotes the continuous-time Bayesian PWE student (Appendix A.2) fitted without borrowing ($\eta = 0$), and *BayCSD* adds the distillation criterion to this specification. Unless otherwise stated, BayCSD selects $\widehat{\eta}$ from a prespecified grid by the widely applicable information criterion (WAIC) under the target-data likelihood (Watanabe, 2010; Gelman et al., 2014). Posterior computation and weight selection are detailed in Appendices C and D.

Evaluation. The integrated Brier score (IBS) measures survival-probability prediction error, while discrimination is measured by Harrell's C-index in the synthetic and pooled-reference experiments and by survival-curve concordance C_{td} in the KI-to-KP application (Graf et al., 1999; Harrell et al., 1982; Antolini et al., 2005). Synthetic experiments additionally report survival RMSE against the known target law and predictive deviance (Burnham & Anderson, 2002; Tutz & Schmid, 2016). Posterior uncertainty is summarized by interval width and truth inclusion in synthetic experiments or pooled-reference inclusion in the recovery benchmarks. Appendix E gives the full conventions.

4.2 Controlled synthetic experiments

Synthetic experiments examine when compatible guidance makes student flexibility useful and how teacher mismatch affects borrowing. We first vary student capacity under a compatible population-oracle teacher, then vary teacher alignment with the student fixed. Appendices F and G detail the designs.

Compatible borrowing supports flexible target learning. Under a common nonlinear target law, we compare a linear risk score, a small neural student, and a richer neural student at $\eta \in \{0, 3, 10\}$. Without borrowing, the richest student has the greatest across-training variability and the worst mean IBS, survival RMSE, and predictive deviance. At $\eta = 10$, ComplexNN performs best on average across all four metrics, while the paired IBS reduction from $\eta = 0$ grows from 0.0044 for Linear to 0.0310 for SimpleNN and 0.0470 for ComplexNN. Appendix Figures 4–5 show that the richest student benefits from lower across-training variance and squared bias. Compatible guidance thus makes student flexibility useful when target failures are sparse.

Teacher compatibility reveals what probability-scale borrowing transfers. We next vary only the teacher. The **compatible** teacher is fitted to a large source cohort from the target law; the **calibration-shifted** teacher rescales the compatible baseline hazard by 1.5, preserving risk ranking while changing the absolute survival scale; and the **risk-function shifted** teacher alters covariate effects while retaining the baseline mechanism. With the compatible teacher, WAIC-selected BayCSD

approaches teacher performance and improves all four metrics over target-only fitting in all 20 replicates, raising the C-index by 0.147 and reducing survival RMSE by 0.160 (Figure 2). Under calibration shift, BayCSD retains 96% of the compatible C-index gain but only 56% and 48% of the IBS and survival-RMSE gains, respectively: probability-scale distillation transfers both absolute survival and relative-risk information. Under risk-function shift, selected borrowing still improves on the target-only fit, but the fixed-η path has an interior optimum because excessive borrowing pulls the student toward shifted covariate effects (Appendix Figure 6).

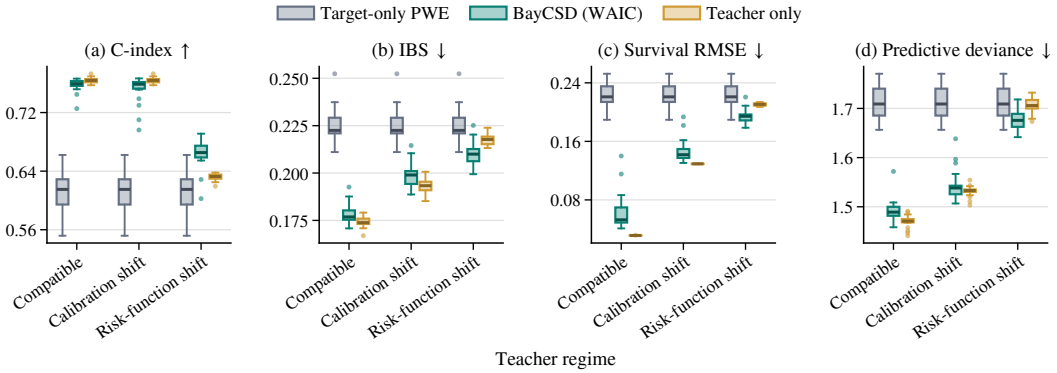


Figure 2: **Borrowing improves prediction across teacher regimes.** Target-only Bayesian PWE, WAIC-selected BayCSD, and direct teacher deployment over 20 paired replicates. Calibration shift changes the absolute survival scale but preserves ranking; risk-function shift alters covariate effects. Boxes show medians and interquartile ranges, with 1.5-IQR whiskers and outliers. The target-only fit is shared across teacher regimes; arrows indicate the preferred direction.

Teacher alignment separates posterior concentration from centering. Along the fixed-η paths, stronger borrowing narrows posterior intervals across teacher regimes, but posterior centering depends on alignment. Compatible guidance keeps bias small and raises truth inclusion; calibration and risk-function shifts displace the center in opposite directions and, under strong borrowing, reduce truth inclusion (Appendix Figures 7 and 8). Median WAIC-selected $\widehat{\eta}$ values are 100, 75, and 5 under compatibility, calibration shift, and risk-function shift, respectively. After selection, the compatible and calibration-shifted teachers yield narrow interval widths while calibration shift produces greater center displacement and lower truth inclusion (Appendix Table 5).

Selection and supporting analyses. WAIC and event-stratified fivefold cross-validation yield similar predictive accuracy, with WAIC closer to the best fitted candidate under risk-function shift (Appendix Table 6). Compatible borrowing remains effective under local structural-prior perturbations (Appendix Table 9). Appendix G.4 empirically illustrates Propositions 1 and 2: the posterior–sampling covariance distinction and stable prediction under comparison-partition refinement.

4.3 Within-population recovery of pooled information

Design. We next assess how much pooled-data predictive information can be recovered when external access is limited to model predictions. SUPPORT (Knaus et al., 1995), METABRIC (Curtis et al., 2012), and a kidney-alone cohort from the Organ Procurement and Transplantation Network (OPTN) are each split into disjoint held-out test, small internal, and large external subsets from the same population. The teacher is fitted to the external subset, target-only Bayesian PWE and BayCSD to the internal subset, and a pooled-data reference to their union. The pooled fit indicates what access to the external raw records would provide. Appendix H gives split sizes and model specifications.

Reference recovery accompanies better outcome prediction. BayCSD improves pooled-reference recovery and held-out prediction (Table 1). Relative to target-only Bayesian PWE, reference survival RMSE decreases by 55–72% and reference RMST MAE by 58–73%. IBS and C-index improve on all three datasets, with the largest C-index gain on OPTN KI, from 0.678 to 0.772. Appendix Table 10 adds teacher and pooled-reference results.

Table 1: **Held-out prediction and recovery of pooled-data information.** IBS and C-index assess held-out prediction, while reference survival RMSE and RMST mean absolute error (MAE) measure recovery of the split-matched pooled-data reference. RMST MAE is reported in days for SUPPORT and OPTN KI and in months for METABRIC. Entries are means ± standard deviations across three splits; bold indicates better performance within each dataset.

Dataset	Method	IBS (10^{-2}) ↓	C-index ↑	Ref. surv. RMSE (10^{-2}) ↓	Ref. RMST MAE ↓
SUPPORT	Target-only Bayesian PWE	24.00 ± 0.52	0.5493 ± 0.0096	13.90 ± 0.71	40.39 ± 2.57
	BayCSD	$\mathbf{22.18 \pm 0.45}$	$\mathbf{0.6060 \pm 0.0097}$	$\mathbf{5.67 \pm 0.91}$	$\mathbf{14.88 \pm 3.31}$
METABRIC	Target-only Bayesian PWE	14.13 ± 1.03	0.6245 ± 0.0163	7.38 ± 2.65	6.08 ± 1.97
	BayCSD	$\mathbf{13.79 \pm 0.57}$	$\mathbf{0.6327 \pm 0.0119}$	$\mathbf{2.03 \pm 0.61}$	$\mathbf{1.67 \pm 0.46}$
OPTN KI	Target-only Bayesian PWE	3.58 ± 0.14	0.6781 ± 0.0285	6.82 ± 0.51	11.18 ± 0.74
	BayCSD	$\mathbf{3.13 \pm 0.16}$	$\mathbf{0.7716 \pm 0.0108}$	$\mathbf{3.05 \pm 0.11}$	$\mathbf{4.69 \pm 0.11}$

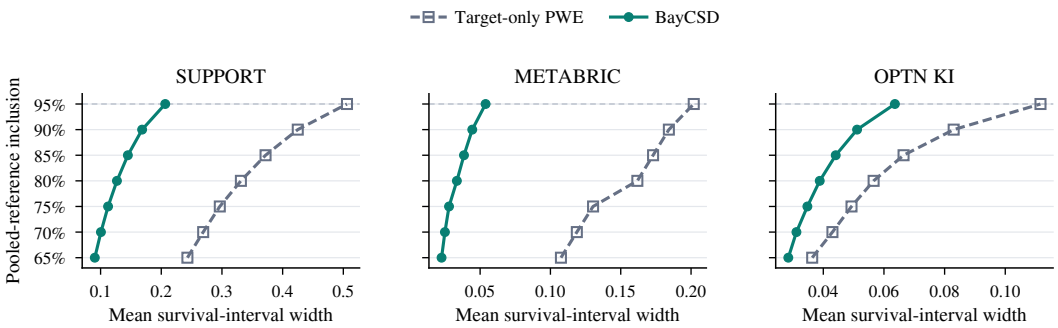


Figure 3: **BayCSD yields narrower survival-probability intervals at matched pooled-reference inclusion.** Columns correspond to datasets. Curves trace posterior interval width against achieved inclusion of the pooled-data reference as the nominal credible level varies; displayed widths are interpolated within each split at common achieved inclusion levels from 65% to 95%. Leftward curves indicate narrower intervals at matched inclusion, and the dashed line marks 95% achieved inclusion. Width is measured on the survival-probability scale.

Borrowing improves the pooled-reference inclusion–width trade-off. Because interval width alone ignores where the posterior is centered, we compare widths at common achieved inclusion of the pooled-data reference, varying the nominal credible level. Throughout the 65–95% inclusion range, BayCSD yields narrower intervals than target-only Bayesian PWE on all three datasets, both for individualized survival probability (Figure 3) and for RMST (Appendix Figure 9). At 95% achieved inclusion, widths decrease by 43–73% (Appendix Table 11).

4.4 Population shift: OPTN kidney-to-kidney–pancreas transfer

Setting and matched design. We evaluate transfer from kidney-alone (KI) to kidney–pancreas (KP) transplantation using de-identified OPTN registry data, with death or kidney graft failure within one year as the endpoint. The KI source contains 24,445 recipients and 1,371 events, and the KP target 1,257 recipients and 51 events; a neural PWE model fitted to KI serves as the teacher. We compare three matched target-only/distillation pairs—LogisticHazard/DiSKD, DeepSurv/component-wise KL, and Bayesian PWE/BayCSD—under five prespecified repeats of event-stratified 5-fold cross-validation with shared outer partitions. Two controlled Bayesian PWE variants separately replace the distillation criterion or the target likelihood. Appendix I gives full design details.

Distillation improves each matched target learner. All three distillation methods improve both IBS and C_{td} over their matched target-only learners in every cross-validation repeat (Table 2). BayCSD attains the lowest IBS and highest C_{td} among the distilled procedures and improves on direct deployment of the KI teacher in both metrics, combining the external predictive signal with KP outcomes rather than reproducing the source predictor.

Table 2: **OPTN KI-to-KP transfer under repeated cross-validation.** IBS (units 10^{-3}) and time-dependent concordance are means $\pm$ standard deviations across five prespecified repeats of 5-fold cross-validation. The three matched target-only/distillation pairs are LogisticHazard/DiSKD, DeepSurv/component-wise KL, and target-only Bayesian PWE/BayCSD. All Bayesian PWE-based methods are evaluated using posterior-mean survival predictions. The two controlled variants retain the Bayesian PWE framework while changing either the distillation criterion or the target likelihood.

Category	Method	IBS (10^{-3}) $\downarrow$	C_{td} $\uparrow$
External model	KI teacher	21.882 ± 0.003	0.7048 ± 0.0023
Target only	LogisticHazard	25.434 ± 0.011	0.5172 ± 0.0395
	DeepSurv	25.401 ± 0.118	0.5443 ± 0.0453
	Bayesian PWE	25.081 ± 0.099	0.6374 ± 0.0197
	Grouped Bayesian PWE	25.963 ± 0.860	0.5653 ± 0.0259
Distillation	DiSKD	22.422 ± 0.152	0.6882 ± 0.0073
	Component-wise KL	23.779 ± 0.627	0.6819 ± 0.0497
	BayCSD	$\mathbf{21.424 \pm 0.047}$	$\mathbf{0.7379 \pm 0.0048}$
Controlled variants	Bayesian PWE + risk-set distillation	23.491 ± 0.052	0.7252 ± 0.0021
	Grouped Bayesian PWE + interval distillation	22.989 ± 0.049	0.7027 ± 0.0020

Controlled variants isolate the two design choices. Two variants each replace one component of BayCSD (Appendix I.3). *Bayesian PWE + risk-set distillation* keeps the continuous-time target likelihood but replaces interval-probability distillation with event-time risk-set KL, which normalizes away the teacher's absolute survival scale. *Grouped Bayesian PWE + interval distillation* keeps interval-probability guidance but replaces the continuous-time likelihood with a monthly grouped likelihood. Both improve over their target-only counterparts, yet neither matches full BayCSD. The risk-set variant retains most of the discrimination ($C_{\mathrm{td}} = 0.725$ versus 0.738) but has higher IBS (23.49 versus 21.42, in units of 10^{-3}), consistent with losing the absolute-scale channel. The grouped variant is worse on both metrics (IBS 22.99, $C_{\mathrm{td}} = 0.703$), a gap already visible between the two target-only fits, reflecting the value of retaining the original event and censoring times.

Predictive gains are accompanied by posterior concentration. Across the 30-, 90-, 180-, and 365-day horizons, BayCSD reduces mean nominal 95% survival-probability interval width by 75.4% relative to target-only Bayesian PWE, and 97.7% of recipients have narrower intervals (Appendix Table 12 and Figure 10). Appendix Table 13 reports predictive-functional MCMC diagnostics.

5 Discussion

BayCSD transfers external survival predictions to support flexible learning and individualized posterior inference. Its conditional interval-probability interface separates teacher comparison from continuous-time target learning. Probability-scale guidance informs both the risk function and cumulative baseline hazard, while joint posterior draws propagate their uncertainty to survival probabilities and RMST.

In controlled experiments, compatible guidance makes the richest student perform best by reducing both across-training variance and squared bias, whereas risk-function shift produces an interior borrowing trade-off. In within-cohort benchmarks, BayCSD approaches an estimated pooled-data reference while improving held-out prediction and yielding narrower survival-probability and RMST intervals at matched reference inclusion. Under KI-to-KP population shift, BayCSD's posterior-mean predictions achieve the lowest mean IBS and highest mean concordance among the three distillation methods; its conditional posterior intervals are narrower than those of target-only Bayesian PWE.

The studied realization assumes proportional covariate effects and conditionally independent right censoring. Generalized-posterior credible intervals need not have nominal frequentist coverage (Miller, 2021). Future work can propagate teacher-training and weight-selection uncertainty and develop local borrowing, non-proportional-hazards students, and competing-risk extensions.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing and organization, and to provide feedback on the presentation of methodology, experiments, results, and scientific figures. All methodological decisions, analyses, numerical results, and scientific conclusions were independently verified by the authors against the derivations, code, experimental outputs, and cited sources. All AI-assisted content was reviewed by the authors, who take responsibility for the final manuscript.

REFERENCES

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

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.

Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, 2002. doi: 10.1007/b97636.

Ziqi Chen, Jing Ning, Yu Shen, and Jing Qin. Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics*, 77(3):1024–1036, 2021.

Yu-Jen Cheng, Yen-Chun Liu, Chang-Yu Tsai, and Chiung-Yu Huang. Semiparametric estimation of the transformation model by leveraging external aggregate data in the presence of population heterogeneity. *Biometrics*, 79(3):1996–2009, 2023.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, 2012.

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

Feiyang Deng, Lingfeng Luo, Jiayu Zhou, and Kevin He. Component-wise composite likelihood distillation for censored time-to-event data. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=WgrRyHYbo5.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, pp. 46–60, 2000.

Joseph G Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer Series in Statistics. Springer, New York, 2001. doi: 10.1007/978-1-4757-3447-8.

Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: Theory and applications. *Statistics in Medicine*, 34(28):3724–3749, 2015.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203, 1995.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

Iwan Paolucci, Yuan-Mao Lin, Jessica Albuquerque Marques Silva, Kristy K Brock, and Bruno C Odisio. Bayesian parametric models for survival prediction in medical applications. *BMC Medical Research Methodology*, 23(1):250, 2023.

Richard D Riley, Kym IE Snell, Joie Ensor, Danielle L Burke, Frank E Harrell Jr, Karel GM Moons, and Gary S Collins. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. *Statistics in Medicine*, 38(7):1276–1296, 2019.

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.

Patrick Royston and Mahesh KB Parmar. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13(1):152, 2013.

Ying Sheng, Yifei Sun, Chiung-Yu Huang, and Mi-Ok Kim. Synthesizing external aggregated information in the penalized Cox regression under population heterogeneity. *Statistics in Medicine*, 40(23):4915–4930, 2021.

Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew G Wilson. Pre-train your loss: Easy Bayesian transfer learning with informative priors. *Advances in Neural Information Processing Systems*, 35:27706–27715, 2022.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

Gerhard Tutz and Matthias Schmid. *Modeling Discrete Time-to-Event Data*. Springer Series in Statistics. Springer, Cham, 2016. doi: 10.1007/978-3-319-28158-2.

Ben Van Calster, David J McLernon, Maarten van Smeden, Laure Wynants, Ewout W Steyerberg, and Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019.

Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

Di Wang, Wen Ye, Randall Sung, Hui Jiang, Jeremy MG Taylor, Lisa Ly, and Kevin He. Kullback-Leibler-based discrete failure time models for integration of published prediction models with new time-to-event dataset. *The Annals of Applied Statistics*, 19(2):1167, 2025.

Di Wang, Wen Ye, Ji Zhu, Gongjun Xu, Weijing Tang, Matthew Zawistowski, Lars G Fritsche, and Kevin He. Incorporating external risk information with the Cox model under population heterogeneity: applications to trans-ancestry polygenic hazard scores. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 2026. doi: 10.1093/jrsssa/qnag042.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116): 3571–3594, 2010. URL `http://jmlr.org/papers/v11/watanabe10a.html`.

Andrew Wey, Nicholas Salkowski, Walter K Kremers, Cory R Schaffhausen, Bertram L Kasiske, Ajay K Israni, and Jon J Snyder. A kidney offer acceptance decision tool to inform the decision to accept an offer or wait for a better kidney. *American Journal of Transplantation*, 18(4):897–906, 2018.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, 2024.

Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International Journal of Epidemiology*, 39(5):1372–1382, 2010.

A Method details

A.1 Comparison partition and censoring mask

The comparison partition $\mathcal{T} = \{0 = \tau_0 < \tau_1 < \cdots < \tau_K\}$ is chosen within the time range over which the teacher provides valid survival predictions. For a queryable teacher, the terminal comparison horizon is fixed from the target training cohort before fitting; for a teacher reported at prespecified horizons, the reported horizons or a prespecified subset define the comparison endpoints. We take $S_i^e(0) = 1$ and assume that the teacher survival predictions are finite, nonincreasing, and lie in $(0, 1]$ on $\mathcal{T}$.

The mask in equation 4 determines which interval comparisons are resolved by the observed follow-up. A subject contributes on $\mathcal{J}_k = (\tau_{k-1}, \tau_k]$ when follow-up reaches τ_k or when an observed failure occurs within $\mathcal{J}_k$. Censoring within the interval leaves that binary comparison unresolved. In particular, $Y_i = \tau_k$ contributes to interval k through the first case in equation 4. The mask is fixed throughout optimization and posterior sampling.

Teacher and student conditional failure probabilities are computed from the survival ratios in equation 3. Probabilities at the boundary values zero or one are evaluated using the corresponding continuous extensions of the Bernoulli KL divergence.

A.2 PWE realization of the continuous-time student

For the empirical implementation, let

$$\mathcal{B} = \{0 = s_0 < s_1 < \cdots < s_{J-1} < s_J = \infty\}, \qquad \mathcal{I}_j = (s_{j-1}, s_j],$$

denote the PWE baseline partition, with $h_0(t) = \lambda_j$ for $t \in \mathcal{I}_j$. Thus the generic baseline parameter $\boldsymbol{\gamma}$ in equation 1 is represented by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)$.

Define the observed exposure

$$E_{ij} = |(0, Y_i] \cap \mathcal{I}_j|$$

and the event-interval index $b(y) = \min\{j : y \le s_j\}$. The continuous-time log likelihood in equation 2 becomes

$$\ell_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \Delta_i \{f_{\boldsymbol{\theta}}(\mathbf{X}_i) + \log \lambda_{b(Y_i)}\} - \sum_{i=1}^{n} \exp\{f_{\boldsymbol{\theta}}(\mathbf{X}_i)\} \sum_{j=1}^{J} \lambda_j E_{ij}. \tag{14}$$

The exposure terms E_{ij} retain each subject's observed follow-up, including partial exposure within the terminal PWE interval.

For a comparison interval $\mathcal{J}_k = (\tau_{k-1}, \tau_k]$, the cumulative baseline-hazard increment appearing in equation 6 is

$$\Delta H_{0k} = \int_{\tau_{k-1}}^{\tau_k} h_0(u)\, \mathrm{d}u = \sum_{j=1}^{J} \lambda_j |\mathcal{I}_j \cap \mathcal{J}_k|. \tag{15}$$

Hence the PWE baseline partition $\mathcal{B}$ and comparison partition $\mathcal{T}$ enter the distillation criterion through their interval overlaps and may be specified independently.

Because a constant shift in the log-relative-risk function can be absorbed into the baseline hazard, we identify their decomposition by reference centering,

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = g_{\boldsymbol{\theta}}(\mathbf{x}) - g_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{ref}}), \tag{16}$$

where $g_{\boldsymbol{\theta}}$ is the unconstrained risk-function output and $\mathbf{x}_{\mathrm{ref}}$ is fixed before fitting. This sets $f_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{ref}}) = 0$, so the baseline hazard is interpreted relative to the reference profile.

A.3 Fractional Bernoulli representation and generalized-Bayesian connections

For $q, p \in (0, 1)$, let

$$d_{\mathrm{B}}(q, p) = D_{\mathrm{KL}}\{\mathrm{Bern}(q) \| \mathrm{Bern}(p)\}.$$

Expanding the Bernoulli KL gives

$$d_{\mathrm{B}}(q, p) = q \log q + (1 - q) \log(1 - q) - q \log p - (1 - q) \log(1 - p). \tag{17}$$

The first two terms are constant with respect to the student parameters. Substituting $q = p^e_{ik}$ and $p = p_{ik}(\boldsymbol{\vartheta})$, multiplying by ηW_{ik}, and summing over subjects and comparison intervals yields equation 8. Boundary probabilities follow by continuity with $0 \log 0 = 0$. This yields the fractional representation in equation 8; the beta-kernel form in equation 9 follows by collecting the powers of p_{ik} and $1 - p_{ik}$.

More generally, let $R_n(\boldsymbol{\vartheta}; \mathcal{D}_n)$ denote an empirical target criterion on the average per-subject scale. The BayCSD update takes the loss-based form

$$\Pi_\eta(d\boldsymbol{\vartheta} \mid \mathcal{D}_n, S^e) \propto \exp\{-nR_n(\boldsymbol{\vartheta}; \mathcal{D}_n) - n\eta Q_n(\boldsymbol{\vartheta}; S^e)\}\, \pi_0(d\boldsymbol{\vartheta}). \tag{18}$$

This is the generalized-Bayesian update associated with the target criterion, teacher–student KL criterion, and structural prior (Bissiri et al., 2016). The likelihood-based BayCSD model studied in this paper uses $R_n = -n^{-1} \log L_n$.

The pair (R_n, Q_n) separates the target learning criterion from the teacher–student comparison criterion. With a flat structural prior, the posterior mode minimizes

$$R_n(\boldsymbol{\vartheta}; \mathcal{D}_n) + \eta Q_n(\boldsymbol{\vartheta}; S^e).$$

A grouped survival likelihood paired with interval-probability KL yields the single-event DiscreteKL construction and the corresponding DiSKD objective (Wang et al., 2025; Deng et al., 2026a). A Cox partial likelihood paired with risk-set KL yields CoxKL and the Cox special case of component-wise distillation (Wang et al., 2026; Deng et al., 2026b). BayCSD combines the continuous-time target likelihood in equation 2 with the censor-aware interval-probability criterion in equation 5. The controlled variants in Section 4.4 replace one of these two criteria at a time while retaining the same Bayesian student model.

B STATISTICAL PROPERTIES AND PROOFS

This appendix gives the conditions and proofs for the local borrowing results in Section 3.4. The analysis separates four objects: the sampling variability of the fitted center under compatible borrowing, the local spread of the generalized posterior, first-order displacement under teacher mismatch, and refinement of the comparison partition. The fractional representation of the teacher factor is given separately in Appendix A.3.

B.1 LOCAL SETUP AND REGULARITY CONDITIONS

Local asymptotic scope. The deployed teacher mapping, finite comparison partition $\mathcal{T}$, student specification, and borrowing weight $\eta \geq 0$ are held fixed as n grows and are taken to be constructed independently of the target sample used for inference. Conditional on these fixed objects, the target records augmented by their teacher predictions are independent and identically distributed under a law P_0. The parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^d$ is represented in locally identifiable coordinates with fixed dimension d.

Write

$$\begin{aligned} R_n(\boldsymbol{\vartheta}) &= \frac{1}{n}\sum_{i=1}^{n} r_i(\boldsymbol{\vartheta}), & Q_n(\boldsymbol{\vartheta}) &= \frac{1}{n}\sum_{i=1}^{n} q_i(\boldsymbol{\vartheta}), \\ q_i(\boldsymbol{\vartheta}) &= \sum_{k=1}^{K} W_{ik} d_{\mathrm{B}}\{p^e_{ik}, p_{ik}(\boldsymbol{\vartheta})\}, & M_{n,\eta}(\boldsymbol{\vartheta}) &= R_n(\boldsymbol{\vartheta}) + \eta Q_n(\boldsymbol{\vartheta}). \end{aligned} \tag{19}$$

Let $R(\boldsymbol{\vartheta}) = \mathbb{E}\{r_i(\boldsymbol{\vartheta})\}$, $Q(\boldsymbol{\vartheta}) = \mathbb{E}\{q_i(\boldsymbol{\vartheta})\}$, and $M_\eta(\boldsymbol{\vartheta}) = R(\boldsymbol{\vartheta}) + \eta Q(\boldsymbol{\vartheta})$. The masks W_{ik} do not depend on $\boldsymbol{\vartheta}$; in derivative expressions, summands with $W_{ik} = 0$ are zero. For the continuous-time likelihood realization, r_i is the negative subject log-likelihood contribution from equation 2.

Sampling conditions. The compatible sampling results use the following local conditions.

1. $\boldsymbol{\vartheta}_0$ is a locally unique interior minimizer of R and $H_R := \nabla^2 R(\boldsymbol{\vartheta}_0) \succ 0$.
2. The losses r_i and q_i are twice continuously differentiable on a convex neighborhood of $\boldsymbol{\vartheta}_0$. Their expectations are finite there, and their derivatives admit integrable envelopes sufficient for differentiation under expectation and uniform laws of large numbers for the Hessians on a smaller neighborhood. On included components, $p_{ik}(\boldsymbol{\vartheta}) \in (0, 1)$ and the derivatives of the Bernoulli KL satisfy the same domination conditions.
3. The target score $\mathbf{s}_{R,i} := \nabla r_i(\boldsymbol{\vartheta}_0)$ has $\mathbb{E}(\mathbf{s}_{R,i}) = 0$ and finite second moment. Write $J_R := \mathrm{Var}_{P_0}(\mathbf{s}_{R,i})$.
4. For each fixed weight under consideration, including $\eta = 0$, there is a consistent interior local minimizer $\widetilde{\boldsymbol{\vartheta}}_{n,\eta}$ of $M_{n,\eta}$, with $\widetilde{\boldsymbol{\vartheta}}_{n,\eta} \xrightarrow{P_0} \boldsymbol{\vartheta}_0$.

For the likelihood specialization, r_i is the negative log-likelihood contribution of a correctly specified regular target model, so the information identity gives $J_R = H_R$. For a general target criterion, J_R replaces H_R.

Posterior conditions. For the generalized-posterior conclusions, assume that the posterior is proper and that the fixed structural prior has a positive, continuously differentiable density near $\boldsymbol{\vartheta}_0$. Write $\mathbf{v} = \sqrt{n}(\boldsymbol{\vartheta} - \widetilde{\boldsymbol{\vartheta}}_{n,\eta})$. Assume root-$n$ posterior tightness: for every deterministic sequence $a_n \to \infty$,

$$\Pi_\eta(\|\mathbf{v}\| > a_n \mid \mathcal{D}_n, S^e) \xrightarrow{P_0} 0. \tag{20}$$

The sampling conditions imply the compact-uniform quadratic expansion used in Appendix B.3; tightness additionally excludes posterior mass on distant branches.

Moment conclusions require more than convergence in distribution. For posterior parameter covariances, assume uniform integrability of $\|\mathbf{v}\|^2$ in P_0-probability: for every $\epsilon > 0$,

$$\lim_{a\to\infty} \limsup_{n\to\infty} P_0\left[\mathbb{E}_{\Pi_\eta}\left\{\|\mathbf{v}\|^2 \mathbb{I}(\|\mathbf{v}\| > a) \mid \mathcal{D}_n, S^e\right\} > \epsilon\right] = 0. \tag{21}$$

For the posterior mean or variance of a scalar functional ψ, impose the same condition with $\|\mathbf{v}\|$ replaced by $|\xi_{n,\psi}|$, where $\xi_{n,\psi} = \sqrt{n}\{\psi(\boldsymbol{\vartheta}) - \psi(\widetilde{\boldsymbol{\vartheta}}_{n,\eta})\}$. These conditions explicitly separate local Gaussian approximation from convergence of the posterior moments used for prediction.

B.2 Compatible borrowing

Under the sampling conditions above and teacher compatibility, $p_{ik}^e = p_{ik}(\boldsymbol{\vartheta}_0)$ almost surely on $\{W_{ik} = 1\}$, the general-criterion sampling limit is

$$\sqrt{n}\big(\widetilde{\boldsymbol{\vartheta}}_{n,\eta} - \boldsymbol{\vartheta}_0\big) \Rightarrow N\big(0, H_\eta^{-1} J_R H_\eta^{-1}\big), \qquad H_\eta = H_R + \eta G. \tag{22}$$

We prove this result and the likelihood specialization in Proposition 1 together.

Population center and distillation score. Compatibility gives $q_i(\boldsymbol{\vartheta}_0) = 0$ almost surely, while $q_i(\boldsymbol{\vartheta}) \geq 0$ for every $\boldsymbol{\vartheta}$. Hence $Q(\boldsymbol{\vartheta}_0) = 0 = \inf_{\boldsymbol{\vartheta}} Q(\boldsymbol{\vartheta})$, and the strict local minimizer $\boldsymbol{\vartheta}_0$ of R remains a strict local minimizer of M_η for every $\eta \geq 0$. For $q, p \in (0,1)$, differentiation of equation 17 gives

$$\partial_p d_{\mathrm{B}}(q,p) = \frac{p-q}{p(1-p)}, \qquad \partial_p^2 d_{\mathrm{B}}(q,p) = \frac{q}{p^2} + \frac{1-q}{(1-p)^2}. \tag{23}$$

At $p = q$, these derivatives are zero and $\{q(1-q)\}^{-1}$, respectively. Writing $\dot{p}_{ik}(\boldsymbol{\vartheta}) = \nabla_{\boldsymbol{\vartheta}} p_{ik}(\boldsymbol{\vartheta})$, the chain rule therefore yields

$$\nabla q_i(\boldsymbol{\vartheta}_0) = \sum_{k=1}^{K} W_{ik} \partial_p d_{\mathrm{B}}\{p_{ik}^e, p_{ik}(\boldsymbol{\vartheta}_0)\} \dot{p}_{ik}(\boldsymbol{\vartheta}_0) = 0 \quad \text{almost surely.} \tag{24}$$

The KL score therefore vanishes subjectwise at compatibility.

Borrowing curvature. For an included component,

$$\begin{aligned} \nabla_{\boldsymbol{\vartheta}}^2 d_{\mathrm{B}}\{p_{ik}^e, p_{ik}(\boldsymbol{\vartheta})\} &= \partial_p^2 d_{\mathrm{B}}\{p_{ik}^e, p_{ik}(\boldsymbol{\vartheta})\} \dot{p}_{ik}(\boldsymbol{\vartheta}) \dot{p}_{ik}(\boldsymbol{\vartheta})^\top \\ &\quad + \partial_p d_{\mathrm{B}}\{p_{ik}^e, p_{ik}(\boldsymbol{\vartheta})\} \nabla_{\boldsymbol{\vartheta}}^2 p_{ik}(\boldsymbol{\vartheta}). \end{aligned} \tag{25}$$

The second term vanishes at compatibility. Differentiating under expectation gives

$$G = \nabla^2 Q(\boldsymbol{\vartheta}_0) = \mathbb{E}\left[\sum_{k=1}^{K} W_{ik} \frac{\dot{p}_{ik}(\boldsymbol{\vartheta}_0) \dot{p}_{ik}(\boldsymbol{\vartheta}_0)^\top}{p_{ik}(\boldsymbol{\vartheta}_0)\{1 - p_{ik}(\boldsymbol{\vartheta}_0)\}}\right] \succeq 0. \tag{26}$$

Each included summand is a nonnegative multiple of an outer product. Consequently, $\nabla^2 M_\eta(\boldsymbol{\vartheta}_0) = H_R + \eta G = H_\eta \succ 0$.

Sampling limit. At the local minimizer, $\nabla M_{n,\eta}(\widetilde{\boldsymbol{\vartheta}}_{n,\eta}) = 0$. An integral-form Taylor expansion gives

$$\begin{aligned} 0 &= \nabla M_{n,\eta}(\boldsymbol{\vartheta}_0) + \overline{H}_{n,\eta}(\widetilde{\boldsymbol{\vartheta}}_{n,\eta} - \boldsymbol{\vartheta}_0), \\ \overline{H}_{n,\eta} &= \int_0^1 \nabla^2 M_{n,\eta}\{\boldsymbol{\vartheta}_0 + t(\widetilde{\boldsymbol{\vartheta}}_{n,\eta} - \boldsymbol{\vartheta}_0)\}\, \mathrm{d}t \xrightarrow{P_0} H_\eta. \end{aligned} \tag{27}$$

Here the convergence follows from consistency and the uniform Hessian law of large numbers. By equation 24, $\nabla M_{n,\eta}(\boldsymbol{\vartheta}_0) = n^{-1} \sum_i \mathbf{s}_{R,i}$, so

$$\sqrt{n}(\widetilde{\boldsymbol{\vartheta}}_{n,\eta} - \boldsymbol{\vartheta}_0) = -\overline{H}_{n,\eta}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}_{R,i}. \tag{28}$$

The multivariate central limit theorem and Slutsky's theorem prove equation 22. The teacher term changes the limiting Jacobian, but contributes no first-order random score.

Likelihood variance ordering. For a correctly specified regular target likelihood, $J_R = H_R$. Set $B_\eta = H_R^{-1/2} H_\eta H_R^{-1/2} = I_d + \eta H_R^{-1/2} G H_R^{-1/2}$. Its eigenvalues are at least one, so $B_\eta^{-2} \preceq I_d$ and

$$\Sigma_{\mathrm{samp},\eta} = H_\eta^{-1} H_R H_\eta^{-1} = H_R^{-1/2} B_\eta^{-2} H_R^{-1/2} \preceq H_R^{-1}. \tag{29}$$

The likelihood information identity $J_R = H_R$ is the step that yields the matrix ordering; the general-criterion limit in equation 22 retains J_R.

B.3 Survival functionals and generalized-posterior implications

Local generalized-posterior limit. Under teacher compatibility and the sampling and additional posterior conditions in Appendix B.1,

$$\left\|\Pi_\eta(\mathbf{v} \in \cdot \mid \mathcal{D}_n, S^e) - N(0, H_\eta^{-1})\right\|_{\mathrm{TV}} \xrightarrow{P_0} 0, \qquad \mathbf{v} = \sqrt{n}(\boldsymbol{\vartheta} - \widetilde{\boldsymbol{\vartheta}}_{n,\eta}), \tag{30}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes total variation distance. To see this, the Hessian conditions and the stationarity of $\widetilde{\boldsymbol{\vartheta}}_{n,\eta}$ give, uniformly over compact sets of $\mathbf{v}$,

$$n\left\{M_{n,\eta}\left(\widetilde{\boldsymbol{\vartheta}}_{n,\eta} + \frac{\mathbf{v}}{\sqrt{n}}\right) - M_{n,\eta}(\widetilde{\boldsymbol{\vartheta}}_{n,\eta})\right\} = \frac{1}{2}\mathbf{v}^\top H_\eta \mathbf{v} + o_{P_0}(1). \tag{31}$$

The local prior-density ratio $\pi_0(\widetilde{\boldsymbol{\vartheta}}_{n,\eta} + \mathbf{v}/\sqrt{n})/\pi_0(\widetilde{\boldsymbol{\vartheta}}_{n,\eta})$ converges to one uniformly on such sets. Together with equation 20, this identifies the normalized Gaussian limit in equation 30; see also Miller (2021) for the general local generalized-posterior argument. Under equation 21, $\mathrm{Var}_{\Pi_\eta}(\boldsymbol{\vartheta} \mid \mathcal{D}_n, S^e) = n^{-1}H_\eta^{-1} + o_{P_0}(n^{-1})$. Thus, under correct likelihood specification, the root-n posterior and sampling covariances satisfy

$$\Sigma_{\mathrm{post},\eta} - \Sigma_{\mathrm{samp},\eta} = H_\eta^{-1} - H_\eta^{-1}H_R H_\eta^{-1} = \eta H_\eta^{-1} G H_\eta^{-1} \succeq 0, \qquad \Sigma_{\mathrm{post},\eta} := H_\eta^{-1}. \tag{32}$$

Individualized survival functionals. Let ψ be continuously differentiable near $\boldsymbol{\vartheta}_0$ and write $\mathbf{g}_0 = \nabla\psi(\boldsymbol{\vartheta}_0)$. Examples are $S_{\boldsymbol{\vartheta}}(t \mid \mathbf{x})$ at fixed $(t, \mathbf{x})$ and $\int_0^\tau S_{\boldsymbol{\vartheta}}(t \mid \mathbf{x})\,\mathrm{d}t$ at fixed $(\tau, \mathbf{x})$, provided differentiation may be passed through the integral. The ordinary delta method applied to equation 22 yields

$$\sqrt{n}\{\psi(\widetilde{\boldsymbol{\vartheta}}_{n,\eta}) - \psi(\boldsymbol{\vartheta}_0)\} \Rightarrow N\left(0, \mathbf{g}_0^\top H_\eta^{-1} J_R H_\eta^{-1}\mathbf{g}_0\right). \tag{33}$$

For $J_R = H_R$, equation 29 gives

$$\mathbf{g}_0^\top H_\eta^{-1} H_R H_\eta^{-1}\mathbf{g}_0 \le \mathbf{g}_0^\top H_R^{-1}\mathbf{g}_0. \tag{34}$$

Applying the same local linearization under equation 30 gives the following weak convergence of random posterior measures in P_0-probability:

$$\Pi_\eta(\xi_{n,\psi} \in \cdot \mid \mathcal{D}_n, S^e) \Rightarrow N\left(0, \mathbf{g}_0^\top H_\eta^{-1}\mathbf{g}_0\right). \tag{35}$$

The additional moment condition for $\xi_{n,\psi}$ then implies

$$\mathbb{E}_{\Pi_\eta}\{\psi(\boldsymbol{\vartheta}) \mid \mathcal{D}_n, S^e\} - \psi(\widetilde{\boldsymbol{\vartheta}}_{n,\eta}) = o_{P_0}(n^{-1/2}), \tag{36}$$

$$n\,\mathrm{Var}_{\Pi_\eta}\{\psi(\boldsymbol{\vartheta}) \mid \mathcal{D}_n, S^e\} \xrightarrow{P_0} \mathbf{g}_0^\top H_\eta^{-1}\mathbf{g}_0. \tag{37}$$

Consequently, the posterior-mean functional shares the sampling limit in equation 33. For survival probability, the corresponding functional is exactly $\overline{S}_\eta(t \mid \mathbf{x})$ in equation 10, obtained by averaging $S_{\boldsymbol{\vartheta}}(t \mid \mathbf{x})$ over the generalized posterior.

Under the compatibility and regularity conditions above, these results establish first-order sampling-variance reduction for posterior-mean survival functionals while separately characterizing their posterior spread.

B.4 First-order displacement under teacher mismatch

Drop teacher compatibility but retain the target population minimizer $\boldsymbol{\vartheta}_0$ and its positive-definite Hessian H_R. Assume that R and Q are twice continuously differentiable near $\boldsymbol{\vartheta}_0$, and set $\mathbf{b}_0 = \nabla Q(\boldsymbol{\vartheta}_0)$. This population calculation uses only local smoothness of R and Q and $H_R \succ 0$. For sufficiently small $\eta \ge 0$, let $\boldsymbol{\vartheta}_\eta^\star$ denote the local minimizer of $R + \eta Q$ on the branch satisfying $\boldsymbol{\vartheta}_0^\star = \boldsymbol{\vartheta}_0$. Then, as $\eta \downarrow 0$,

$$\boldsymbol{\vartheta}_\eta^\star - \boldsymbol{\vartheta}_0 = -\eta H_R^{-1}\mathbf{b}_0 + o(\eta), \tag{38}$$

$$\psi(\boldsymbol{\vartheta}_\eta^\star) - \psi(\boldsymbol{\vartheta}_0) = -\eta\nabla\psi(\boldsymbol{\vartheta}_0)^\top H_R^{-1}\mathbf{b}_0 + o(\eta) \tag{39}$$

for any differentiable scalar functional ψ.

Indeed, apply the implicit function theorem to $\nabla R(\boldsymbol{\vartheta}_\eta^\star) + \eta \nabla Q(\boldsymbol{\vartheta}_\eta^\star) = 0$ at $(\boldsymbol{\vartheta}_0, 0)$. Since the parameter derivative is $H_R \succ 0$, the theorem gives a differentiable local stationary branch with derivative $\partial \boldsymbol{\vartheta}_\eta^\star / \partial \eta\big|_{\eta=0} = -H_R^{-1}\mathbf{b}_0$. Continuity of the Hessian makes this branch locally minimizing for sufficiently small η. Taylor expansion proves equation 38, and the chain rule proves equation 39.

Under compatibility, $\mathbf{b}_0 = 0$ and Appendix B.2 gives exact preservation of the local population center. Equation 38–equation 39 isolate the first-order centering effect of mismatch from the sampling covariance and posterior spread characterized above.

B.5 Comparison-partition refinement and posterior consequence

We prove Proposition 2 with the target cohort, teacher curves, and student model held fixed. This is a fixed-sample refinement limit, separate from the local sampling analysis in Appendices B.1–B.4.

Setup. For a fixed $\tau > 0$, write

$$\mathcal{T}_m = \{0 = \tau_{m0} < \tau_{m1} < \cdots < \tau_{mK_m} = \tau\},$$
$$d_{mk} = \tau_{mk} - \tau_{m,k-1}, \qquad \delta_m = \max_{1 \le k \le K_m} d_{mk} \longrightarrow 0.$$

The result applies to arbitrary partition sequences with vanishing mesh. Their endpoints query the same teacher and student curves for all m. For each fixed state $\boldsymbol{\vartheta}$ under consideration, these curves have the representations

$$S_i^e(t) = \exp\Big\{-\int_0^t h_i^e(u)\,\mathrm{d}u\Big\}, \qquad S_{\boldsymbol{\vartheta}}(t \mid \mathbf{X}_i) = \exp\Big\{-\int_0^t h_{\boldsymbol{\vartheta}}(u \mid \mathbf{X}_i)\,\mathrm{d}u\Big\}. \tag{40}$$

Assume the hazards are piecewise continuous on $[0, \tau]$ and satisfy $0 < \underline{h} \le h_i^e(t), h_{\boldsymbol{\vartheta}}(t \mid \mathbf{X}_i) \le \overline{h} < \infty$. The bounds may depend on the fixed cohort and state, but not on m. These conditions cover positive finite-piece PWE hazards on a compact horizon, with $\mathcal{B}$ fixed while the comparison partition $\mathcal{T}_m$ is refined.

Let W_{imk} be defined as in equation 4 with endpoints $(\tau_{m,k-1}, \tau_{mk}]$, and let p_{imk}^e and $p_{imk}(\boldsymbol{\vartheta})$ denote the corresponding teacher and student conditional failure probabilities. Using d_{B} from Appendix A.3, define

$$q_{i,m}(\boldsymbol{\vartheta}) = \sum_{k=1}^{K_m} W_{imk} d_{\mathrm{B}}\{p_{imk}^e, p_{imk}(\boldsymbol{\vartheta})\}, \qquad Q_{n,m}(\boldsymbol{\vartheta}) = \frac{1}{n}\sum_{i=1}^{n} q_{i,m}(\boldsymbol{\vartheta}). \tag{41}$$

Thus $q_{i,m}$ aggregates the resolved interval-wise discrepancies for subject i, and $Q_{n,m}$ averages these subject-level contributions over the target cohort, matching the scale of equation 5.

Small-interval expansion. For $a, b \in [\underline{h}, \overline{h}]$ and $\delta \downarrow 0$, the full Bernoulli KL satisfies

$$\begin{aligned} d_{\mathrm{B}}(1 - e^{-\delta a}, 1 - e^{-\delta b}) &= (1 - e^{-\delta a}) \log \frac{1 - e^{-\delta a}}{1 - e^{-\delta b}} + \delta e^{-\delta a}(b - a) \\ &= \delta\{a \log(a/b) - a + b\} + O(\delta^2), \end{aligned} \tag{42}$$

uniformly over a, b in this compact interval. Indeed, $1 - e^{-\delta a} = \delta a + O(\delta^2)$ and $\log\{(1 - e^{-\delta a})/(1 - e^{-\delta b})\} = \log(a/b) + O(\delta)$ uniformly. For each comparison interval, let

$$\overline{h}_{imk}^e = \frac{1}{d_{mk}} \int_{\tau_{m,k-1}}^{\tau_{mk}} h_i^e(t)\,\mathrm{d}t, \qquad \overline{h}_{imk}(\boldsymbol{\vartheta}) = \frac{1}{d_{mk}} \int_{\tau_{m,k-1}}^{\tau_{mk}} h_{\boldsymbol{\vartheta}}(t \mid \mathbf{X}_i)\,\mathrm{d}t.$$

Equation 40 gives the exact identities $p_{imk}^e = 1 - e^{-d_{mk}\overline{h}_{imk}^e}$ and $p_{imk}(\boldsymbol{\vartheta}) = 1 - e^{-d_{mk}\overline{h}_{imk}(\boldsymbol{\vartheta})}$. Applying equation 42 and summing yields

$$\begin{aligned} &q_{i,m}(\boldsymbol{\vartheta}) = \sum_{k=1}^{K_m} W_{imk} d_{mk} d_{\mathrm{I}}\{\overline{h}_{imk}^e, \overline{h}_{imk}(\boldsymbol{\vartheta})\} + O\left(\sum_{k=1}^{K_m} d_{mk}^2\right), \\ &\sum_{k=1}^{K_m} d_{mk}^2 \le \tau \delta_m. \end{aligned} \tag{43}$$

Integrated discrepancy and the observation boundary. Fix i and $\boldsymbol{\vartheta}$. On $(\tau_{m,k-1}, \tau_{mk}]$, set $F_{im}(t) = d_{\mathrm{I}}\{\overline{h}^{e}_{imk}, \overline{h}_{imk}(\boldsymbol{\vartheta})\}$, and write $F_i(t) = d_{\mathrm{I}}\{h^e_i(t), h_{\boldsymbol{\vartheta}}(t \mid \mathbf{X}_i)\}$. At every common continuity point of the hazards, interval averages converge to their values as $\delta_m \to 0$, so $F_{im}(t) \to F_i(t)$ almost everywhere. Since d_{I} is continuous and bounded on $[\underline{h}, \overline{h}]^2$, dominated convergence gives $\int_0^\tau |F_{im}(t) - F_i(t)|\,\mathrm{d}t \to 0$.

Let U_{im} be the union of intervals with $W_{imk} = 1$. If $Y_i < \tau$ is strictly inside an interval, censoring omits that interval whereas an observed failure includes it. Relative to $(0, Y_i \wedge \tau]$, the omitted or added segment therefore has length at most δ_m. If Y_i is a grid endpoint or $Y_i \geq \tau$, the intervals agree up to endpoints. Consequently, $|U_{im} \triangle (0, Y_i \wedge \tau]| \leq \delta_m$, where $\triangle$ denotes symmetric difference and $|\cdot|$ denotes length. Together with equation 43, this gives

$$\left| q_{i,m}(\boldsymbol{\vartheta}) - \int_0^{Y_i \wedge \tau} F_i(t)\,\mathrm{d}t \right| \leq \int_0^\tau |F_{im}(t) - F_i(t)|\,\mathrm{d}t + O(\delta_m) \longrightarrow 0. \tag{44}$$

Averaging over the fixed n subjects proves equation 13. The limit is the integrated hazard discrepancy accumulated over observed target follow-up for any partition sequence with vanishing mesh.

Fixed-sample posterior consequence. Fix n, $\eta \geq 0$, the target criterion R_n, the structural prior π_0, and the student model. Let $\Pi_0(\cdot \mid \mathcal{D}_n)$ be the proper target-only posterior obtained from equation 18 at $\eta = 0$. Assume the comparison criteria are measurable in $\boldsymbol{\vartheta}$ and the curve conditions of Proposition 2 hold for Π_0-almost every state. Then $Q_{n,m}(\boldsymbol{\vartheta}) \to Q_{n,\infty}(\boldsymbol{\vartheta}) < \infty$ for Π_0-almost every $\boldsymbol{\vartheta}$. For $r \in \mathbb{N} \cup \{\infty\}$, define

$$w_{\eta,r}(\boldsymbol{\vartheta}) = e^{-n\eta Q_{n,r}(\boldsymbol{\vartheta})}, \qquad Z_{\eta,r} = \int w_{\eta,r}(\boldsymbol{\vartheta})\,\Pi_0(d\boldsymbol{\vartheta} \mid \mathcal{D}_n). \tag{45}$$

The full KL is nonnegative, so $0 < w_{\eta,r} \leq 1$ almost everywhere. Dominated convergence therefore gives $\int |w_{\eta,m} - w_{\eta,\infty}|\,d\Pi_0 \to 0$ and $Z_{\eta,m} \to Z_{\eta,\infty} > 0$. The finite-grid posteriors and their limit are

$$\Pi_{\eta,r}(d\boldsymbol{\vartheta} \mid \mathcal{D}_n, S^e) = \frac{w_{\eta,r}(\boldsymbol{\vartheta})}{Z_{\eta,r}}\,\Pi_0(d\boldsymbol{\vartheta} \mid \mathcal{D}_n), \qquad r \in \mathbb{N} \cup \{\infty\}. \tag{46}$$

Suppressing the fixed conditioning arguments, convergence of the weights and normalizing constants implies

$$\|\Pi_{\eta,m} - \Pi_{\eta,\infty}\|_{\mathrm{TV}} = \frac{1}{2}\int \left| \frac{w_{\eta,m}}{Z_{\eta,m}} - \frac{w_{\eta,\infty}}{Z_{\eta,\infty}} \right| d\Pi_0 \longrightarrow 0. \tag{47}$$

Posterior expectations of every bounded measurable functional converge; this includes $S_{\boldsymbol{\vartheta}}(t \mid \mathbf{x})$ at fixed $(t, \mathbf{x})$ and $\int_0^\tau S_{\boldsymbol{\vartheta}}(t \mid \mathbf{x})\,\mathrm{d}t$ at fixed $(\tau, \mathbf{x})$. Thus the generalized posterior is stable under comparison-partition refinement at fixed n and η.

C PWE PRIORS AND POSTERIOR COMPUTATION

C.1 STRUCTURAL PRIORS

Covariates are standardized using the target training cohort, and $\mathbf{x}_{\mathrm{ref}}$ in equation 16 is the standardized image of the zero profile on the original covariate scale. Let d_θ be the number of sampled score parameters and $\sigma_r > 0$ the specified prior standard deviation for parameter r. We use independent Gaussian priors

$$\theta_r \sim N(0, \sigma_r^2), \qquad r = 1, \ldots, d_\theta. \tag{48}$$

These structural priors are specified independently of the teacher predictions.

For a finite PWE interval of width $d_j = s_j - s_{j-1}$, let $m_0 > 0$ be the prior mean baseline rate and $\kappa > 0$ the concentration parameter. Under the shape–rate parameterization,

$$A_j = \lambda_j d_j \sim \mathrm{Gamma}(\kappa m_0 d_j, \kappa), \qquad \lambda_j \sim \mathrm{Gamma}(\kappa m_0 d_j, \kappa d_j). \tag{49}$$

Every finite-interval rate therefore has prior mean m_0, while the cumulative-hazard increment scales with interval width, as in Gamma-process constructions for Bayesian survival models (Ibrahim et al., 2001). The final open interval uses the proper prior $\lambda_J \sim \mathrm{Gamma}(1, 10)$ in the synthetic experiments. Dataset-specific prior settings are stated with the corresponding experiments.

C.2 Posterior parameterization and computation

Sampling uses unconstrained log rates $u_j = \log \lambda_j$, so $\boldsymbol{\gamma} = \mathbf{u}$ in the PWE realization. If a_j and b_j denote the shape and rate of the corresponding Gamma prior, the negative log generalized posterior, up to an additive constant, is

$$U_\eta(\boldsymbol{\theta}, \mathbf{u}) = -\ell_n(\boldsymbol{\theta}, e^{\mathbf{u}}) + n\eta Q_n((\boldsymbol{\theta}, \mathbf{u}); S^e) + \frac{1}{2}\sum_{r=1}^{d_\theta} \frac{\theta_r^2}{\sigma_r^2} + \sum_{j=1}^{J} \{b_j e^{u_j} - a_j u_j\}. \tag{50}$$

The term $-a_j u_j$ includes the Jacobian of the transformation from λ_j to u_j, and the teacher contribution is $n\eta Q_n$ under the $1/n$ scaling in equation 5.

For each fixed η, a joint maximum a posteriori (MAP) fit initializes independently seeded full-gradient, diagonally preconditioned Metropolis-adjusted Langevin algorithm (MALA) chains with a Metropolis–Hastings correction (Roberts & Rosenthal, 1998). Proposal-scale adaptation is restricted to burn-in and frozen before retained draws. Experiment-specific chain lengths, burn-in, thinning, and prediction subsampling are given in Appendices G–I. Survival probabilities and RMST are evaluated jointly from the risk function and baseline hazard at each posterior state used for prediction.

D Borrowing-weight selection

The borrowing weight $\eta \geq 0$ controls the contribution of the teacher–student KL criterion to the generalized posterior. We select η by predictive fit to the target outcomes. For a candidate weight, let

$$\ell_i(\boldsymbol{\vartheta}) = \Delta_i \log h_{\boldsymbol{\vartheta}}(Y_i \mid \mathbf{X}_i) + \log S_{\boldsymbol{\vartheta}}(Y_i \mid \mathbf{X}_i) \tag{51}$$

denote the target log-likelihood contribution of subject i. Given posterior draws $\boldsymbol{\vartheta}_\eta^{(1)}, \ldots, \boldsymbol{\vartheta}_\eta^{(M)}$, define

$$\widehat{\mathrm{lppd}}_\eta = \sum_{i=1}^{n} \log \left\{ \frac{1}{M} \sum_{m=1}^{M} \exp\big(\ell_i(\boldsymbol{\vartheta}_\eta^{(m)})\big) \right\}, \tag{52}$$

$$\widehat{p}_{\mathrm{WAIC},\eta} = \sum_{i=1}^{n} \mathrm{Var}_m\Big\{ \ell_i(\boldsymbol{\vartheta}_\eta^{(m)}) \Big\}, \qquad \widehat{\mathrm{WAIC}}(\eta) = -2\left\{ \widehat{\mathrm{lppd}}_\eta - \widehat{p}_{\mathrm{WAIC},\eta} \right\}. \tag{53}$$

This is the pointwise target-likelihood WAIC (Watanabe, 2010; Gelman et al., 2014). The candidate generalized posteriors incorporate both target outcomes and teacher guidance, while their borrowing weights are compared through predictive fit under the target likelihood.

The selected weight is

$$\widehat{\eta} = \arg\min_{\eta \in \mathcal{E}} \widehat{\mathrm{WAIC}}(\eta), \tag{54}$$

with numerical ties resolved in favor of the smaller weight. Unless otherwise specified in the corresponding experiment, we use the prespecified candidate grid

$$\mathcal{E} = \{0, 0.5, 1, 2, 3, 5, 10, 20, 50, 100, 200\}. \tag{55}$$

The candidate $\eta = 0$ recovers the matched target-only Bayesian PWE posterior. Experiment-specific fixed-weight analyses or alternative candidate grids are stated locally with their experimental designs.

E Prediction and evaluation estimators

This appendix defines the estimators used in Section 4. The synthetic experiments report IBS, survival RMSE, predictive deviance, Harrell's C-index, and interval summaries relative to the known target law. The pooled-reference benchmarks report IBS, Harrell's C-index, and recovery and interval summaries relative to the pooled-data reference. The OPTN application reports IBS, C_{td}, and interval widths. Throughout, posterior uncertainty describes variation in survival functionals within a fitted model, whereas sampling variability describes changes in fitted predictions across training cohorts. Interval inclusion records whether a specified functional lies inside an interval. Probability calibration instead concerns agreement between predicted probabilities and event frequencies; neither interval inclusion nor an overall prediction-error metric such as IBS isolates calibration.

E.1 Posterior prediction rule

For a fixed fitted η, let $\boldsymbol{\vartheta}_\eta^{(m)}$, $m = 1, \ldots, M$, be the M joint posterior states used for prediction after the prespecified thinning. Each state contains score parameters $\boldsymbol{\theta}_\eta^{(m)}$ and baseline parameters $\boldsymbol{\gamma}_\eta^{(m)}$. The Monte Carlo predictors are

$$\widehat{S}_\eta(t \mid \mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} S_{\boldsymbol{\vartheta}_\eta^{(m)}}(t \mid \mathbf{x}), \qquad \widehat{f}_\eta(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} f_{\boldsymbol{\theta}_\eta^{(m)}}(\mathbf{x}). \tag{56}$$

The first approximates $\overline{S}_\eta$ in equation 10; the second is the posterior-mean log-risk score used for Harrell's C-index in the synthetic experiments and pooled-reference benchmarks. IBS, RMSE, and concordance are computed from these predictors. Neither per-draw metric averaging nor prediction at posterior-mean parameters is used for primary performance. Predictive deviance uses the distinct likelihood-mixture rule below. For deterministic comparators, $\widehat{S}$ and $\widehat{f}$ below denote their fitted predictions, and the same evaluation estimators apply.

E.2 Integrated Brier score

Let $\widehat{G}_{\mathrm{ref}}$ denote the Kaplan–Meier estimate of the censoring survival function from the designated censoring-reference sample. Because $\widehat{G}_{\mathrm{ref}}$ is estimated marginally, this IPCW evaluation additionally assumes that the censoring distribution is independent of the predictors. For an evaluation subject with observed time Y_i and event indicator Δ_i, the inverse-probability-of-censoring-weighted (IPCW) Brier score at time t is

$$\widehat{\mathrm{BS}}(t) = \frac{1}{N_{\mathrm{ev}}} \sum_{i=1}^{N_{\mathrm{ev}}} \left[\frac{\mathbb{I}(Y_i \le t, \Delta_i = 1) \widehat{S}_\eta(t \mid \mathbf{x}_i)^2}{\widehat{G}_{\mathrm{ref}}(Y_i)} + \frac{\mathbb{I}(Y_i > t)\{1 - \widehat{S}_\eta(t \mid \mathbf{x}_i)\}^2}{\widehat{G}_{\mathrm{ref}}(t)} \right]. \tag{57}$$

Subjects censored by t contribute neither term. Over a prespecified interval $[a, b]$, the integrated Brier score is

$$\widehat{\mathrm{IBS}} = \frac{1}{b-a} \int_a^b \widehat{\mathrm{BS}}(t)\, \mathrm{d}t, \tag{58}$$

which is evaluated by trapezoidal integration on the declared prediction grid; lower values indicate more accurate survival-probability prediction (Graf et al., 1999; Gerds & Schumacher, 2006; Kvamme et al., 2019). In the synthetic experiments, the independent evaluation cohort supplies $\widehat{G}_{\mathrm{ref}}$, avoiding truncation of the fixed evaluation horizon by the small target training cohort; it does not enter fitting or WAIC/CV tuning. In the pooled-reference benchmarks, a split-specific censoring-reference rule is fixed before evaluation and shared across all paired methods within that split. In the OPTN application, $\widehat{G}_{\mathrm{ref}}$ is estimated from the corresponding KP training fold. The same weights and time grid are used for all methods in a paired comparison.

E.3 Survival-function root mean squared error

Write $S^\star(t \mid \mathbf{x})$ for the true target survival function, $\mathbf{x}_i^{\mathrm{ev}}$, $i = 1, \ldots, N_{\mathrm{ev}}$, for the N_{ev} evaluation profiles, and a_l, $l = 1, \ldots, L_{\mathrm{ev}}$, for the L_{ev} evaluation times. Synthetic-data survival RMSE is

$$\widehat{\mathrm{RMSE}}_S = \left[\frac{1}{N_{\mathrm{ev}} L_{\mathrm{ev}}} \sum_{i=1}^{N_{\mathrm{ev}}} \sum_{l=1}^{L_{\mathrm{ev}}} \{\widehat{S}_\eta(a_l \mid \mathbf{x}_i^{\mathrm{ev}}) - S^\star(a_l \mid \mathbf{x}_i^{\mathrm{ev}})\}^2 \right]^{1/2}. \tag{59}$$

The evaluation profiles and time points receive equal weight. This quantity is reported only for synthetic data, where $S^\star$ is known.

E.4 Posterior-predictive deviance

If $\ell_i^{\mathrm{ev}}(\boldsymbol{\vartheta})$ denotes the exact event/censoring log-likelihood contribution of evaluation subject i, the reported predictive deviance per subject is

$$\widehat{D}_{\mathrm{pred}} = -\frac{2}{N_{\mathrm{ev}}} \sum_{i=1}^{N_{\mathrm{ev}}} \log \left[\frac{1}{M} \sum_{m=1}^{M} \exp\{\ell_i^{\mathrm{ev}}(\boldsymbol{\vartheta}_\eta^{(m)})\} \right]. \tag{60}$$

Thus the likelihood is averaged over posterior states before taking its logarithm; the teacher discrepancy is not part of the evaluation likelihood (Gelman et al., 2014). Lower values indicate better held-out predictive fit. We compare this quantity only among models with the common continuous-time likelihood and do not use it to compare different likelihood families in the OPTN application.

E.5 Harrell's concordance index

For the synthetic experiments and pooled-reference benchmarks, define the set of comparable ordered pairs

$$\mathcal{P}_H = \{(i,j) : Y_i < Y_j,\ \Delta_i = 1\}. \tag{61}$$

Because larger $\widehat{f}_\eta$ denotes greater event risk, Harrell's concordance estimate is

$$\widehat{C}_H = \frac{1}{|\mathcal{P}_H|} \sum_{(i,j)\in\mathcal{P}_H} \left[\mathbb{I}\{\widehat{f}_\eta(\mathbf{x}_i) > \widehat{f}_\eta(\mathbf{x}_j)\} + \frac{1}{2}\mathbb{I}\{\widehat{f}_\eta(\mathbf{x}_i) = \widehat{f}_\eta(\mathbf{x}_j)\} \right]. \tag{62}$$

Pairs with equal observed times are excluded, and tied risk scores receive half credit (Harrell et al., 1982). The metric is computed from the posterior-mean score in equation 56; higher values indicate better risk ranking.

E.6 Time-dependent concordance

The OPTN application uses the survival-curve concordance of Antolini et al. (2005), because its discrete-time comparators do not produce a single time-invariant risk score; all predictors in that application are therefore compared through their survival curves. With the same comparable-pair set, both subjects' posterior-mean survival curves are evaluated at the earlier observed event time:

$$\widehat{C}_{\mathrm{td}} = \frac{1}{|\mathcal{P}_H|} \sum_{(i,j)\in\mathcal{P}_H} \mathbb{I}\{\widehat{S}_\eta(Y_i \mid \mathbf{x}_i) < \widehat{S}_\eta(Y_i \mid \mathbf{x}_j)\}. \tag{63}$$

The implementation evaluates each curve at the original event time, excludes equal observed times, and uses a strict comparison: tied survival predictions receive zero credit. No IPCW weights enter this concordance calculation. Higher values indicate better dynamic ranking.

E.7 Posterior interval summaries

For method m, credible level c, and $\alpha_c = (1-c)/2$, let

$$L^{(c)}_{i,m}(t) = q_{\alpha_c}\{S_{\boldsymbol{\vartheta}^{(b)}_m}(t \mid \mathbf{x}_i)\}_{b=1}^M, \qquad U^{(c)}_{i,m}(t) = q_{1-\alpha_c}\{S_{\boldsymbol{\vartheta}^{(b)}_m}(t \mid \mathbf{x}_i)\}_{b=1}^M \tag{64}$$

be the pointwise posterior credible interval for the survival probability (Ibrahim et al., 2001). Its mean width at time t is

$$\widehat{W}_m(c,t) = \frac{1}{N_{\mathrm{ev}}} \sum_{i=1}^{N_{\mathrm{ev}}} \{U^{(c)}_{i,m}(t) - L^{(c)}_{i,m}(t)\}. \tag{65}$$

When the target law is known, as in the synthetic experiments, the nominal 95% pointwise truth inclusion and signed survival bias are

$$\begin{aligned} \widehat{\mathrm{Inc}}^{\star}_m(t) &= \frac{1}{N_{\mathrm{ev}}} \sum_{i=1}^{N_{\mathrm{ev}}} \mathbb{I}\{L^{(0.95)}_{i,m}(t) \le S^\star(t \mid \mathbf{x}_i) \le U^{(0.95)}_{i,m}(t)\}, \\ \widehat{\mathrm{Bias}}_m(t) &= \frac{1}{N_{\mathrm{ev}}} \sum_{i=1}^{N_{\mathrm{ev}}} \{\widehat{S}_m(t \mid \mathbf{x}_i) - S^\star(t \mid \mathbf{x}_i)\}. \end{aligned} \tag{66}$$

The reported synthetic-experiment summaries average these quantities over prespecified horizons when stated. Across-replicate Monte Carlo intervals summarize variation over independently generated cohorts and are distinct from the posterior interval for one fitted cohort.

For a pooled-reference benchmark, let $\widehat{S}_{\mathrm{ref}}(t \mid \mathbf{x}_i)$ be the posterior-mean prediction from the pooled-data reference fitted without test subjects. We report

$$\widehat{\mathrm{Inc}}_m^{\mathrm{ref}}(c,t) = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \mathbb{I}\Big\{ L_{i,m}^{(c)}(t) \le \widehat{S}_{\mathrm{ref}}(t \mid \mathbf{x}_i) \le U_{i,m}^{(c)}(t) \Big\} \tag{67}$$

and the corresponding mean width from equation 65. Overall survival-functional summaries average subject–horizon indicators and widths over a prespecified common grid. We evaluate $c \in \{0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.975, 0.99\}$. The width $W_m^{95,\mathrm{ref}}$ is obtained by linear interpolation of the width against the nondecreasing pooled-reference-inclusion curve at inclusion 0.95. If the observed curve does not bracket 0.95, this summary is reported as not estimable rather than extrapolated. The primary interval-efficiency contrast is $W_{\mathrm{BayCSD}}^{95,\mathrm{ref}} / W_{\mathrm{Internal}}^{95,\mathrm{ref}}$. The superscript $95, \mathrm{ref}$ denotes achieved inclusion, not nominal credible level: the methods may require different c to attain it. This is a retrospective comparison against the fitted reference held at its posterior mean.

E.8 Individualized survival and RMST reference recovery

On evaluation times $t_1, \ldots, t_L$, individualized recovery of the pooled-data reference is measured by

$$\mathrm{RMSE}_{m,\mathrm{ref}} = \left[\frac{1}{N_{\mathrm{test}} L} \sum_{i=1}^{N_{\mathrm{test}}} \sum_{\ell=1}^{L} \{ \widehat{S}_m(t_\ell \mid \mathbf{x}_i) - \widehat{S}_{\mathrm{ref}}(t_\ell \mid \mathbf{x}_i) \}^2 \right]^{1/2}. \tag{68}$$

For a dataset-specific horizon τ, posterior draw b induces the individualized restricted mean survival time

$$\mathrm{RMST}_{m,i}^{(b)}(\tau) = \int_0^\tau S_{\boldsymbol{\vartheta}_m^{(b)}}(t \mid \mathbf{x}_i)\, \mathrm{d}t, \qquad \widehat{\mathrm{RMST}}_{\mathrm{ref},i}(\tau) = \int_0^\tau \widehat{S}_{\mathrm{ref}}(t \mid \mathbf{x}_i)\, \mathrm{d}t. \tag{69}$$

Posterior quantiles across the draw-wise RMST values preserve dependence between the risk score and baseline hazard (Royston & Parmar, 2013). We compare posterior-mean RMST with the pooled-reference RMST using subject-level MAE or RMSE and apply the same pooled-reference-inclusion–width construction to RMST intervals. These quantities measure recovery and interval efficiency relative to the estimated pooled-data reference.

The empirical prediction-function variance in the student-capacity experiment uses one common reference cohort across training replicates and is defined separately in equation 73.

F Synthetic student-capacity experiment: design and supplementary analyses

This appendix gives the full design and results for the compatible-borrowing student-flexibility experiment summarized in Section 4.2. It first reports the complete prediction paths across student specifications and then decomposes the resulting prediction error into across-training variance and squared bias.

Target distribution and compatible teacher. We generated $X_1, \ldots, X_{12} \overset{\mathrm{iid}}{\sim} N(0,1)$, with X_{10}, X_{11}, X_{12} serving as noise variables. Let $a_0 = 0.75$, with positive normalizing standard deviations σ_s and σ_c defined below, and form

$$\begin{aligned} H_1 &= (0.8X_1 - 0.6X_2 + 0.5X_3)/\sqrt{1.25}, \\ H_2 &= \sin(a_0 X_4 X_5)/\sigma_s, && \sigma_s^2 = \tfrac{1}{2}\{1 - (1+4a_0^2)^{-1/2}\}, \\ H_3 &= \left[\cos\{a_0(X_6 - X_7)/\sqrt{2}\} - 1\right]/\sigma_c, && \sigma_c^2 = \tfrac{1}{2}(1 + e^{-2a_0^2}) - e^{-a_0^2}. \\ H_4 &= (X_8^2 + X_9^2)/2, \end{aligned} \tag{70}$$

The target log risk combines linear, interaction, periodic, and quadratic structure,

$$f_{\mathrm{cap}}(\mathbf{X}) = 1.25 \left(\sqrt{0.35} H_1 + \sqrt{0.30} H_2 + \sqrt{0.20} H_3 + \sqrt{0.15} H_4 \right). \tag{71}$$

The four independent components have unit variance, so the population log-risk standard deviation is 1.25.

For $v_j = 10(j/20)^{1.35}$, $j = 1, \ldots, 20$, the population baseline was a ten-piece PWE model with cutpoints $v_2, v_4, \ldots, v_{18}$ and rates

$$(0.05729, 0.04439, 0.02340, 0.01664, 0.01498, 0.01677, 0.02296, 0.03366, 0.04210, 0.03947), \tag{72}$$

giving an early peak, middle trough, and late rebound. Censoring was independent with $C \sim \mathrm{Exp}(\mathrm{rate} = 0.16)$. The population observed-event fraction is approximately 0.222. Each of the ten target cohorts contains 300 subjects and is paired with an independent target-distribution test cohort of 20,000.

We use the population survival law as the compatible teacher so that the experiment focuses directly on how external guidance interacts with student capacity. The teacher supplies population conditional failure probabilities on the 20 comparison intervals with endpoints $v_1, \ldots, v_{20}$.

Student capacity ladder. All students use the same continuous-time right-censored likelihood and the same 20-interval distillation interface, while differing in risk-function and baseline capacity:

1. **Linear:** a reference-centered linear risk score and a five-piece baseline with cutpoints v_4, v_8, v_{12}, v_{16};
2. **SimpleNN:** one width-4 SiLU hidden layer with a linear skip and the same five-piece baseline;
3. **ComplexNN:** a 64–32 SiLU network with a linear skip and a twenty-piece baseline with cutpoints $v_1, \ldots, v_{19}$.

The ladder therefore increases overall student flexibility through both the risk function and baseline representation; the corresponding joint state dimensions are 17, 73, and 2,976. The Linear and SimpleNN score priors use standard deviation 0.25, and the ComplexNN score prior uses 0.35; all three use equation 49 on their respective baseline grids. With ten baseline pieces in the truth, this ladder jointly varies nonlinear risk capacity, temporal resolution, and part of the prior specification. Each positive-weight fit is matched to its own student's $\eta = 0$ fit.

All three students are evaluated at fixed borrowing weights $\eta \in \{0, 3, 10\}$ on the same ten paired target cohorts, with $\eta = 0$ giving the corresponding target-only Bayesian PWE fit. Posterior prediction follows Appendices C and E. The two positive weights characterize the response to moderate and stronger compatible borrowing.

F.1 Compatible borrowing across student specifications

Figure 4 shows the absolute predictive outcomes for all three student specifications over $\eta \in \{0, 3, 10\}$. Without borrowing, greater flexibility does not pay off: ComplexNN achieves only a small improvement in discrimination while performing worst in mean IBS, survival RMSE, and predictive deviance, with the greatest variability across target cohorts; Linear and SimpleNN perform similarly. Borrowing from the compatible teacher reverses this pattern. At $\eta = 10$, ComplexNN attains the best mean value of all four metrics, both neural students continue to improve from $\eta = 3$ to 10, and Linear changes comparatively little.

Table 3 quantifies two complementary contrasts at $\eta = 10$: each student's gain over its matched target-only fit and the final accuracy differences across student specifications. Contrasts are computed within replicate before averaging, with positive values favoring borrowing or the first named student; brackets are paired Student-t 95% Monte Carlo intervals over the ten replicates. Borrowing gains increase markedly with student flexibility for both IBS and survival RMSE. From $\eta = 0$ to 10, the paired IBS reduction increases from 0.0044 for Linear to 0.0310 for SimpleNN and 0.0470 for ComplexNN. At $\eta = 10$, final accuracy is ordered ComplexNN, SimpleNN, and Linear, reversing the target-only pattern. Compatible guidance therefore allows additional student flexibility to translate into improved prediction rather than increased estimation instability.

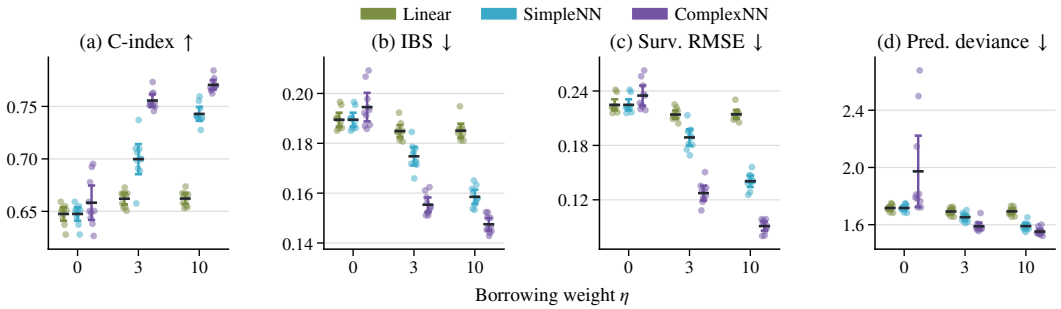


Figure 4: **Compatible borrowing makes greater student capacity useful.** Prediction of the three students at $\eta \in \{0, 3, 10\}$; $\eta = 0$ is each student's target-only Bayesian PWE fit. Colored points show ten paired replicates; dark ticks mark means and whiskers give Student-t 95% Monte Carlo intervals. Metrics use posterior-mean scores or survival curves, with predictive deviance based on the posterior-predictive likelihood.

Table 3: **Larger students benefit more from compatible borrowing.** Entries are mean error reductions in units of 10^{-2} at $\eta = 10$, with paired 95% Monte Carlo intervals across ten student-capacity replicates. The upper block compares final accuracy between students; the lower block measures each student's gain from $\eta = 0$ to 10. Positive values favor the first named student or borrowing, respectively.

Model / contrast	IBS (10^{-2})	Survival RMSE (10^{-2})
Final accuracy advantage at $\eta = 10$: first over second		
SimpleNN vs. Linear	2.66 [2.43, 2.89]	7.38 [6.72, 8.05]
ComplexNN vs. SimpleNN	1.10 [0.94, 1.26]	4.94 [4.34, 5.55]
Gain from target-only fitting ($\eta = 0$ to 10)		
Linear	0.44 [0.10, 0.78]	1.04 [0.36, 1.73]
SimpleNN	3.10 [2.75, 3.45]	8.43 [7.69, 9.17]
ComplexNN	4.70 [4.15, 5.25]	14.39 [13.25, 15.53]

F.2 Why larger students benefit: across-training variance and bias

We evaluate all fitted models on a common target-distribution reference cohort of 20,000 subjects, so that the decomposition reflects variation across repeated training fits rather than variation in the evaluation sample. Let $\widehat{S}_r(t \mid \mathbf{x})$ be the posterior-mean survival prediction from training replicate r and let $\overline{S} = 10^{-1} \sum_r \widehat{S}_r$. We compute

$$V = \left\langle \frac{1}{10} \sum_{r=1}^{10} (\widehat{S}_r - \overline{S})^2 \right\rangle, \qquad B^2 = \left\langle (\overline{S} - S^\star)^2 \right\rangle, \tag{73}$$

where $\langle \cdot \rangle$ averages equally over reference subjects and 103 fixed evaluation times. The grid augments 101 uniformly spaced points from v_1 to 20.6798 with the prespecified horizons 2.9112, 9.6977, and 20.6798 (rounded here). Using divisor 10 in the empirical variance gives the exact decomposition $\langle 10^{-1} \sum_r (\widehat{S}_r - S^\star)^2 \rangle = V + B^2$, attributing prediction error to across-training variance and squared bias.

Figure 5 shows why compatible guidance is most useful for the richest student. At $\eta = 0$, ComplexNN has substantially greater across-training variance than the simpler students despite lower squared bias, and borrowing sharply reduces both components. SimpleNN benefits mainly through lower squared bias, with comparatively little change in variance, whereas Linear retains substantial bias even after its variance decreases, reflecting the limits of a linear risk score for the nonlinear target law. The dominant source of predictive improvement therefore depends on student capacity.

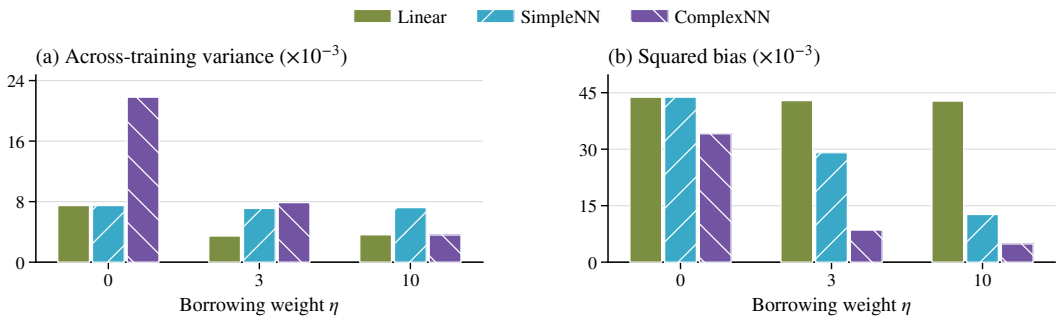


Figure 5: **ComplexNN gains through lower variance and bias.** Posterior-mean survival functions from the student-capacity experiment are evaluated on a common reference cohort. Panel (a) measures variation across ten independently trained fits; panel (b) measures squared bias of their average relative to truth. Bars show equation 73 multiplied by 10^3. Colors match Figure 4.

G SYNTHETIC TEACHER-MISMATCH EXPERIMENT: DESIGN AND SUPPLEMENTARY ANALYSES

This appendix gives the full design and supporting analyses for the teacher-compatibility experiment summarized in Section 4.2. It characterizes predictive borrowing paths, posterior centering and concentration, target-guided weight selection, and the behavior of the selected procedure. We also connect the empirical patterns to Propositions 1 and 2 and examine sensitivity to the structural prior.

G.1 DESIGN

Target distribution. We generated $X_1, \ldots, X_{12} \overset{\text{iid}}{\sim} N(0,1)$ and formed three unit-variance block summaries

$$Z_b = \frac{1}{2} \sum_{j=4b-3}^{4b} X_j, \qquad b = 1, 2, 3. \tag{74}$$

Following the nonlinear block construction used in earlier survival-distillation experiments (Deng et al., 2026a;b), the target log risk was

$$f_0(\mathbf{X}) = c_0 \sum_{b=1}^{3} w_b (0.6 Z_b + Z_b^2), \qquad \mathbf{w} = (0.6, 0.8, 1), \qquad c_0 = \left\{ (0.6^2 + 2) \sum_{b=1}^{3} w_b^2 \right\}^{-1/2}. \tag{75}$$

Hence $f_0(\mathbf{0}) = 0$ and $\mathrm{sd}\{f_0(\mathbf{X})\} = 1$. Writing $\mu_f = c_0 \sum_b w_b = 1.1047$, event times followed the PH model in equation 1 with PWE cutpoints $(2, 5, 10)$ and baseline rates

$$\boldsymbol{\lambda}_0 = e^{-\mu_f}(0.08, 0.04, 0.12, 0.06). \tag{76}$$

Censoring was independent, $C \sim \mathrm{Exp}(\mathrm{rate} = 0.24)$. Each of 20 replicates paired a target cohort of $n = 200$ with an independent target-distribution test cohort of 20,000 subjects. The target cohorts contained a mean of 51.5 observed failures (range 41–61), and each target–test pair was shared across teacher regimes and borrowing weights.

Teacher regimes. The compatible and risk-function-shifted teachers were fitted once to independent source cohorts of 50,000 subjects using a reference-centered 64–32–16 SiLU risk network with a linear skip and the same four-piece PWE representation. The compatible source followed the target law, yielding an estimated teacher aligned with the target population. For the risk-function-shifted source, only the block weights in equation 75 changed, to $(0.6, 0.8, -0.1)$, while the scale and baseline mechanism were retained.

The calibration-shifted teacher was constructed from the fitted compatible teacher by retaining its risk network and multiplying every baseline hazard by 1.5. Writing $S^e_{\mathrm{comp}}(t \mid \mathbf{x})$ and $S^e_{\mathrm{cal}}(t \mid \mathbf{x})$ for the two survival functions,

$$S^e_{\mathrm{cal}}(t \mid \mathbf{x}) = \{S^e_{\mathrm{comp}}(t \mid \mathbf{x})\}^{1.5}. \tag{77}$$

This transformation preserves the teacher's risk ordering and locally normalized risk-set probabilities while shifting its absolute survival scale. Each teacher was then held fixed across target replicates.

Student and borrowing path. The Bayesian student used a reference-centered 64–32–16 SiLU architecture with a linear skip, score-parameter prior standard deviation 0.25, and four jointly sampled baseline rates. The distillation interface used $K = 20$ intervals with endpoints at the target-training observed-duration quantiles $k/20$, while the target likelihood used the original observed event and censoring times.

For each teacher regime, we fitted the full borrowing path over the default grid in Appendix D, sharing the $\eta = 0$ target-only Bayesian PWE fit across regimes within each replicate. Posterior computation used four chains, each with 20,000 burn-in transitions and 60,000 retained post-thinning states (thinning factor 2). Posterior prediction and evaluation follow Appendices C and E.

G.2 Borrowing paths under teacher compatibility and mismatch

Predictive borrowing paths. Figure 6 traces prediction as borrowing strength increases. Under compatibility, all four mean metrics improve steadily toward direct teacher performance. Calibration shift leaves the discrimination path nearly unchanged while selectively degrading probability-scale prediction: IBS, survival RMSE, and predictive deviance move toward the higher error levels of the calibration-shifted teacher. Because the calibration transformation preserves teacher ranking, this contrast makes the role of the teacher's absolute survival scale directly visible.

Risk-function shift produces a different borrowing path. Prediction initially improves as the student receives useful external guidance, reaches its best mean performance near $\eta = 20$, and then deteriorates as stronger borrowing moves the student further toward the shifted covariate effects. Teacher compatibility therefore determines whether increasing borrowing continues to improve prediction or creates an interior trade-off.

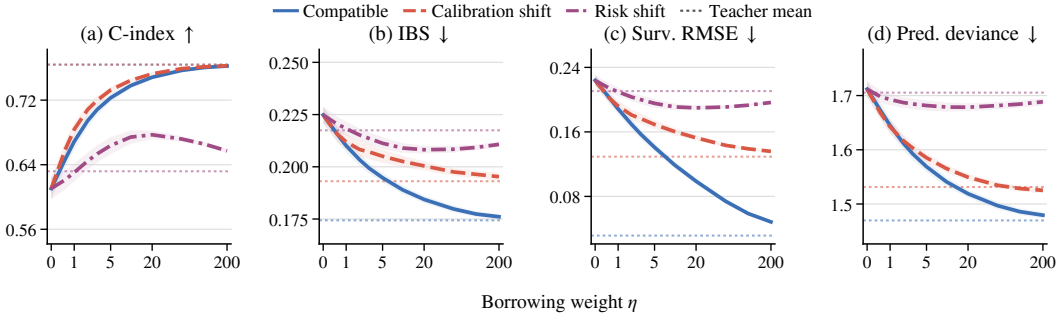


Figure 6: **Teacher compatibility shapes the predictive borrowing path.** C-index, IBS, survival-function RMSE, and predictive deviance are shown over fixed borrowing weights. Curves and bands give means and ±1.96 Monte Carlo standard errors across 20 replicates; dotted lines show direct teacher performance.

Posterior concentration and centering. Figure 7 first aggregates the fixed-η posterior behavior over the population event-time quartiles t_{25}, t_{50}, and t_{75}. Increasing borrowing strength narrows posterior intervals across all three teacher regimes. With the compatible teacher, this concentration is accompanied by small bias and higher truth inclusion. Under calibration and risk-function shift, the posterior centers move in opposite directions as η increases, and sufficiently strong borrowing reduces truth inclusion. These patterns separate two effects of borrowing: posterior concentration increases with teacher influence, whereas centering reflects teacher alignment.

Posterior centering and concentration across horizons. Figure 8 resolves the posterior borrowing paths at the population event-time quartiles $t_{25} = 3.692$, $t_{50} = 8.624$, and $t_{75} = 25.299$. Increasing borrowing narrows posterior intervals in all three teacher regimes, but the posterior centers respond differently to the information supplied by each teacher. Compatible guidance combines concentration

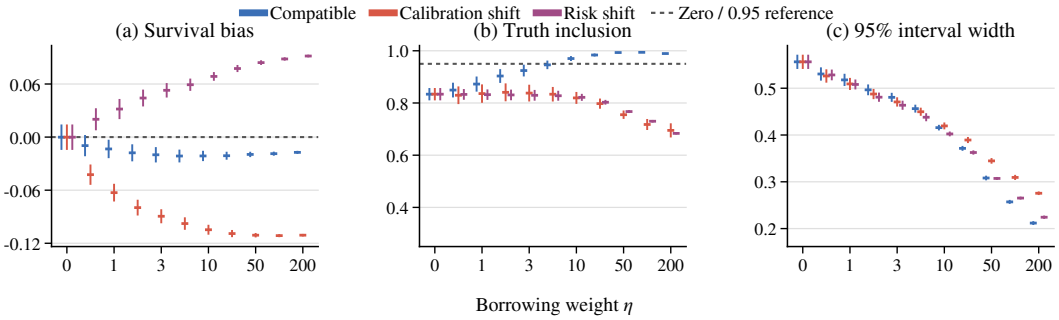


Figure 7: **Teacher alignment shapes posterior concentration and centering.** For individualized survival probability $S(t \mid \mathbf{x})$, the panels show signed bias, truth inclusion in nominal 95% pointwise posterior intervals, and mean interval width, averaged over t_{25}, t_{50}, and t_{75}. Ticks and bars give means and Student-t 95% Monte Carlo intervals across 20 replicates; dashed lines mark zero bias and 0.95 inclusion.

with small signed bias and high truth inclusion. Calibration shift produces increasingly negative survival bias later in follow-up, whereas risk-function shift produces positive bias across all three horizons. Under sufficiently strong mismatch, narrower intervals therefore accompany displacement of the posterior center and lower truth inclusion.

The fixed-weight paths separate two consequences of borrowing: within each teacher regime, stronger borrowing increases posterior concentration, while teacher alignment governs how the center moves as external guidance becomes stronger.

G.3 Target-guided borrowing selection

G.3.1 Predictive performance after WAIC selection

We next evaluate the procedure obtained when η is selected from the target data using Appendix D. Table 4 compares WAIC-selected BayCSD with the matched target-only Bayesian PWE fit within each replicate. Selected borrowing improves mean prediction across all three teacher regimes, with the largest gains under compatibility and progressively smaller gains as teacher mismatch becomes more consequential.

The calibration-shifted regime is particularly informative. Relative to the compatible teacher, it retains 96% of the C-index gain but only 56% of the IBS gain and 48% of the survival-RMSE gain. Thus, target-guided borrowing preserves most of the teacher's ranking benefit while probability-scale prediction remains sensitive to the shifted absolute survival information.

G.3.2 Selected borrowing strength and individualized posterior behavior

Table 5 connects the fixed-weight paths to the procedure obtained after target-guided selection. It reports the selected borrowing weight together with signed survival bias, truth inclusion, and posterior interval width for individualized $S(t \mid \mathbf{x})$.

Target-guided selection adapts the strength of borrowing. The selected weights differ sharply across teacher regimes. Under compatibility and calibration shift, WAIC favors strong borrowing, with median selected weights of 100 and 75, respectively. Under risk-function shift, the median selected weight falls to 5. Positive borrowing is selected in all replicates, but the substantially smaller weights under risk-function shift place the selected procedure closer to the weak-borrowing region of the fixed-weight paths.

Posterior concentration and centering reveal complementary effects of borrowing. Both the compatible and calibration-shifted regimes substantially narrow posterior intervals relative to the target-only fit, with mean widths of 0.2653 and 0.3460, respectively, compared with 0.5564 for target-only fitting. Their posterior locations differ more sharply. Compatible borrowing retains small

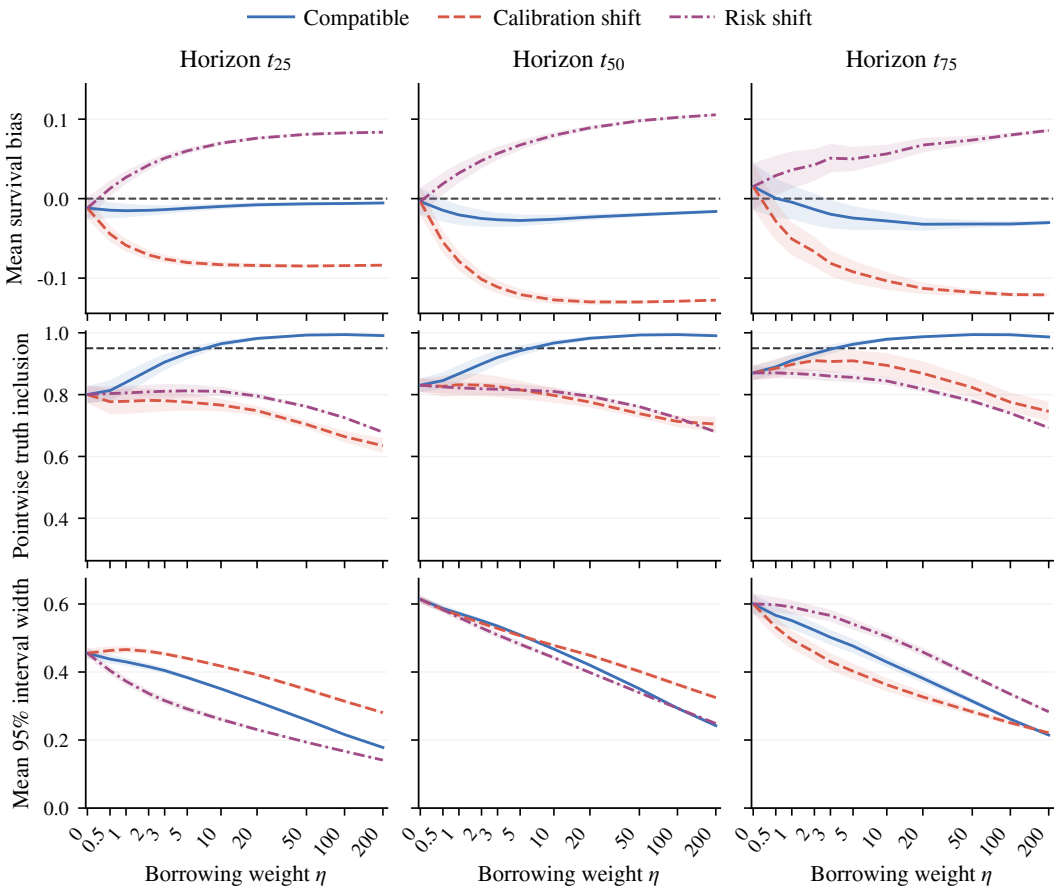


Figure 8: **Teacher compatibility shapes posterior centering and concentration across horizons.** For individualized survival probability $S(t \mid \mathbf{x})$, rows show signed bias, truth inclusion in nominal 95% pointwise posterior intervals, and mean interval width at t_{25}, t_{50}, and t_{75}. Curves and bands give means and ± 1.96 Monte Carlo standard errors across 20 replicates; dashed lines mark zero bias and 0.95 inclusion.

signed survival bias (-0.0188) and increases truth inclusion to 98.83%, whereas calibration shift produces a larger negative bias (-0.1062) and 75.57% truth inclusion. This contrast makes the transfer of absolute survival information directly visible, while indicating that concentration is driven mainly by borrowing strength and centering by teacher alignment. Under risk-function shift, the smaller selected weight yields weaker concentration (mean width 0.4190) and more limited center displacement, with truth inclusion remaining close to the target-only fit.

These results complement the fixed-weight analysis and the first-order mismatch result in Appendix B.4: target-guided selection determines how strongly teacher predictions enter the fitted procedure, while teacher alignment determines the centering of individualized survival estimates.

G.3.3 Comparison with cross-validation

We compare WAIC with event-stratified fivefold cross-validation on the first ten target replicates under all three teacher regimes in Table 6. Both selectors use the common candidate grid

$$\mathcal{E}_{\text{sel}} = \{0, 0.5, 1, 2, 3, 5, 10, 20, 50, 100, 200\}.$$

Cross-validation refits the generalized Bayesian student on four fifths of the target observations and selects η by held-out posterior-predictive negative log likelihood under the target model. Prediction is then evaluated using the corresponding full-$n = 200$ posterior fit.

Table 4: **Prediction gains from WAIC-selected borrowing.** BayCSD is compared with the shared target-only Bayesian PWE fit across 20 paired teacher-mismatch replicates. Positive gains favor borrowing; brackets give Student-t 95% Monte Carlo intervals, and counts give replicate-level improvements.

Teacher regime	Metric	WAIC-selected mean	Mean paired gain (95% CI)	Improved replicates
Compatible	C-index ↑	0.7574	0.1469 [0.1328, 0.1610]	20/20
	IBS ↓	0.1787	0.0461 [0.0415, 0.0508]	20/20
	Survival RMSE ↓	0.0637	0.1601 [0.1464, 0.1739]	20/20
	Predictive deviance ↓	1.4916	0.2202 [0.2037, 0.2367]	20/20
Calibration shift	C-index ↑	0.7515	0.1410 [0.1268, 0.1551]	20/20
	IBS ↓	0.1990	0.0258 [0.0225, 0.0292]	20/20
	Survival RMSE ↓	0.1469	0.0769 [0.0671, 0.0867]	20/20
	Predictive deviance ↓	1.5453	0.1666 [0.1479, 0.1853]	20/20
Risk-function shift	C-index ↑	0.6638	0.0533 [0.0375, 0.0691]	19/20
	IBS ↓	0.2102	0.0146 [0.0107, 0.0184]	20/20
	Survival RMSE ↓	0.1941	0.0297 [0.0226, 0.0368]	20/20
	Predictive deviance ↓	1.6766	0.0352 [0.0215, 0.0489]	19/20

Table 5: **Target-guided borrowing and individualized posterior behavior across teacher regimes.** Bias is the signed error of posterior-mean $S(t \mid \mathbf{x})$; width and truth inclusion concern nominal 95% pointwise posterior intervals. Bias and width are in units of 10^{-2} and inclusion is a percentage. Metrics are averaged equally over t_{25}, t_{50}, and t_{75} within each replicate and summarized across 20 replicates as means [Student-t 95% Monte Carlo intervals]. Selected weights are medians [IQRs].

Method / teacher regime	Selected η median [IQR]	Survival-probability bias (10^{-2})	Truth inclusion (%)	95% interval width (10^{-2})
Target-only Bayesian PWE	0 (fixed)	-0.01 [-1.45, 1.43]	83.37 [81.01, 85.73]	55.64 [54.12, 57.17]
Compatible	100 [50, 200]	-1.88 [-2.16, -1.59]	98.83 [98.20, 99.47]	26.53 [23.39, 29.67]
Calibration shift	75 [42.5, 200]	-10.62 [-11.17, -10.08]	75.57 [72.15, 78.99]	34.60 [31.12, 38.07]
Risk-function shift	5 [2, 20]	5.83 [4.51, 7.15]	82.11 [79.25, 84.97]	41.90 [37.44, 46.35]

Table 6: **Prediction after WAIC or cross-validation selects the borrowing weight.** Selected-model and best-path values are shown in units of 10^{-2}; gaps to the best fitted candidate are in units of 10^{-3}. Entries are means over ten paired replicates.

Teacher regime	Metric	WAIC selected (10^{-2})	5-fold CV selected (10^{-2})	Best fitted candidate (10^{-2})	WAIC gap (10^{-3})	CV gap (10^{-3})
Compatible	IBS	17.82	17.74	17.70	1.2	0.3
	Survival RMSE	5.50	5.00	4.75	7.5	2.5
Calibration shift	IBS	19.80	19.70	19.60	2.0	1.0
	Survival RMSE	14.15	13.82	13.52	6.3	3.0
Risk-function shift	IBS	21.12	21.37	20.70	4.2	6.6
	Survival RMSE	19.86	20.51	18.81	10.5	17.0

For reference, we also report the best fitted candidate on each held-out test path and the gap between the selected predictor and that candidate.

The two selectors produce similar predictions overall, but their relative behavior changes with teacher alignment. Under compatibility and calibration shift, cross-validation is slightly closer to the best fitted candidate. Under risk-function shift, where the fixed-weight paths exhibit a clear interior borrowing optimum, WAIC is closer to the best fitted candidate for both IBS and survival RMSE. Thus, WAIC gives up little when the teacher remains useful while showing a more conservative predictive response when stronger teacher mismatch makes aggressive borrowing less favorable.

G.4 Empirical counterparts to the theoretical results

G.4.1 Posterior and sampling variability under compatible borrowing

Proposition 1 distinguishes the sampling variability of the fitted center from the spread of the generalized posterior. Table 7 illustrates this distinction in a correctly specified finite-dimensional model. The seven-parameter student uses the three nonlinear features in equation 75 and the true four-piece baseline, with the same structural-prior settings as the primary experiment. The teacher is the population survival law.

A $K = 20$ comparison partition, defined by observed-duration quantiles in an independent design cohort of 200,000 subjects, is held fixed across fits. At each of $n = 200$ and 1000, we generate 100 independent target-training cohorts and fit each at $\eta \in \{0, 20, 200\}$. Every fit is evaluated on the same independent set of 20,000 reference covariates at t_{25}, t_{50}, and t_{75}.

To visualize the covariance distinction in Proposition 1, we compare raw posterior intervals with intervals obtained after rescaling posterior draws to the estimated sampling covariance. Let $\bar{\boldsymbol{\vartheta}}$ and $\widehat{C}_{\mathrm{post}}$ denote the empirical mean and covariance of the joint posterior draws, and let $\widehat{I}_R$ be the summed target negative-log-likelihood Hessian evaluated at the MAP. We set

$$\widehat{C}_{\mathrm{samp}} = \widehat{C}_{\mathrm{post}} \widehat{I}_R \widehat{C}_{\mathrm{post}},$$

and transform each posterior draw as

$$\boldsymbol{\vartheta}_{\mathrm{res}}^{(m)} = \bar{\boldsymbol{\vartheta}} + \widehat{C}_{\mathrm{samp}}^{1/2} \widehat{C}_{\mathrm{post}}^{-1/2} \left(\boldsymbol{\vartheta}^{(m)} - \bar{\boldsymbol{\vartheta}} \right),$$

using symmetric matrix roots. The transformation preserves the fitted center while replacing the empirical posterior covariance with its sampling-covariance counterpart.

Table 7: **Posterior and sampling-covariance-rescaled intervals under compatible borrowing.** Each entry reports truth inclusion (%) ± Monte Carlo standard error, followed by mean interval width in parentheses. Inclusion is evaluated across 100 independent training replicates at each fixed reference covariate and then averaged over 20,000 reference covariates.

η	Interval	t_{25}	t_{50}	t_{75}
$n = 200$				
0	Raw posterior	87.4 ± 2.0 (0.1832)	89.1 ± 2.0 (0.2605)	95.2 ± 1.6 (0.3470)
	Rescaled	85.9 ± 2.1 (0.1787)	87.4 ± 2.1 (0.2496)	94.6 ± 1.7 (0.3517)
20	Raw posterior	100.0 ± 0.0 (0.0400)	100.0 ± 0.0 (0.0627)	100.0 ± 0.0 (0.1775)
	Rescaled	83.5 ± 2.4 (0.0090)	87.2 ± 2.3 (0.0147)	92.2 ± 2.1 (0.1112)
200	Raw posterior	100.0 ± 0.0 (0.0129)	100.0 ± 0.0 (0.0204)	100.0 ± 0.0 (0.1130)
	Rescaled	83.8 ± 2.4 (0.0010)	86.6 ± 2.3 (0.0016)	91.1 ± 2.3 (0.0888)
$n = 1000$				
0	Raw posterior	95.0 ± 1.3 (0.0760)	96.5 ± 1.1 (0.1151)	97.0 ± 1.2 (0.2058)
	Rescaled	94.9 ± 1.3 (0.0757)	96.3 ± 1.1 (0.1143)	97.1 ± 1.2 (0.2097)
20	Raw posterior	100.0 ± 0.0 (0.0168)	100.0 ± 0.0 (0.0269)	100.0 ± 0.0 (0.0583)
	Rescaled	94.8 ± 1.3 (0.0038)	96.6 ± 1.2 (0.0064)	95.6 ± 1.5 (0.0153)
200	Raw posterior	100.0 ± 0.0 (0.0055)	100.0 ± 0.0 (0.0087)	100.0 ± 0.0 (0.0192)
	Rescaled	94.1 ± 1.4 (0.0004)	96.5 ± 1.2 (0.0007)	94.4 ± 1.7 (0.0016)

Compatible borrowing sharply narrows the raw posterior intervals at both sample sizes while their repeated truth inclusion remains conservative at the positive borrowing weights. At $n = 1000$ and t_{50}, for example, mean raw width decreases from 0.1151 at $\eta = 0$ to 0.0269 and 0.0087 at $\eta = 20$ and 200. This pattern reflects the distinction in Proposition 1: posterior spread and sampling variability of the fitted center are governed by different covariance operators under borrowing.

After sampling-covariance rescaling, inclusion at positive weights ranges from 83.5% to 92.2% at $n = 200$ and from 94.1% to 96.6% at $n = 1000$. The larger-sample results approach nominal 95% inclusion, whereas the smaller-sample results show stronger finite-sample departure from the local approximation. The experiment therefore makes the posterior–sampling covariance distinction visible directly on individualized survival probabilities.

Table 8: **Finite-resolution behavior under comparison-partition refinement.** At fixed $\eta = 20$ and with the student model unchanged, the table reports prediction error, posterior-mean survival RMSE relative to the $K = 40$ fit, and nominal 95% pointwise posterior-interval behavior. Truth inclusion and interval width are averaged over t_{25}, t_{50}, and t_{75}. Entries are means ± 1.96 Monte Carlo standard errors over ten paired replicates.

K	IBS (10^{-2})	Survival RMSE (10^{-2})	RMSE to $K = 40$ prediction (10^{-2})	Truth inclusion (%)	95% width (10^{-2})
5	18.63 ± 0.22	10.34 ± 0.38	2.54 ± 0.55	98.29 ± 0.65	38.57 ± 0.57
10	18.54 ± 0.24	9.85 ± 0.43	1.44 ± 0.19	98.58 ± 0.57	37.43 ± 0.84
20 (primary)	18.52 ± 0.25	9.75 ± 0.38	1.24 ± 0.18	98.64 ± 0.53	37.03 ± 0.82
40	18.53 ± 0.26	9.73 ± 0.41	0.00 ± 0.00	98.66 ± 0.49	36.39 ± 0.52

G.4.2 Comparison-partition refinement

Proposition 2 characterizes the limiting distillation criterion as the comparison partition is refined while the student model remains fixed. We examine the corresponding finite-resolution behavior in Table 8 using the first ten primary replicates, the compatible teacher, and fixed $\eta = 20$. We compare $K \in \{5, 10, 20, 40\}$, with endpoints at the target-training observed-duration quantiles k/K and a common terminal support. The target and test cohorts, teacher, student architecture, four-piece baseline partition, continuous-time target likelihood, priors, posterior computation, and evaluation grid are unchanged across K.

In addition to prediction and posterior-interval summaries, we measure the root-mean-square difference between the posterior-mean survival prediction at each K and the prediction obtained at $K = 40$, evaluated on the same 20,000 test subjects and survival-RMSE time grid.

Prediction and posterior summaries change little from $K = 10$ onward. At the same time, posterior-mean survival predictions move closer to those from the finest evaluated partition as K increases. These results provide a finite-resolution counterpart to Proposition 2: refining the probability-comparison interface produces stable prediction and interval summaries without changing the student's temporal representation.

G.5 Structural-prior sensitivity

The generalized update separates external predictive guidance from structural regularization. We examine their interaction by perturbing the structural prior under the compatible teacher. The primary specification uses score-parameter prior standard deviation $\sigma_\theta = 0.25$ and the baseline prior in equation 49 with mean $m_0 = 0.1$ and concentration $\kappa = 5$. We vary one component at a time: $\sigma_\theta \in \{0.2, 0.3\}$, $m_0 \in \{0.05, 0.2\}$, and $\kappa \in \{2.5, 10\}$.

Table 9 evaluates each specification on the same ten paired target cohorts at fixed $\eta \in \{0, 20, 200\}$. Each positive-weight fit is compared with the matched target-only Bayesian PWE fit under the same structural prior.

Table 9: **Prediction across structural-prior specifications under compatible borrowing.** Mean IBS and survival-function RMSE (both in units of 10^{-2}) are reported over ten paired replicates. Each alternative changes one component of the default $(\sigma_\theta, m_0, \kappa) = (0.25, 0.1, 5)$.

		$\eta = 0$		$\eta = 20$		$\eta = 200$	
Varied component	Setting	IBS	Surv. RMSE	IBS	Surv. RMSE	IBS	Surv. RMSE
Default	$(\sigma_\theta, m_0, \kappa) = (0.25, 0.1, 5)$	22.31	22.39	18.52	9.75	17.70	4.75
Score prior SD	$\sigma_\theta = 0.2$	22.98	23.80	18.61	10.02	17.68	4.50
	$\sigma_\theta = 0.3$	21.78	21.19	18.56	10.04	17.78	5.48
Baseline mean	$m_0 = 0.05$	22.11	21.96	18.49	9.67	17.71	4.76
	$m_0 = 0.2$	22.87	23.50	18.60	10.03	17.72	4.83
Baseline concentration	$\kappa = 2.5$	22.23	22.21	18.51	9.69	17.71	4.79
	$\kappa = 10$	22.43	22.66	18.56	9.90	17.71	4.78

Structural-prior perturbations produce visible differences in the sparse-data target-only fits, whereas compatible borrowing brings predictive performance into a substantially narrower range across prior specifications. Both $\eta = 20$ and $\eta = 200$ improve IBS and survival RMSE over the matched target-only fit for every tested specification and all ten paired cohorts. Across-prior ranges under borrowing are only 2–6% of the corresponding default-prior gains over $\eta = 0$. Thus, compatible teacher guidance improves prediction while reducing the target learner's practical sensitivity to local structural-prior choices.

H REAL-DATA POOLED-REFERENCE BENCHMARK

This appendix provides the full design for Section 4.3. The benchmark evaluates how much of the predictive information available under raw-data pooling can be recovered when the target learner instead has access only to predictions from an external model.

H.1 WITHIN-POPULATION BENCHMARK CONSTRUCTION AND INFORMATION ROLES

The benchmark uses SUPPORT (Knaus et al., 1995), METABRIC (Curtis et al., 2012), and the 2017–2018 OPTN kidney-alone (KI) cohort. Within each dataset, the internal, external, and test subsets are disjoint samples from the same underlying cohort rather than from distinct source and target populations. This within-population design allows a model fitted to the union of the internal and external subsets to serve as a high-information reference for what could be learned if the external raw records were available, while separating information loss from prediction-only access from the population-shift setting studied in Section 4.4.

The held-out test set contains 20% of the records. SUPPORT contains 1,775 test subjects (1,207 events), 100 internal subjects (68 events), and 6,998 external subjects (4,761 events); METABRIC contains 381 test subjects (221 events), 110 internal subjects (64 events), and 1,413 external subjects (818 events); and OPTN KI contains 4,889 test recipients (274 events), 750 internal recipients (42 events), and 18,806 external recipients (1,055 events). The SUPPORT and METABRIC internal subsets are sampled without outcome-dependent undersampling, preserving the event fractions of the original cohorts.

The four information roles are defined within each split. The external teacher is fitted to the external subset; target-only Bayesian PWE and BayCSD are fitted to the internal subset; the pooled-data reference is fitted to the union of the external and internal subsets; and all predictors are evaluated on the same held-out test set. Preprocessing and model-specific transformations are estimated from the non-test records available to the corresponding information role and then applied to query or test subjects.

Individualized RMST uses restriction horizons of 365 days for SUPPORT and OPTN KI and 120 months for METABRIC. SUPPORT and METABRIC use the processed 14- and 9-covariate DeepSurv representations, respectively, whereas OPTN KI uses the same 129-field clinical representation as the KI-to-KP application. One censored METABRIC record has reported time zero and is retained at 0.05 months, half the 0.1-month resolution of the processed data; there are no zero-time events.

H.2 MATCHED MODELS AND BORROWING

Within each dataset, the pooled-data reference, target-only Bayesian PWE, and BayCSD use the same prespecified reference-centered neural PWE family, baseline-partition rule, and structural prior. The risk network has hidden widths 32–16, SiLU activations, and a linear skip. Model time is native time divided by the dataset-specific RMST horizon. Seven baseline cuts occur at fractions

$$\{0.02, 0.04, 0.08, 0.16, 0.25, 0.50, 0.75\}$$

of that horizon, while 20 comparison endpoints and 50 evaluation times use a common power-1.5 grid through the horizon. The score-prior standard deviation is 0.25, and the baseline prior uses mean 0.1 per horizon and concentration 5, with a $\mathrm{Gamma}(1, 10)$ tail-rate prior.

The external teacher is a fixed joint-MAP PWE predictor fitted to the external subset. The pooled-data reference is fitted to the union of the external and internal subsets, with posterior-mean individualized survival function $\widehat{S}_{\mathrm{ref}}(t \mid \mathbf{x})$. Target-only Bayesian PWE and BayCSD are fitted to the internal

Table 10: **Held-out prediction for the external teacher and pooled-data reference.** The teacher is fitted to the external subset, whereas the pooled-data reference is fitted to the union of the external and internal raw records. Entries are means ± sample standard deviations across the same three held-out splits as Table 1.

Dataset	Predictor	IBS (10^{-2}) ↓	C-index ↑
SUPPORT	Teacher	21.87 ± 0.37	0.6185 ± 0.0077
	Pooled reference	21.75 ± 0.43	0.6214 ± 0.0088
METABRIC	Teacher	13.84 ± 0.55	0.6339 ± 0.0103
	Pooled reference	13.69 ± 0.58	0.6413 ± 0.0090
OPTN KI	Teacher	3.34 ± 0.06	0.7671 ± 0.0094
	Pooled reference	3.04 ± 0.10	0.7857 ± 0.0039

subset, with BayCSD additionally receiving predictions from the external teacher. The pooled-data fit therefore provides the high-information reference against which recovery from prediction-level distillation is assessed.

The pooled-data reference, target-only Bayesian PWE, and BayCSD use four preconditioned MALA chains, each with 8,000 burn-in transitions and 32,000 retained post-thinning states (thinning factor 2). Every tenth retained state is used for prediction and WAIC summaries. BayCSD selects $\widehat{\eta}$ using the target-guided WAIC rule in Appendix D. Dataset-specific comparison and evaluation grids are restricted to common follow-up and censoring support.

H.3 Recovering pooled-data information

We assess recovery on two complementary scales. IBS and Harrell's C-index measure prediction of held-out outcomes. Recovery of the pooled-data reference is measured by the distance between posterior-mean individualized survival functions using equation 68 and by the corresponding individualized RMST error from equation 69. All summaries are computed within split and then averaged equally across the three splits.

Table 10 completes the information ladder in Table 1 by reporting held-out prediction for the external teacher and pooled-data reference. On SUPPORT and METABRIC, the two high-information predictors have similar held-out performance. On OPTN KI, combining the internal records with the external data improves both IBS and C-index relative to the teacher fitted to the external subset alone. Together with Table 1, these results locate BayCSD between sparse internal-only learning and the predictors available from substantially more information: prediction-level distillation moves the internal learner toward the high-information end of the benchmark while retaining evaluation against independent outcomes.

H.4 Posterior intervals at matched pooled-reference inclusion

We next compare posterior concentration relative to the pooled-data reference. Because interval width and reference recovery are distinct quantities, we compare widths at common achieved inclusion of the split-matched pooled-reference prediction. Subject-specific survival and RMST draws generate central posterior intervals over nominal levels

$$c \in \{0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.975, 0.99\}.$$

Pooled-reference inclusion and interval width follow equation 64–equation 67. Figure 3 in the main text traces the resulting inclusion–width curves for individualized survival probability. The corresponding RMST analysis is shown in Figure 9 below, and Table 11 reports interpolated widths at 95% achieved pooled-reference inclusion for both functionals.

The comparison is performed within each split before averaging across splits. Survival summaries average equally over test subject–time pairs on the 50-point evaluation grid, whereas RMST summaries average over test subjects. The pooled-data reference is represented by its posterior-mean survival prediction throughout this recovery analysis.

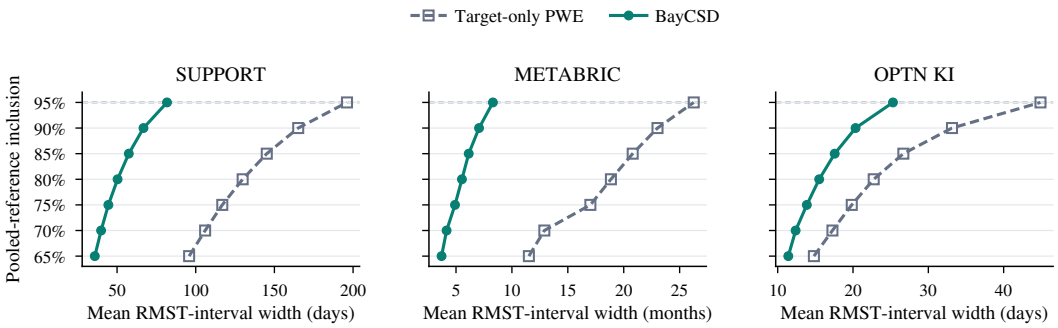


Figure 9: **BayCSD yields narrower RMST intervals at matched pooled-reference inclusion.** Columns correspond to datasets. Curves trace posterior RMST-interval width against achieved inclusion of the pooled-data reference as the nominal credible level varies; displayed widths are interpolated within each split at common achieved inclusion levels from 65% to 95%. Leftward curves indicate narrower intervals at matched inclusion, and the dashed line marks 95% achieved inclusion. RMST width is reported in days for SUPPORT and OPTN KI and in months for METABRIC.

Table 11: **Posterior interval width at 95% achieved pooled-reference inclusion.** Survival-probability widths are in units of 10^{-2}; RMST widths are in days for SUPPORT and OPTN KI and months for METABRIC. Entries are means $\pm$ standard deviations across three matched splits.

Dataset	Method	Survival $W^{95,\mathrm{ref}}$ (10^{-2})	RMST $W^{95,\mathrm{ref}}$
SUPPORT	Target-only Bayesian PWE	50.6 ± 4.4	196.1 ± 14.6
	BayCSD	20.7 ± 5.9	81.7 ± 23.9
METABRIC	Target-only Bayesian PWE	20.2 ± 4.5	26.3 ± 5.4
	BayCSD	5.4 ± 1.2	8.3 ± 2.0
OPTN KI	Target-only Bayesian PWE	11.2 ± 0.7	44.9 ± 1.7
	BayCSD	6.4 ± 0.7	25.3 ± 2.8

At matched 95% pooled-reference inclusion, BayCSD yields narrower intervals for both individualized survival probability and RMST on all three datasets. These reductions complement the point-prediction recovery in Table 1: prediction-level distillation moves the sparse target learner toward the pooled-data reference while concentrating individualized posterior predictions around that reference more efficiently.

I OPTN APPLICATION: COHORT, MODELS, EVALUATION, AND SUPPLEMENTARY RESULTS

This appendix provides the cohort construction, KI teacher, KP learners, controlled BayCSD variants, repeated cross-validation and evaluation protocol, and posterior-computation diagnostics for Section 4.4.

I.1 COHORTS AND KI TEACHER

The source and target cohorts include adult deceased-donor recipients transplanted from January 1, 2017 through December 31, 2018. KI recipients are restricted to those without a previous kidney transplant. Records with kidney graft-failure dates before transplantation are excluded, and KP records additionally exclude pre-transplant pancreas-failure dates. The endpoint is the first occurrence of death or kidney graft failure within one year after transplantation; pancreas graft failure alone is not an endpoint. Follow-up is administratively capped at 365.25 days, and recorded same-day events are assigned 0.5 days. Center eligibility is determined separately within KI and KP using a minimum of more than 10 eligible recipients over the two-year window. The resulting KI source

contains 24,445 recipients and 1,371 events across 201 centers, and the KP target contains 1,257 recipients and 51 events across 54 centers.

The input schema contains 129 clinical fields before training-specific constant-column removal. Center indicators and identifiers, transplant dates, and record identifiers are excluded from model inputs. Fixed-category filling and mean imputation are followed by constant-column removal and scaling estimated from the corresponding model-training records.

The KI teacher is a neural PWE model fitted with the continuous-time right-censored likelihood and a Gaussian prior on network weights. An event-stratified 80/20 KI split compares risk networks with hidden widths 32–16, 64–64, and 64–64–64, prior standard deviations in $\{0.1, 0.25, 0.5\}$, and two optimization seeds per configuration. Validation negative log likelihood selects the architecture, prior scale, seed, and training checkpoint. The selected 32–16 network with prior standard deviation 0.25 is then refitted to the complete KI cohort for the selected training duration. Teacher development uses KI data only.

I.2 KP LEARNERS AND POSTERIOR COMPUTATION

Bayesian PWE family. The BayCSD student is a reference-centered 64–32 SiLU network with score-prior standard deviation 0.25. The PWE baseline has cuts at 7, 14, 30, 60, 90, 180, and 270 days, followed by an open tail. Computation measures time in years using 365.25 days per year. In the prior of Appendix C, we set $m_0 = 0.05$ per year and $\kappa = 5$, with

$$\lambda_J \sim \mathrm{Gamma}(1, 20)$$

for the tail rate under the shape–rate parameterization.

All interval-based distillation in the OPTN experiment uses the common monthly comparison partition

$$\tau_m = 365.25\, m/12, \qquad m = 1, \ldots, 12.$$

This comparison partition is distinct from the PWE baseline partition. BayCSD compares teacher and student conditional failure probabilities on the monthly intervals while the target likelihood uses the original observed event and censoring times.

Target-only Bayesian PWE uses the same Bayesian student model without distillation. Grouped Bayesian PWE uses the same student architecture, baseline representation, and structural prior, but fits target outcomes through the monthly grouped likelihood defined in Appendix I.3. These models provide the target-only controls for BayCSD and the grouped-likelihood controlled variant, respectively.

Other matched learners. LogisticHazard (Kvamme & Borgan, 2021) models KP outcomes on the common 12-month partition, and DiSKD (Deng et al., 2026a) augments the same learner with KL distillation between teacher and student conditional event probabilities.

DeepSurv (Katzman et al., 2018) uses a scalar Cox log-risk output and the original event-time risk sets. A Breslow baseline estimated from the corresponding outer-training records converts its fitted score into survival probabilities for evaluation. Component-wise KL (Deng et al., 2026b) augments the same Cox learner with KL divergence between teacher- and student-implied risk-set distributions.

Posterior computation. Bayesian PWE-based methods use the joint posterior computation described in Appendix C. A joint MAP fit initializes four preconditioned MALA chains. Each chain uses 8,000 burn-in transitions and retains 32,000 post-thinning states (thinning factor 2), with step-size adaptation restricted to the first 4,000 burn-in transitions. The initial step size is 10^{-4}; the diagonal preconditioner is capped at 2 before its baseline block is multiplied by 10. Every tenth retained state is used for prediction, yielding 12,800 joint score–baseline posterior states per fitted model. All reported Bayesian PWE predictions use posterior-mean survival functions computed from these states.

I.3 CONTROLLED BAYCSD VARIANTS

The two controlled variants modify one component of BayCSD while retaining the remaining Bayesian PWE specification.

Bayesian PWE + risk-set distillation. This variant retains the continuous-time PWE target likelihood but replaces BayCSD's interval-probability distillation criterion with event-time risk-set KL. For each observed training event i, let

$$\mathcal{R}_i = \{j : Y_j \geq Y_i\}$$

denote its risk set. Writing r_j^e for the fixed KI teacher's log-risk prediction,

$$\begin{aligned} \pi^e_{j|i} &= \frac{\exp(r_j^e)}{\sum_{\ell\in\mathcal{R}_i}\exp(r_\ell^e)}, \qquad & \pi_{j|i}(\boldsymbol{\theta}) &= \frac{\exp\{f_{\boldsymbol{\theta}}(\mathbf{X}_j)\}}{\sum_{\ell\in\mathcal{R}_i}\exp\{f_{\boldsymbol{\theta}}(\mathbf{X}_\ell)\}}, \qquad & j\in\mathcal{R}_i, \\ Q_n^{\mathrm{risk}}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i:\Delta_i=1} D_{\mathrm{KL}}\Big\{\pi^e_{\cdot|i} \,\|\, \pi_{\cdot|i}(\boldsymbol{\theta})\Big\}. \end{aligned} \tag{78}$$

Tied events use the common risk-set KL with Breslow multiplicity, and subjects censored at the event time remain in the risk set. The generalized posterior uses $n\eta Q_n^{\mathrm{risk}}$ in place of BayCSD's interval-probability criterion, with the same target likelihood and structural prior. Risk-set normalization removes the teacher's baseline scale from the distillation criterion.

Grouped Bayesian PWE + interval distillation. This variant retains BayCSD's interval-probability distillation on the common 12-month comparison partition but replaces the continuous-time target likelihood with a grouped target likelihood on that partition. An observed event contributes survival through the left endpoint of its event month and the conditional failure probability within that month. A censored observation contributes survival through the last completed monthly endpoint. Monthly cumulative-hazard increments are computed by overlap with the unchanged PWE baseline intervals, as in equation 15.

Apart from the stated changes, BayCSD and the two controlled variants share the KI teacher, student architecture, PWE baseline partition, structural prior, posterior-computation settings, cross-validation splits, and borrowing-weight candidate grid. At $\eta = 0$, Bayesian PWE + risk-set distillation reduces to Bayesian PWE, whereas Grouped Bayesian PWE + interval distillation reduces to Grouped Bayesian PWE.

I.4 Repeated cross-validation and predictive evaluation

We use five independently prespecified event-stratified 5-fold partitions of the KP cohort, shared across all methods. Each recipient is held out once within each repeat. Training folds contain 1,005–1,006 recipients and 40–41 events, and test folds contain 251–252 recipients and 10–11 events.

All model selection is performed within the corresponding outer-training fold. Distillation weights use the common candidate set

$$\mathcal{E}_{\mathrm{OPTN}} = \{0, 0.1, 0.3, 1, 3, 10, 30, 100, 200\}.$$

BayCSD and the two Bayesian controlled variants select η by WAIC evaluated under their respective target-data likelihoods. LogisticHazard, DiSKD, DeepSurv, and component-wise KL use a 64–32 SiLU trunk without normalization or dropout, a method-specific output head, minibatches of 32, and at most 200 epochs. A shared event-stratified inner 5-fold partition selects learning rate from $\{10^{-4}, 10^{-3}\}$, weight decay from $\{10^{-4}, 10^{-2}\}$, one of two prespecified initializations, and stopping epoch by mean validation predictive deviance. DiSKD and component-wise KL select their distillation weights from $\mathcal{E}_{\mathrm{OPTN}}$ using the same inner validation scheme. Selected optimization-based models are refitted on the complete outer-training fold before held-out evaluation.

IPCW weights are estimated from the corresponding KP training fold. IBS is evaluated from predicted survival functions over the continuous one-year follow-up scale using Appendix E, and time-dependent concordance compares predicted survival at the original observed event times using the same appendix. Within each cross-validation repeat, fold-level metrics are averaged equally over the five held-out folds. Table 2 reports the mean and sample standard deviation across the five repeat-level estimates.

I.5 Posterior interval analysis

We compare posterior uncertainty in individualized survival probabilities under target-only Bayesian PWE and BayCSD. For each held-out recipient, the 12,800 retained joint posterior states are used

to evaluate survival probability at 30, 90, 180, and 365 days. Pointwise 95% posterior intervals are defined by the 2.5th and 97.5th percentiles of the corresponding draw-wise survival probabilities.

Within each cross-validation repeat, the five mutually exclusive held-out folds provide one out-of-fold prediction for every KP recipient. At each horizon, we summarize the mean interval width under target-only Bayesian PWE and BayCSD, the relative reduction in mean width, and the proportion of recipients whose BayCSD interval is narrower. Table 12 reports the mean and sample standard deviation of these summaries across the five repeated cross-validation partitions.

For Figure 10(b), each recipient's 365-day interval width is averaged over its five out-of-fold predictions, yielding one target-only Bayesian PWE–BayCSD width pair for each of the 1,257 KP recipients.

Table 12: **Posterior interval widths for individualized survival probabilities in OPTN KP.** Widths refer to nominal 95% pointwise posterior intervals for $S(t \mid \mathbf{x})$. Entries are means $\pm$ standard deviations across five repeated cross-validation partitions.

Quantity	30 days	90 days	180 days	365 days	Average
Target-only Bayesian PWE: interval width (10^{-2})	4.00 ± 0.03	7.11 ± 0.05	11.02 ± 0.06	15.80 ± 0.07	9.48 ± 0.05
BayCSD: interval width (10^{-2})	1.23 ± 0.05	2.01 ± 0.09	2.57 ± 0.11	3.53 ± 0.15	2.33 ± 0.10
Relative width reduction	$(69.3 \pm 1.3)\%$	$(71.8 \pm 1.2)\%$	$(76.7 \pm 0.9)\%$	$(77.7 \pm 0.9)\%$	$(75.4 \pm 1.0)\%$
Recipients with narrower BayCSD interval	$(96.9 \pm 0.3)\%$	$(97.3 \pm 0.2)\%$	$(98.1 \pm 0.2)\%$	$(98.5 \pm 0.2)\%$	$(97.7 \pm 0.2)\%$

I.6 Matched predictive performance and recipient-level posterior concentration

Table 2 reports the repeat-level means for the three matched target-only/distillation pairs and the two controlled variants. Figure 10(a) complements those summaries by showing the matched changes across the five prespecified cross-validation repeats. Each distillation procedure moves its target-only learner toward lower IBS and higher C_{td}, despite the different target-only starting points. BayCSD finishes with the lowest mean IBS and highest mean concordance among the distilled procedures, while the repeat-level points show that the gain is not driven by a single partition.

Panel (b) visualizes the individualized posterior contraction quantified across horizons in Table 12. Each point compares a recipient's mean 365-day interval width under target-only Bayesian PWE and BayCSD across its five out-of-fold predictions. A total of 98.5% of recipients fall below the diagonal, while Table 12 shows mean width reductions of 69.3–77.7% over the four reported horizons. Thus, the predictive improvements under KI guidance are accompanied by broad recipient-level posterior concentration rather than a reduction confined to a small subset of the KP cohort.

I.7 Predictive-functional MCMC diagnostics

We assess MCMC mixing directly for the individualized survival probabilities used in prediction and interval summaries. Across all 25 outer-training-fold fits for target-only Bayesian PWE and WAIC-selected BayCSD, Table 13 reports the worst rank-normalized split $\widehat{R}$, bulk and tail effective sample size (ESS), and Monte Carlo standard error (MCSE) of the posterior mean over monitored recipient–horizon quantities at 30, 90, 180, and 365 days.

Table 13: **Predictive-functional MCMC diagnostics in OPTN KP.** Diagnostics summarize individualized survival probabilities across all 25 outer-training-fold fits per method; MCSE is on the survival-probability scale.

Fit	Max. $\widehat{R}$	Min. bulk ESS	Min. tail ESS	Max. mean MCSE
Target-only Bayesian PWE	1.0037	2502	4127	2.87×10^{-3}
BayCSD (WAIC-selected)	1.0049	1297	1742	6.93×10^{-4}

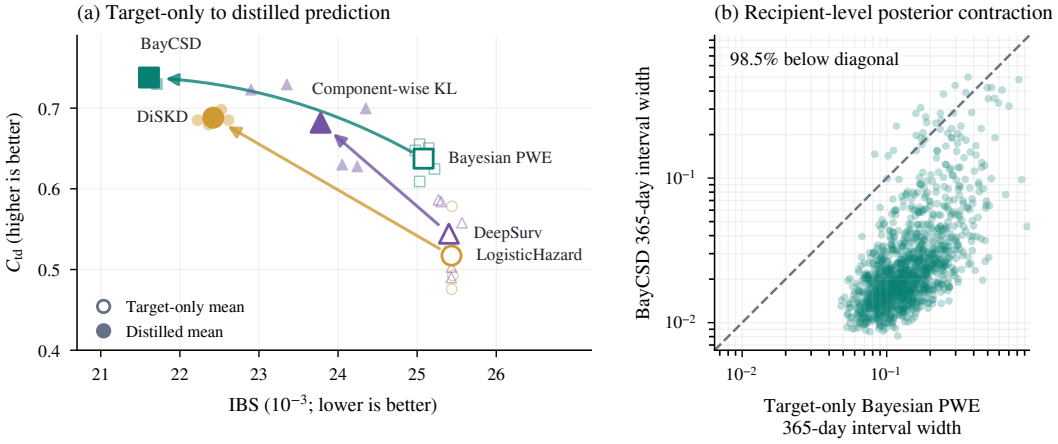


Figure 10: **Prediction and posterior concentration in KI-to-KP transfer.** (a) Held-out IBS and C_{td} for the three matched target-only/distillation pairs. Open markers denote target-only learners and filled markers their distilled counterparts; small markers are cross-validation repeat means and large markers are means over five repeats. Arrows connect the matched means, showing both the starting point and final prediction; upper-left positions indicate better performance. (b) Recipient-level widths of nominal 95% posterior intervals for 365-day survival probability under target-only Bayesian PWE and BayCSD. Points below the diagonal have narrower BayCSD intervals.