

Supplementary Material for “A Unified Survival Benchmark for Temporal Dropout Risk Prediction in Learning Analytics”

Rafael da Silva^{1*}, Jeff Eicher¹ and Gregory Longo¹

^{1*}Applied Data Science Program, Eastern University, St. Davids, PA, USA.

*Corresponding author(s). E-mail(s): rafael.dasilva@eastern.edu;
Contributing authors: jeff.eicher@eastern.edu; gregory.longo@eastern.edu;

Supplementary Material

This document contains the supplementary audits and reliability diagrams referenced in the main manuscript. Tables and figures are numbered S1, S2, ... and are cited in the main text as “Supplementary Table S_n ” and “Supplementary Figure S_n ”.

Evaluation protocol audit

Table [S1](#) summarizes the main operational conventions used in the benchmark. Its purpose is not to introduce a second evaluation framework, but to make explicit the conventions already used in the empirical pipeline: the enrollment-level unit of analysis, the event definition, the censoring treatment for IBS and horizon-specific Brier score, the concordance convention reported in the benchmark tables, the distinction between dynamic and static early-window prediction settings, the enrollment-level identity leakage result, the contextual scope of the split, the operational definition of horizon-specific calibration, the inferential role of bootstrap, and the proportional-hazards scope boundary.

Calibration evidence by horizon

To complement the calibration summary table in the main manuscript, this supplement retains the horizon-wise calibration evidence for Family B. In the canonical final export, no single model held the lowest calibration gap at all three horizons: *Random Survival*

Table S1 Evaluation protocol audit for the unified benchmark.

Component	Status/value	Details
Unit of analysis	defined: enrollment	All final benchmark comparisons are reported at the enrollment level.
Event definition	defined: withdrawn with valid date (unregistration)	Observed withdrawal event with a valid time stamp (unregistration with a recorded date).
Official horizons	defined: 10, 20, 30	Shared benchmark horizons used for Brier, IBS, and calibration reporting.
Primary discrimination metric	defined: TD concordance	The canonical discrimination metric is time-dependent concordance rather than a static concordance proxy.
Censoring treatment	defined: IPCW with KM	Brier score and IBS use inverse-probability-of-censoring weighting with the Kaplan–Meier censoring estimator.
Dynamic versus comparable rule	defined: dynamic weekly vs early-window comparable	Weekly discrete-time arms update predictions over person-period rows, whereas comparable continuous-time arms use early-window enrollment representations; cross-family comparison occurs only after shared enrollment-level horizon reporting.
Leakage prevention rule	defined: no enrollment identity leakage	The split is enforced at the enrollment level with no identity leakage between train and test.
Split scope boundary	defined: shared curricular context	The benchmark generalizes across enrollments under shared curricular context rather than under context-disjoint transportability conditions.
Primary calibration metric	defined: weighted absolute gap by horizon	The main calibration criterion is the weighted absolute calibration gap at each benchmark horizon.
Calibration strengthening	defined: intercept and slope by horizon	Intercept and slope are retained as strengthening diagnostics rather than as the primary benchmark ranking criterion.
Bootstrap inferential role	defined: ranking support, not formal testing	Bootstrap is used to assess how stable the exported ranking appears, not to claim a formal hypothesis-testing result.
PH scope boundary	defined: formal PH audit for Cox anchor only	Formal classical proportional-hazards auditing is available for the comparable Cox anchor, but not in an equivalent form for DeepSurv.

Forest led at horizon 10 (0.0121), *DeepSurv* at horizon 20 (0.0145), and *Cox Comparable* at horizon 30 (0.0106). *XGBoost AFT* remained the only notable exception, with gaps of 0.1242, 0.1775, and 0.2176 at horizons 10, 20, and 30, consistent with its outlier status across benchmark performance and discrimination metrics. Family A calibration gaps at horizon 10 ranged from 0.089 (*Gradient-Boosted Weekly Hazard*) to 0.098 (*Poisson Piecewise-Exponential*); these are reported in the calibration summary table

Table S2 Preprocessing and tuning audit for the four Family B models retained in the focused comparative export. All families used median numeric imputation, constant-missing categorical imputation, one-hot encoding, and standard scaling fit on training rows only. Tuning was deliberately controlled, not exhaustive.

Model	Input level	Validation strategy	Tuning and complexity control
Cox Comparable	enrollment early window	Enrollment-level split with event stratification when possible (20%).	24 candidates. Selection by highest validation C-index. Restricted regularization grid over penalizer and l_1 -ratio. No early stopping.
DeepSurv	enrollment early window	Internal validation fraction on training rows (20%).	24 candidates. Selection by lowest validation loss. Early stopping (patience 10). Complexity controls: architecture grid, dropout, weight decay, best-epoch refit.
Random Survival Forest	enrollment early window	Enrollment-level split with event stratification when possible (20%).	24 candidates. Selection by lowest validation IBS. No early stopping. Complexity controls: forest size, minimum leaf size, maximum depth, feature subsampling fraction.
Neural-MTLR	enrollment early window	Enrollment-level split with deterministic survival-time discretization (10%).	24 candidates. Selection by lowest validation IBS. Early stopping (patience 8). Complexity controls: discretization grid, architecture, dropout, weight decay.

in the main manuscript and confirm that all five Family A models generate numerically coherent horizon-specific risk estimates.

Slope and intercept estimates are retained as strengthening diagnostics only, not as a second parallel ranking criterion. In the consolidated interpretation, RSF, which led on time-dependent concordance and Brier horizons, also showed the best calibration profile among all Family B models (gaps: 0.0121, 0.0162, 0.0138), supporting the main benchmark hierarchy without contradiction.

Accordingly, this supplement uses the tabulated horizon-wise calibration evidence as the primary robustness material for calibration, without a separate figure.

Preprocessing and tuning audit

Table S2 summarizes the preprocessing and tuning conventions for the four Family B models retained in the focused tuning audit. All four used the same upstream preprocessing discipline: median imputation for numeric variables, constant-missing category for categorical variables, one-hot encoding with unknown handling, and standard scaling fitted exclusively on training rows. The remaining ten tuned families (three in Family A and six additional Family B models) applied the same preprocessing contract; their individual candidate records are documented in the pipeline scripts and available in the open repository. A machine-readable summary of all search grids is provided as Online Resource 1 (`ESM_1.pdf`), auto-generated from pipeline run-time metadata by `dropout_bench_v3_G_13_supplementary_hyperparameter_grids.py`.

XGBoost AFT, the consistent outlier across IBS, TD concordance, and calibration, applied the same upstream preprocessing discipline and was tuned over 24 candidates varying AFT loss distribution (logistic, normal), distribution scale (1.0, 1.5, 2.0), learning rate (0.03–0.08), and boosting rounds (200–400); tree depth was held fixed at `max_depth = 2` throughout the grid. A plausible hypothesis for its systematic underperformance is the adverse interaction between the AFT parametrization (which assumes time-invariant multiplicative covariate effects on the acceleration factor) and the broad proportional-hazards departure observed in this dataset (26.3% of covariates

Table S3 Bootstrap uncertainty for the four Family B models (200 enrollment-level resamples, no refit, frozen survival predictions). Lower IBS and Brier scores indicate better probabilistic accuracy (lower mean squared survival error); higher time-dependent concordance indicates better discrimination.

Model	IBS [95% CI]	TD concordance [95% CI]	Brier@10 [95% CI]	Brier@20 [95% CI]	Brier@30 [95% CI]
RSF	0.1118 [0.109, 0.115]	0.6735 [0.663, 0.684]	0.0947 [0.091, 0.099]	0.1329 [0.129, 0.137]	0.1574 [0.153, 0.161]
DeepSurv	0.1145 [0.111, 0.118]	0.6619 [0.651, 0.673]	0.0978 [0.094, 0.102]	0.1343 [0.131, 0.138]	0.1589 [0.155, 0.163]
Cox Comparable	0.1176 [0.114, 0.121]	0.6512 [0.640, 0.663]	0.1012 [0.097, 0.105]	0.1378 [0.133, 0.142]	0.1623 [0.158, 0.166]
Neural-MTLR	0.1131 [0.109, 0.117]	0.6569 [0.647, 0.668]	0.0954 [0.091, 0.100]	0.1340 [0.129, 0.139]	0.1588 [0.154, 0.163]

showed evidence of non-proportionality in the comparable Cox audit, classified as broad departure from proportionality). When relative hazard ratios exhibit strong temporal heterogeneity, even a gradient-boosted ensemble cannot fully compensate for the misspecification introduced by a fixed AFT time-to-event parametrization, and the persistence of this gap under the expanded 24-candidate search reinforces a structural rather than a tuning-related explanation. The failure mode has not been diagnosed beyond this structural hypothesis; formal diagnosis would require analysis of predicted survival curve shapes, AFT residuals, and risk score distributions across event-time strata, which remain outside the current benchmark scope.

Bootstrap uncertainty for the benchmark hierarchy

Because the leading performance differences were modest in absolute magnitude, uncertainty was quantified on the held-out test set through enrollment-level bootstrap resampling (200 resamples, no model refit within iterations, fixed frozen survival predictions). The bootstrap covers the four Family B models retained in the focused comparative audit. Family A families are excluded because their weekly discrete-time hazard predictions are not directly comparable under the same fixed-horizon survival evaluation footprint. Table S3 reports the point estimate and 95% bootstrap interval for IBS, time-dependent concordance, and the three Brier horizons.

The bootstrap results confirm that margins are modest in absolute terms: IBS confidence intervals for the top four models overlap substantially, so the IBS hierarchy should be treated as a directional signal rather than a stable dominance claim. On time-dependent concordance, RSF maintained rank 1 in 90% of resamples, a largely stable leadership pattern. These bootstrap artifacts do not constitute a formal hypothesis test and should not be narrated as unqualified strict superiority.

Table S4 Bootstrap inferential scope for the main benchmark leadership claims (Family B, 200 enrollment-level resamples, no refit).

Metric	Direction	Leading model	Claim status	Conservative takeaway
Integrated Brier Score	lower is better	RSF	directional signal only	IBS intervals overlap substantially; RSF leads narrowly but the ranking is not stable enough for a strong dominance claim.
Time-dependent concordance	higher is better	RSF	largely stable bootstrap leadership	RSF held rank 1 in 90% of resamples; largely stable but not absolute leadership.

Bootstrap inferential scope

Table S4 clarifies what the bootstrap artifacts do and do not support in the current paper. On time-dependent concordance, *Random Survival Forest* held rank 1 in 90% of resamples, a largely stable leadership pattern, with *DeepSurv* as the main runner-up (10% rank-1 share). On IBS, RSF led with a rank-1 share of 64%; the IBS hierarchy is directional but the intervals overlap, and the ranking should not be treated as strictly dominant (RSF rank-1 share: 64%; Neural-MTLR: 27%).

The inferential reading should remain disciplined. These bootstrap artifacts support cautious directional language within the Family B audit subset, but they do not constitute a formal null-hypothesis test, do not establish a universally fixed ranking across samples, and should not be narrated as unqualified strict superiority. Because the resampling procedure uses frozen survival predictions without refitting the model within each bootstrap iteration, the reported intervals capture enrollment-sampling variance only, not the estimation variance of the fitted model parameters. The intervals should therefore be interpreted as conservative lower-bound estimates of total uncertainty, not as full inferential confidence intervals for the generalization performance of the models.

Split and contextual overlap audit

Because the benchmark uses an enrollment-level train–test split, we audited not only enrollment identity leakage but also the degree of shared curricular context across the partition. This subsection clarifies the scope of the benchmark evaluation: it should be read as generalization across enrollments under shared curricular context, not as a stricter transportability setting across unseen module or presentation contexts.

Table S5 summarizes the compact split/context audit exported from the notebook. The benchmark used 32,593 enrollments in total, with 22,815 in train and 9,778 in test. Event rates were closely matched across the two partitions, and no enrollment identity leakage was detected. At the same time, curricular context was fully shared across train and test: all 7 modules, all 4 presentations, and all 22 module-presentation combinations appeared in both splits. The benchmark should therefore be interpreted as enrollment-level evaluation under shared curricular context, not as a context-disjoint transportability benchmark.

Table S5 Split and contextual overlap audit for the main benchmark.

Split unit	Stratification	Total	Train	Test	Train event rate	Test event rate
enrollment	event status + coarse event-time bucket	32,593	22,815	9,778	0.2266	0.2266
Identity leakage	Shared modules	Shared presentations	Shared module-presentations			
no	7/7	4/4	22/22			

Table S6 Discrete-time diagnostic summary for Family A models.

Model	Representation	Candidates	Rows per enrollment	Row event rate	IBS rank	TD rank
Neural Discrete-Time	dynamic weekly person-period	24	23.77	0.0095	4	1
Linear Discrete-Time	dynamic weekly person-period	24	23.77	0.0095	1	2

Discrete-time diagnostic bridge

The purpose of this subsection is to clarify why the dynamic discrete-time branch underperformed in the final enrollment-level benchmark without collapsing that result into a simplistic claim of absent weekly signal. The frozen diagnostic export shows that both discrete-time anchor models retained non-trivial row-level predictive structure, but that this local signal did not translate as effectively as the early-window comparable representation when the evaluation target shifted to enrollment-level survival ranking, error, and calibration.

Table S6 summarizes the key diagnostic contrast. Both dynamic models were explicitly tuned, used approximately 23.8 weekly rows per enrollment on the test set, and operated under low row event rates near 0.0095. The neural weekly model improved row-level ROC-AUC and log loss relative to the linear weekly model, yet the final time-dependent concordance ranking remained weaker. This pattern is consistent with a harder aggregation problem in which sparse low-base-rate weekly hazards must be converted into stable enrollment-level survival outputs.

Table S7 then records a conservative interpretation of the main diagnostic hypotheses considered in the final evidence audit. Weak tuning alone is not supported as the primary explanation, because both discrete-time anchor models were explicitly tuned under bounded searches. Absent weekly signal is also not supported, because row-level

Table S7 Hypothesis audit for the discrete-time underperformance reading.

Hypothesis	Status	Conservative takeaway
Weak tuning as primary driver	tested, not supported as primary explanation	Family A models were explicitly tuned; weak tuning alone is not the best-supported explanation for the final benchmark gap.
Absent weekly signal	tested, not supported	Weekly rows contain predictive signal; the underperformance emerges downstream when those weekly hazards must support enrollment-level survival ranking and calibration.
Sparse low-base-rate aggregation burden	tested, partially supported	The dynamic representation faces a sparse low-base-rate aggregation problem that is consistent with harder survival reconstruction and calibration, but this remains evidence-consistent rather than causal proof.
Neural weekly instability or overdispersion	tested, partially supported	Family A neural models appear more expressive locally, but that extra flexibility does not translate into a stronger final ranking profile under the benchmark contract.
Representational inefficiency under the current contract	residual interpretation best supported	Family A captures signal, but less efficiently than the early-window comparable design under the present benchmark contract.

discrimination and loss remained informative. The best-supported residual interpretation is therefore representational inefficiency under the present benchmark contract: Family A captures signal, but less efficiently than the early-window comparable design when the benchmark outcome is enrollment-level survival prediction.

Proportional-hazards audit for the comparable Cox model

Because the static early-window family is methodologically anchored by a Cox-type model, the comparable Cox benchmark was subjected to a formal proportional-hazards audit. The purpose of this supplementary section is not to reopen model selection, but to clarify the diagnostic status of the classical Cox assumption within the static early-window branch.

Table S8 summarizes the global result of the audit. Of the 38 tested covariates, 10 showed evidence of possible non-proportionality at $\alpha = 0.05$, corresponding to 26.3% of the tested covariates. The notebook assigns a global classification of broad departure from proportionality: the comparable Cox benchmark shows broad evidence of proportional-hazards departure and should be interpreted as an approximate comparable benchmark rather than a fully assumption-clean Cox specification.

The covariates with the strongest evidence of possible non-proportionality in the current evidence freeze were the following:

- (i) `region` (level: Scotland)
- (ii) `studied_credits`
- (iii) `region` (level: Wales)
- (iv) `age_band` (level: 0–35)

Table S8 Global proportional-hazards audit summary for the comparable Cox model.

Model	Covariates tested	Flagged	Global interpretation
Cox Comparable	38	10 (26.3%)	Broad departure: approximate comparable benchmark, not assumption-clean Cox specification

(v) **region** (level: Ireland)

The enumerated covariates above were identified as the strongest signals of non-proportionality, confirming the broad departure from proportionality classification.

This audit establishes that the comparable Cox benchmark is an approximate comparable specification (broad departure from proportionality classification), not a fully assumption-clean Cox model. The comparable Cox still provides the classical methodological anchor for the static early-window family, but all results from this branch should be narrated with the broad-departure approximation explicitly acknowledged. DeepSurv shares the Cox-type ranking structure, but it was not subjected to an identical classical proportional-hazards diagnostic.

PH scope boundary

Table S9 makes the scope boundary explicit for all nine Family B models. Formal classical proportional-hazards auditing (via Schoenfeld residuals and log-log diagnostic plots) requires a linear predictor with well-defined martingale residuals; this condition is satisfied only by the comparable Cox anchor, which received a formal audit (broad departure from proportionality classification). For families that use an accelerated failure time or discrete-time formulation (Weibull AFT, XGBoost AFT, Neural-MTLR, DeepHit), the proportional-hazards assumption does not apply and the audit is not applicable. For non-parametric or tree-based ensembles that make no distributional assumption (RSF), the audit is likewise not applicable. For the Gradient-Boosted Cox family, which uses a partial-likelihood objective that implies a proportional-hazards structure, a Schoenfeld audit would be applicable in principle but was not executed within this benchmark freeze. For Royston-Parmar, which relaxes PH via splines, dedicated goodness-of-fit tools exist but were not applied. DeepSurv shares the Cox-type ranking structure but its nonlinear risk surface makes classical Schoenfeld diagnostics inapplicable; this is narrated as a boundary on diagnostic coverage, not as an automatic invalidation of DeepSurv’s benchmark performance. Specifically, DeepSurv extends the Cox partial likelihood with a neural predictor $h_0(t) \exp(f_\theta(\mathbf{x}))$, where f_θ is a multi-layer network. Schoenfeld residuals require per-subject score contributions derived from the partial-likelihood gradient, which are not available in closed form when the linear predictor is replaced by a neural function; classical proportional-hazards diagnostics are therefore methodologically inapplicable to DeepSurv, not merely omitted. DeepSurv is evaluated through the same predictive metrics (IBS, time-dependent concordance, horizon-specific Brier scores), calibration gaps, and enrollment-level

Table S9 Proportional-hazards scope boundary across all nine Family B models. Formal classical PH diagnostics (Schoenfeld residuals) require a linear partial-likelihood estimator; this condition is met only by the Cox anchor, which was formally audited.

Model	Assumption class	Formal PH audit	Paper claim boundary
Cox Comparable	PH (linear)	yes	broad departure from proportionality; all comparable Cox results should be narrated as from an approximate benchmark specification, not from a fully assumption-clean Cox model.
DeepSurv	Cox-type neural (nonlinear)	no	Classical Schoenfeld diagnostics require per-subject partial-likelihood score contributions, which are not computable in closed form when the linear predictor is replaced by a neural network. PH diagnostics are therefore methodologically inapplicable (not merely omitted). Evaluated via IBS, concordance, calibration gaps, and bootstrap rank stability.
Random Survival Forest	tree-survival (non-parametric)	N/A	No PH assumption is made; PH audit is not applicable.
Gradient-Boosted Cox	PH (partial likelihood)	in principle; not run	Partial-likelihood objective implies PH; a Schoenfeld audit would be applicable but was not executed in this benchmark.
Weibull AFT	AFT (parametric)	N/A	AFT model; PH assumption does not apply; accelerated time parametrization is used instead.
Royston-Parmar	PH spline (relaxed PH)	no	Spline-based relaxation of PH; dedicated diagnostics exist but were not applied. Narrate as relaxed-PH, not audited-PH.
XGBoost AFT	AFT (tree-based)	N/A	AFT formulation; PH assumption does not apply.
Neural-MTLR	discrete-time	N/A	Discrete-time model; no PH assumption is made or audited.
DeepHit	discrete-time	N/A	Discrete-time model; no PH assumption is made or audited.

bootstrap rank stability applied to all other families; these constitute the applicable scope of performance evidence for a Cox-type neural model.

Early-window sensitivity

Table S10 reports IBS and time-dependent concordance for each Family B tuned family at the canonical window ($w = 4$ weeks), validated against a sensitivity grid ($w \in \{2, 4, 6, 8, 10\}$ weeks). In the pipeline sensitivity grid, performance generally improved with longer observation windows; $w = 4$ was selected as the canonical choice and held fixed for all primary benchmark comparisons.

Table S10 Early-window sensitivity for Family B. IBS (lower is better) and time-dependent concordance (higher is better) for each model, reported at the canonical window ($w = 4$ weeks). Performance generally improves with longer observation windows; full sensitivity results are available in the pipeline outputs.

Model	IBS	TD Concordance
Random Survival Forest	0.1118	0.6735
Neural-MTLR	0.1131	0.6569
DeepSurv	0.1145	0.6619
Gradient-Boosted Cox	0.1148	0.6621
DeepHit	0.1151	0.5835
Cox Comparable	0.1176	0.6512
Weibull AFT	0.1180	0.6131
Royston-Parmar	0.1190	0.6430
XGBoost AFT	0.1495	0.5772

Reliability diagrams

Figures S1, S2, and S3 display the reliability diagrams for Family B in three smaller panels to preserve page layout. Figure S4 retains the Family A reliability view. In every panel, rows correspond to horizons 10, 20, and 30 weeks; each subplot compares mean predicted risk (x-axis) against observed event rate (y-axis), with point size proportional to bin size and the dashed diagonal representing perfect calibration. These figures complement the tabulated calibration evidence in the main manuscript and the calibration-evidence section above.

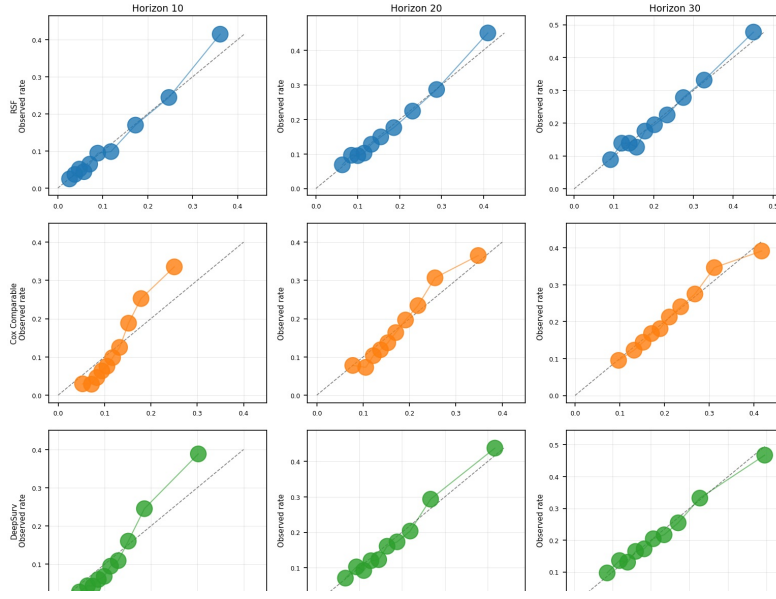


Figure S1 Reliability diagrams for Family B (panel 1 of 3). This panel covers the first subset of static early-window families; most points remain close to the diagonal, consistent with the strong calibration profile of the leading static early-window models.

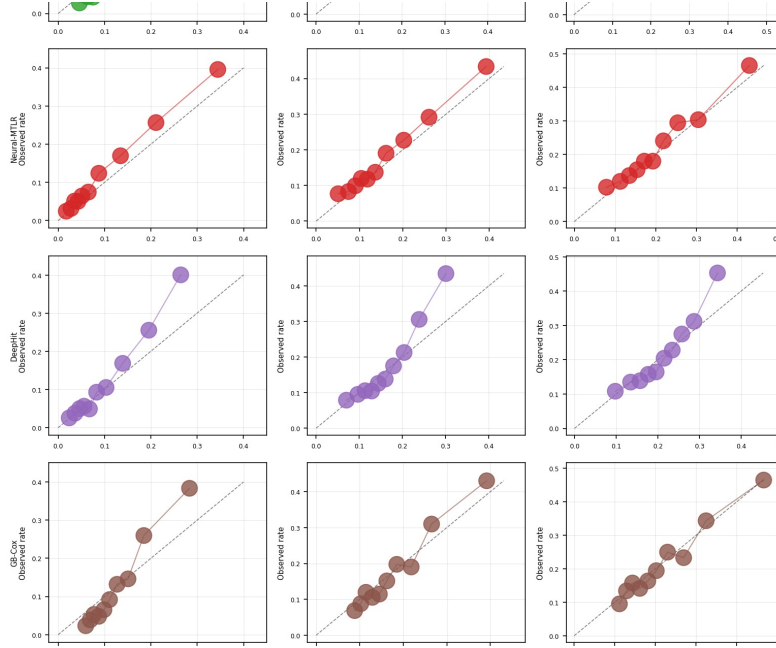


Figure S2 Reliability diagrams for Family B (panel 2 of 3). The middle subset preserves the same three-horizon calibration comparison while avoiding the page overflow of the original full-width grid.

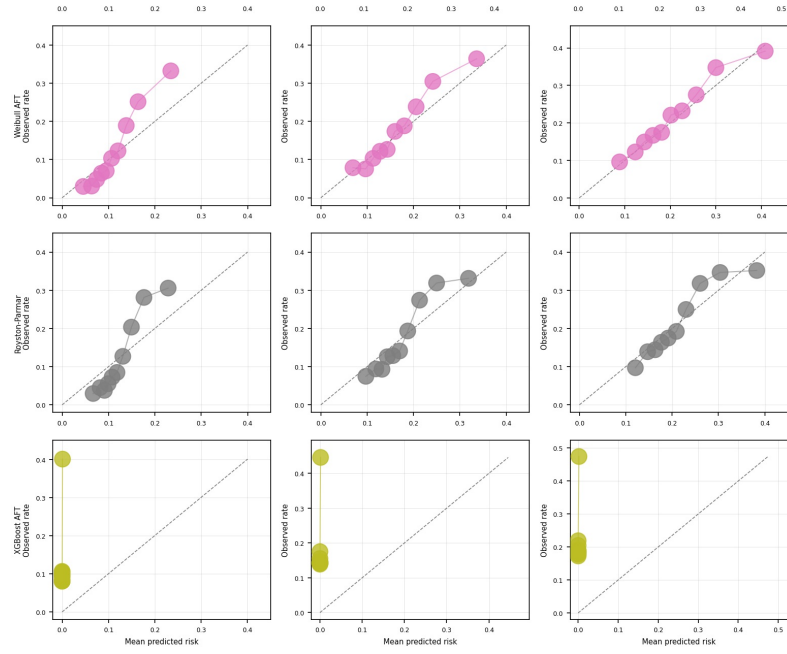


Figure S3 Reliability diagrams for Family B (panel 3 of 3). This final panel includes the remaining static early-window families and retains the visible calibration departure of XGBoost AFT relative to the better-calibrated alternatives.

Reliability Diagrams — Dynamic Arm

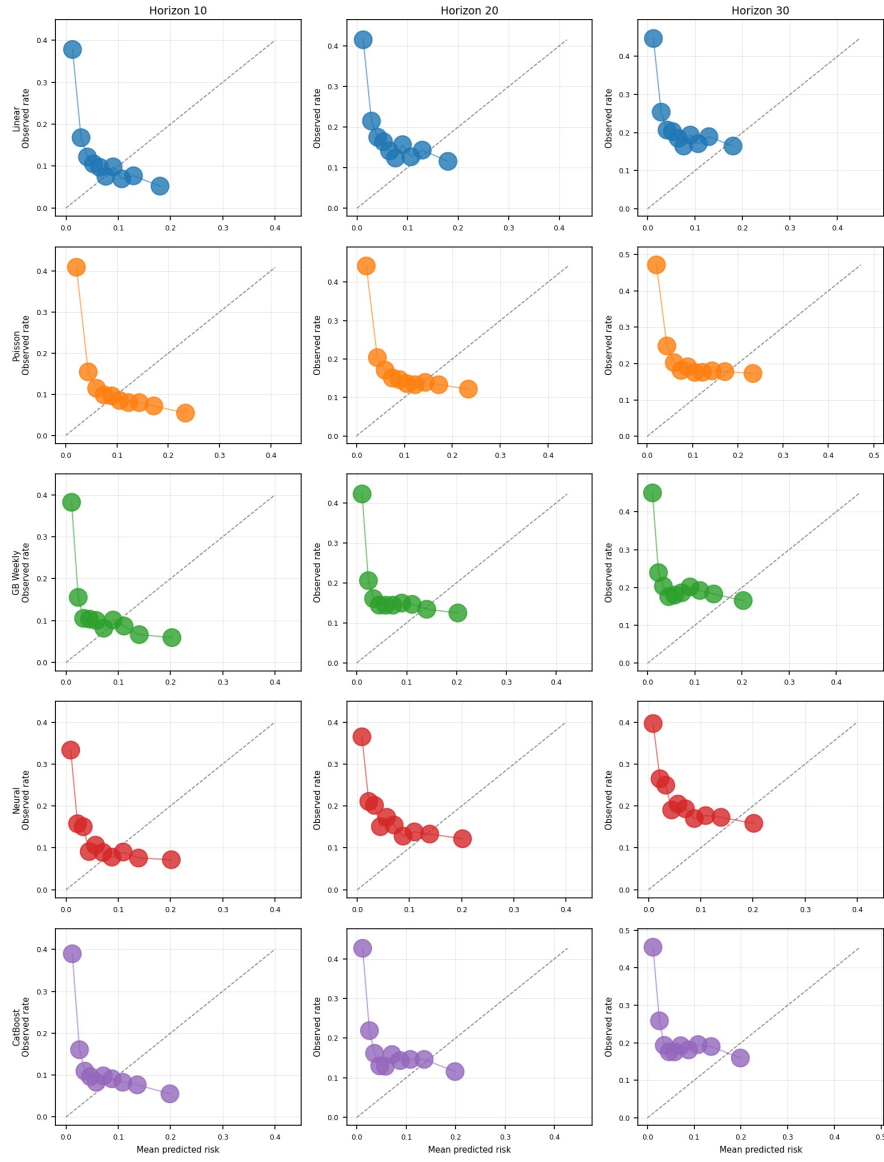


Figure S4 Reliability diagrams for Family A. Rows correspond to horizons 10, 20, and 30 weeks; columns correspond to models. Family A calibration gaps remain wider than those of the strongest Family B models, but the panels still show coherent horizon-wise probability structure.