File S1: Supplementary methods and detailed results
How Breast Cancer Survival Predictions Change Over Time

Ayan Kokhan, Tomiris Zhaksylyk, and Nursultan Kuldeyev

Reading the supplementary materials

Tables S1–S10 are supplied as separate CSV files. The table_part column identifies the sections described in the table index.

BCSM denotes breast-cancer-specific mortality; OCM denotes other-cause mortality. MCB is the signed mean predicted minus observed cumulative incidence, multiplied by 100 and expressed in percentage points. Positive MCB indicates average overprediction. CIF, Brier score, ICI and maximum calibration error are reported on the risk-fraction scale; AUC is unitless. Logistic calibration parameters use clipped risk logits and IPCW weights. The slope-fixed intercept fixes the logit slope at one. A small MCB does not establish individual-level calibration.

Unless otherwise specified, evaluation uses the common 20,000-patient samples. Model development and recent refitting use full eligible cohorts. Development, recalibration and evaluation windows have different roles. Bootstrap intervals are percentile intervals; an evaluation-only bootstrap conditions on fitted updates. The coupled bootstrap adds update-estimation uncertainty for the latest era while holding the anchor and model settings fixed. No table establishes clinical effectiveness or independent-population validation.

All contrasts in Table S2 (parts A and B) are isotonic recalibration minus recent refitting. Negative signed MCB contrasts do not automatically mean lower absolute calibration error. Higher AUC is favorable; lower Brier score is favorable. Individual confidence intervals are pointwise and are not multiplicity adjusted.

Supplementary methods and detailed results

Evaluation-sample sensitivity

We checked evaluation-sample sensitivity against the full cohorts. Maximum absolute differences were 0.322 pp for MCB, 0.00722 for IPCW AUC, 0.00230 for IPCW Brier score, 0.00299 for ICI, and 0.0584 for $E_{\max}$ (Supplementary Table S3 (part A)).

Unknown-cause sensitivity protocol

Excluding deaths of unknown cause may introduce selection bias if missingness depends on patient characteristics or outcomes. We assessed sensitivity by assigning the 5,538 otherwise-eligible deaths to OCM, refitting the encoder and anchor models, and repeating sampling and evaluation with the original seeds (Supplementary Tables S4 (part C) and S5 (part B)). The changed eligible population meant that the sampled patients were not identical to those in the primary analysis. Maximum absolute differences were 0.45 pp for MCB and 0.020 for IPCW AUC across families, eras, and causes. This comparison tests the combined inclusion and cause-assignment policy; it neither identifies the true causes of these deaths nor rules out selection bias.

Cumulative-incidence implementation

Cumulative incidence combines the estimated hazards for both causes with event-free survival. For the Cox-based models, $H_k(u \mid X_i) = H_{0,k}(u)\exp(X_i\beta_k)$ is the patient's cumulative hazard for cause k; $S(u \mid X_i)$ is event-free survival for both causes (manuscript Appendix A, Equation A1). For RSF, we obtained cause-specific cumulative hazards from the ensemble's terminal nodes. All families used the same numerical integration on a shared grid of 256 time points and the same evaluation code. BCSM risk was $F_1(t)$, rather than $1 - S_1(t)$. Primary model settings are listed in manuscript Appendix A.1; alternative RSF and DeepSurv configurations are reported in Supplementary Table S4 (part A).

Recalibration-method implementation

Sensitivity analyses varied the nominal recalibration window (1–5 years, truncated at 2004 in the earliest era), calibrator (isotonic, logistic, or weighted LOESS), survival-model hyperparameters, missing-data handling, and evaluation-sample size and seed. Each calibrator used preceding-window data and was evaluated in the target era. Weighted LOESS used local-linear fits, a tricube kernel, span 0.75, and up to 401 prediction-quantile grid points; interpolated predictions were clipped to [0,1]. The method comparison used identical preceding two-year samples and IPCW weights for all three calibrators (Supplementary Table S6 (part A)). Fitting to target-era outcomes would assess contemporaneous calibration rather than an update available before that era.

Decision-curve weighting

Exploratory decision-curve analysis used hypothetical cause-specific 60-month risk thresholds of 2.5%–30%, with treat-all and treat-none references (Vickers and Elkin, 2006; Sadatsafavi et al., 2021). This broad range was chosen to examine lower and higher predicted risks, without a treatment guideline or validated action cutoff. At threshold p, the hypothetical rule intervenes above predicted risk p and assigns relative false-positive weight $p/(1-p)$. Competing deaths were non-cases for the cause evaluated, and IPCW accounted for right censoring. Because no intervention or treatment effect was specified, the resulting net benefit does not demonstrate clinical benefit.

Censoring weights and assumptions

We estimated the censoring-survival function $\widehat{G}$ separately in each calibration or evaluation sample using a marginal reverse Kaplan–Meier estimator. Both causes of death were removals from the censoring risk set, rather than censoring events. Deaths were removed before censoring at tied times. Observed deaths by the prediction horizon τ received weight $1/\widehat{G}(T_i^-)$, including competing deaths. Patients followed beyond τ received weight $1/\widehat{G}(\tau)$. Records right-censored at or before τ had unknown horizon status and zero weight in the IPCW outcome calculations. Thus, calibration weights used preceding-window patients, whereas evaluation weights used target-era patients; evaluation outcomes did not fit the calibrator.

Each denominator was bounded below by 10^{-8} as a numerical safeguard. We did not apply additional percentile trimming or fit a covariate-dependent censoring model. Censoring weights were re-estimated in bootstrap samples. Marginal weighting assumes censoring is independent of event time and cause within the sample, with positive censoring-survival probabilities at the relevant times. Dependence of censoring on patient characteristics or outcomes may therefore remain a source of bias.

Calibration-parameter estimation

Both CIFs in the equation are fractions; multiplying their difference by 100 expresses MCB in percentage points. We estimated calibration intercept and slope by IPCW-weighted logistic regression of the cause-specific 60-month outcome on the logit of predicted CIF, clipped to

$[10^{-6}, 1-10^{-6}]$. A separate fit with slope fixed at one estimated calibration-in-the-large. ICI and $E_{\max}$ were the mean and maximum deviations of a weighted LOESS calibration curve, respectively, on the risk-fraction scale.

Evaluation bootstrap

Evaluation uncertainty. Primary MCB and IPCW Brier-score intervals used 500 evaluation-bootstrap replicates per era. A further 200 replicates estimated intervals for MCB, IPCW AUC, IPCW Brier score, ICI, and $E_{\max}$ across all three primary strategies. Each replicate resampled evaluation patients with replacement, re-estimated censoring weights, and recalculated metric smoothers and quantile-bin incidence. Survival models and post-hoc calibrators remained fixed. Calibration-curve intervals were pointwise and binwise, not simultaneous. Paired isotonic-minus-refit contrasts used 500 shared evaluation-patient resamples, with original-sample differences as point estimates.

Coupled update-bootstrap algorithm

Update-estimation uncertainty. A complementary 200-replicate bootstrap for 2016–2017 re-estimated updates, conditional on the original anchor model, encoder, and selected calibration and evaluation samples. We split the 2012–2015 refitting cohort into the fixed 20,000-patient 2014–2015 calibration sample and its complement, then resampled each stratum at its original size. The same calibration-stratum draw fitted the isotonic mappings and entered refitting, preserving dependence through shared patients. Evaluation patients were resampled independently of the update cohort, with a common draw for both strategies. Both cause-specific refit models and calibrators were re-estimated; predictions, censoring weights, and metrics were recalculated.

Bootstrap intervals and random seeds

Percentile intervals used the 2.5th and 97.5th percentiles, retaining original-sample point estimates separately from bootstrap means. Hyperparameters and fitting seeds remained fixed. Replicate r used NumPy seed $22200814 + r$, for $r = 0, \ldots, 199$; calibration- and evaluation-sampling seeds were 20260918 and 20261018. This bootstrap does not include initial anchor-development uncertainty. Recalibration and refitting also differ in update data: two years and 20,000 patients versus four years and all eligible patients. The comparison therefore concerns these specified protocols, not an isolated updating mechanism with identical training data. Supplementary Table S1 (part A) and Figure S1 report metric estimates, calibration parameters, and evaluation-bootstrap intervals.

Subgroup definitions

We described calibration by age group and within-family, within-era predicted-risk decile in the full evaluation cohorts. Age groups were 20–39, 40–49, 50–59, 60–69, 70–79, and 80+ years; observed CIF was estimated separately in each subgroup (Supplementary Table S8 (parts A and B)). We recorded the first era to cross each illustrative MCB cutoff without prespecifying a clinical action threshold. Further analyses varied RSF/DeepSurv hyperparameters, missing-data handling, evaluation-sample size and seed, and predictor set (age only).

Full-cohort BCSM drift checks

Observed BCSM declined while fixed models continued to predict risks closer to those of earlier cohorts. Full-cohort MCB slopes were 0.14–0.17 pp per year. Across model families, MCB increased from 0.52–1.22 pp in the first era to 1.55–2.41 pp in the latest era (Table 3), with the smallest bias for RSF and largest for elastic-net Cox. Latest-era observed-to-expected ratios were 0.81–0.87, and full-cohort versus sampled MCB differed by no more than 0.035 pp.

Other-cause metrics and integration sensitivity

OCM calibration errors remained below 0.93 pp for all families and eras (Figure 5a). For RSF, MCB ranged from 0.30 to 0.90 pp, ICI from 0.004 to 0.009, and $E_{\max}$ from 0.007 to 0.081. OCM discrimination was lower than for BCSM, with IPCW AUC values of 0.78–0.82 (Figure 5b). The archived numerical-integration check reported a maximum difference of approximately 0.002 between predictions calculated on alternative time grids.

Baseline-hazard updating

Updating only the baseline hazard reduced late-era MCB to 0.54 pp for ridge Cox, 0.73 pp for elastic-net Cox, and 0.79 pp for DeepSurv. This removed 65%–74% of the original bias, while IPCW AUC differed from frozen predictions by less than 0.0002 across eras. This update is not defined for RSF.

Refitting point estimates

Recent refitting left latest-era residual BCSM MCB of 0.57–0.97 pp, compared with at most 0.36 pp after isotonic recalibration. Refitting increased IPCW AUC by 0.0017–0.0064 over the frozen models. These comparisons describe average signed bias and discrimination; lower MCB does not establish better calibration at every risk level or lower overall prediction error.

AUC rounding and isotonic ties

Some frozen and isotonic-recalibrated AUC values coincide after rounding to three decimals. A strictly increasing mapping preserves ranking, but isotonic plateaus create ties that can change AUC through tie averaging. Latest-era RSF BCSM AUC changed from 0.883583 to 0.883110, reported as 0.884 and 0.883, respectively. The values are close, not identical. Variation in risk distributions and ties can change the size of this difference across eras without a large change in discrimination. Supplementary Table S1 (part A) reports the full-precision comparisons.

Paired strategy contrasts

Latest-era original-sample isotonic-minus-refit BCSM MCB differences were −0.404 pp (95% CI −0.448 to −0.363) for ridge Cox, −0.798 (−0.842 to −0.748) for elastic-net Cox, −0.212 (−0.254 to −0.171) for RSF, and −0.527 (−0.577 to −0.481) for DeepSurv. These paired evaluation-bootstrap intervals condition on fitted updates and omit uncertainty in their estimation. The contrasts concern signed MCB, not absolute calibration error. Supplementary Table S2 (part A) reports paired MCB, AUC, and Brier-score estimates for all 40 family-era-cause contrasts, with pointwise intervals and no multiplicity adjustment.

Coupled-bootstrap results

Re-estimating updates in the coupled bootstrap widened the 95% intervals for BCSM MCB differences: −0.725 to −0.113 pp for ridge Cox, −1.140 to −0.517 for elastic-net Cox, −0.491 to 0.085 for RSF, and −0.881 to −0.129 for DeepSurv. The RSF MCB interval included zero, as did its AUC-difference interval (−0.003125 to 0.000747). Brier-score difference intervals favored refitting for RSF (0.000281 to 0.001170) and DeepSurv (0.000289 to 0.001267). All four OCM MCB-difference intervals included zero. Supplementary Table S2 (part B) separates original-sample contrasts, bootstrap means, and intervals for all eight latest-era family-cause combinations. These intervals remain conditional on the anchor model, encoder, selected samples, hyperparameters, and fitting seeds.

Additional performance metrics

Across families and eras, frozen BCSM models had IPCW Brier scores of 0.0641–0.0734, ICI of 0.0041–0.0235, and $E_{\max}$ of 0.0101–0.1630. Corresponding isotonic-recalibration ranges were 0.0638–0.0724, 0.0020–0.0053, and 0.0079–0.0559. Because these ranges span different cohorts, they are not paired improvement estimates. In the latest era, refitting yielded lower Brier score than isotonic recalibration for RSF (0.06295 versus 0.06375) and DeepSurv (0.06313 versus 0.06389). RSF also had lower $E_{\max}$ after refitting (0.01369 versus 0.04594). Supplementary Table S1 (part A) reports all five metrics, calibration parameters, and conditional intervals for each family, era, cause, and primary strategy.

Joint probability coherence

Separate cause-wise isotonic mappings produced risk sums above one in 0–94 patients per 20,000-patient evaluation sample (0–0.47%); the largest sum was 1.296. Frozen and recently refitted predictions did not exceed one. Proportionally rescaling both updated risks where their sum exceeded one changed MCB by at most 0.011 pp, IPCW AUC by 0.00009, Brier score by 0.00007, and ICI by 0.00019. The maximum change in $E_{\max}$ was 0.097 on the risk-fraction scale, indicating sensitivity of tail calibration to this correction (Supplementary Table S10).

Short-horizon recalibration and lagged associations

Across four successive era pairs per family, the descriptive Pearson correlation between 12-month BCSM MCB and next-era 60-month MCB ranged from −0.32 to 0.07 (Supplementary Table S7 (part D)). Archived recalibration at the same 12- or 24-month horizon left MCB between −0.20 and 0.38 pp. Transferring a 12-month calibrator to 60-month predictions produced underprediction of 2.62–3.88 pp (Supplementary Table S7 (part F)). These estimates do not validate prospective drift prediction or transfer between horizons.

Contemporaneous associations and outcome availability

Contemporaneous 12-, 24-, and 36-month MCB was associated descriptively with 60-month MCB across family-era cells (Spearman $\rho = 0.57$, 0.94, and 0.97 for BCSM; 0.63, 0.68, and 0.69 for OCM). In 2018–2019, outcome status was known for 98.8%, 97.9%, and 97.1% of patients by 12, 24, and 36 months, respectively, and 50.9% by 60 months (Supplementary Table S7 (part E)). Earlier availability may permit faster feedback, but these contemporaneous associations do not establish a surrogate for mature 60-month calibration. Supplementary Table S4 (part C) summarizes the horizon and 2018–2019 analyses.

Table index

Table S1: Performance and calibration over calendar time

Part A. Performance and calibration across models, eras, outcomes, and updating strategies All families, eras, causes and strategies. Metric intervals use 200 evaluation-patient bootstrap replicates, conditional on fitted models and updates. Calibration parameters are point estimates. This part contains 120 rows.

Part B. Observed mortality incidence by diagnosis year and era Full-cohort Aalen–Johansen estimates; annual cohorts are distinct patient groups. This part contains 20 rows.

Part C. Frozen-model calibration bias by diagnosis year Signed MCB for distinct annual diagnosis cohorts and each model family and cause. This part contains 40 rows.

Part D. Binned calibration estimates and pointwise confidence intervals Original-sample quantile-bin mean predictions and observed CIF, with 200-replicate pointwise binwise evaluation-bootstrap intervals. Not simultaneous confidence bands. This part contains 1,200 rows.

Table S2: Paired comparisons of recalibration and refitting

Part A. Paired differences between recalibration and refitting across diagnosis eras All 40 family/era/cause contrasts; original-sample differences and 500-replicate paired evaluation-bootstrap intervals. Conditional on fitted updates; pointwise intervals without multiplicity adjustment. This part contains 40 rows.

Part B. Latest-era recalibration versus refitting with update-estimation uncertainty Latest era only. Original-sample differences, draw means reported separately, and 200-replicate percentile intervals. Shared patient multiplicities preserve overlap between recalibration and refitting windows. Anchor model, encoder, hyperparameters and model seeds remain fixed. Replicate r uses NumPy seed 22200814+r (r=0,...,199). Calibration/evaluation sampling seeds are 20260918/20261018; RSF/DeepSurv fitting seed is 20260814. This part contains 8 rows.

Table S3: Sensitivity to evaluation sampling

Part A. Performance in full evaluation cohorts and shared patient samples Compact comparison of full evaluation cohorts and the common 20,000-patient samples. Model-training sample sizes are unchanged. This part contains 40 rows.

Part B. Metric-by-metric comparison of full-cohort and sampled evaluation All five major metrics for 40 frozen-model family/era/cause cells, with signed full-minus-sample differences. This part contains 40 rows.

Part C. Sensitivity to evaluation-sample size and random seed 5,000, 10,000, 20,000 and 40,000 patients with three seed offsets; models remain fixed. This part contains 480 rows.

Table S4: Model settings and sensitivity-analysis overview

Part A. RSF and DeepSurv configurations and calibration-drift summaries Six RSF and four DeepSurv configurations, including one reference configuration per family. This part contains 10 rows.

Part B. Calibration drift by model configuration and diagnosis era Family, configuration, era and cause estimates for all tested settings. This part contains 100 rows.

Part C. Overview of sensitivity analyses and their findings Overview of assumptions, numerical findings, and limits; detailed estimates are provided in the other table parts. This part contains 13 rows.

Table S5: Sensitivity to missing data, cause of death, and predictor availability

Part A. Performance under alternative missing-data strategies Primary, complete-case and no-indicator analyses. Complete-case comparisons change both fitting and evaluation populations and do not isolate imputation bias. This part contains 120 rows.

Part B. Sensitivity to assigning deaths of unknown cause to other-cause mortality Hypothetically assigns 5,538 otherwise-eligible unknown-cause deaths to OCM, refits anchor models and redraws samples with the same seeds. This changes eligible populations and does not resolve missing-cause selection bias. This part contains 40 rows.

Part C. Temporal performance of age-only models Age is available at diagnosis. These models are not a complete diagnosis-time model and do not correct the primary predictor-timing limitation. This part contains 40 rows.

Table S6: Recalibration methods, window lengths, and monitoring cutoffs

Part A. Comparison of isotonic, logistic, and LOESS recalibration Full-precision isotonic, logistic and LOESS estimates using identical preceding-window samples and IPCW weights. This part contains 120 rows.

Part B. Calibration bias by recalibration-window length Nominal one- to five-year windows, with actual year boundaries. The earliest nominal five-year window is truncated at 2004 and spans four years. This part contains 200 rows.

Part C. Latest-era calibration bias by recalibration window and method, with confidence intervals 16 BCSM cells: isotonic one-, two- and three-year windows and logistic two-year recalibration for all four families. 500 shared evaluation-patient bootstrap draws, seed 20260814, conditional on fitted models/calibrators. This part contains 16 rows.

Part D. First crossing of illustrative calibration-bias cutoffs First frozen-model BCSM absolute-MCB crossing for 1.5, 2.0 and 2.5 percentage-point cutoffs; no optimized clinical trigger. This part contains 12 rows.

Table S7: Prediction horizons and outcome availability

Part A. Frozen-model performance by prediction horizon Frozen-model estimates at 12, 24, 36 and 60 months. This part contains 160 rows.

Part B. Associations between shorter-horizon and five-year calibration bias Full-precision rank associations across 20 family/era cells per cause. Contemporaneous associations do not validate a prospective lead indicator. This part contains 6 rows.

Part C. Short-horizon performance and outcome availability in 2018–2019 Prediction summaries from the original analysis at 12, 24 and 36 months, including outcome availability. Prediction models were not refitted. Five-year availability is reported separately in S7 (part E). This part contains 24 rows.

Part D. Association of 12-month calibration bias with next-era five-year bias Family-specific Pearson correlation of 12-month BCSM MCB with next-era 60-month MCB; only four era pairs per family. This is a descriptive association, not a validated predictive assessment. This part contains 4 rows.

Part E. Availability of outcome status at each horizon in 2018–2019 Complete pooled cohort of 125,925 patients. Known status at each horizon includes a recorded death before that horizon or follow-up reaching it; a competing-cause death also gives known status. Counts and percentages are calculated from endpoint data. These are not complete-case performance estimates. This part contains 4 rows.

Part F. Short-horizon recalibration and transfer to five-year predictions Original-analysis point estimates at 12 and 24 months and transfer of a 12-month calibrator to 60-month predictions. The 40 frozen-model estimates also appear in Table S7 (part A). Recalibration estimates were retained from the original analysis without refitting. This part contains 100 rows.

Table S8: Calibration by age and predicted risk

Part A. Full-cohort calibration bias by age group Descriptive age-group mean calibration bias for all families and eras. Predictions use the frozen anchor models; subgroup observed CIF uses Aalen–Johansen. This part contains 120 rows.

Part B. Full-cohort calibration bias by predicted-risk decile Descriptive within-family, within-era risk-decile mean calibration bias in the first and latest eras. These are distinct cohorts; decile membership changes between eras. This part contains 80 rows.

Table S9: Exploratory decision-curve analysis

Part A. Net benefit across hypothetical risk thresholds Threshold-specific hypothetical net benefit. No specified intervention or established treatment benefit; competing deaths are non-cases and right censoring is handled by IPCW. This part contains 1,440 rows.

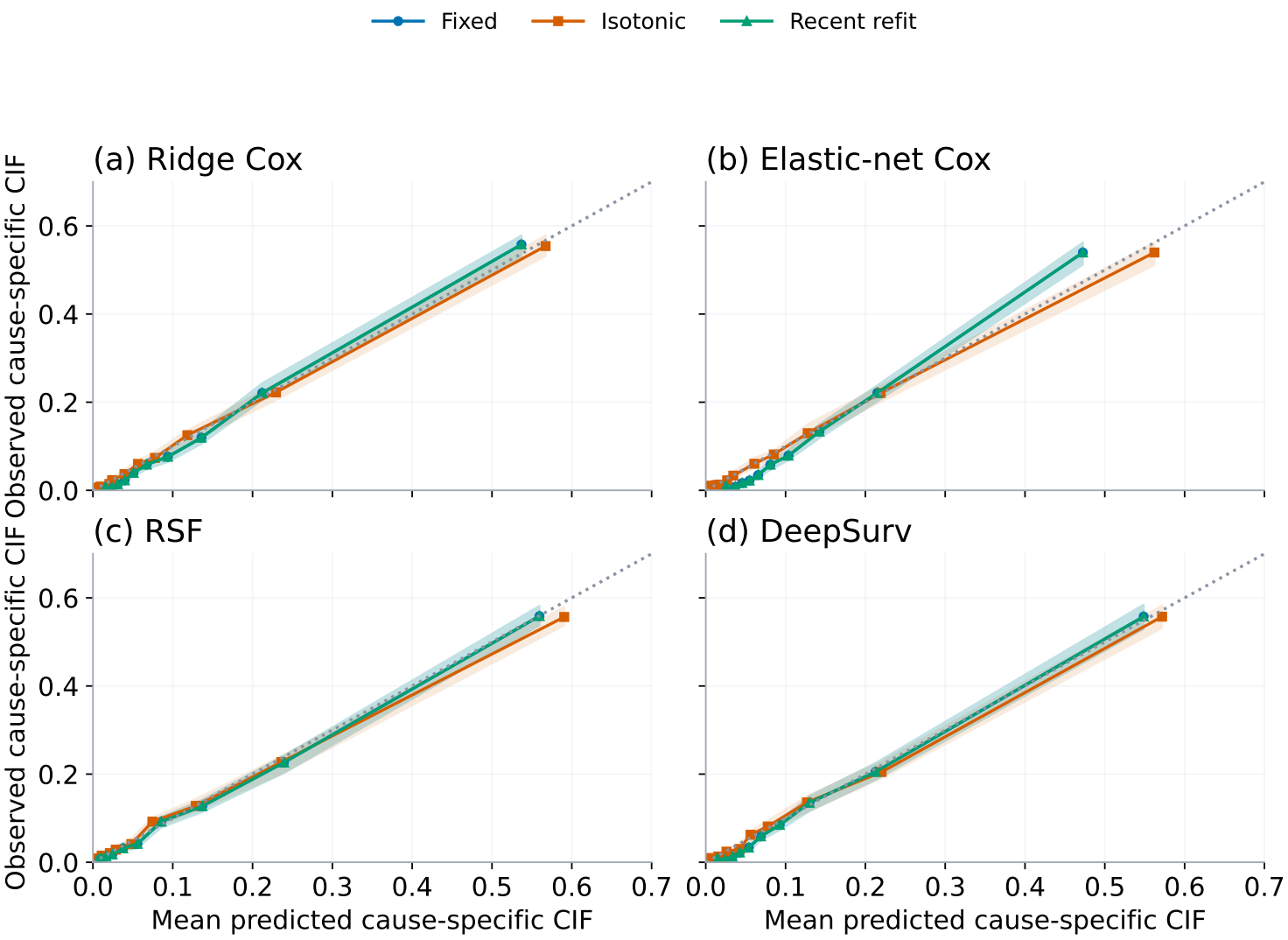
Table S10: Coherence of recalibrated cause-specific risks

Part A. Sensitivity to rescaling incompatible cause-specific risks Proportional rescaling where independently recalibrated cause risks sum above one. Primary results remain unchanged. This descriptive check does not validate a jointly coherent prospective updater. This part contains 40 rows.

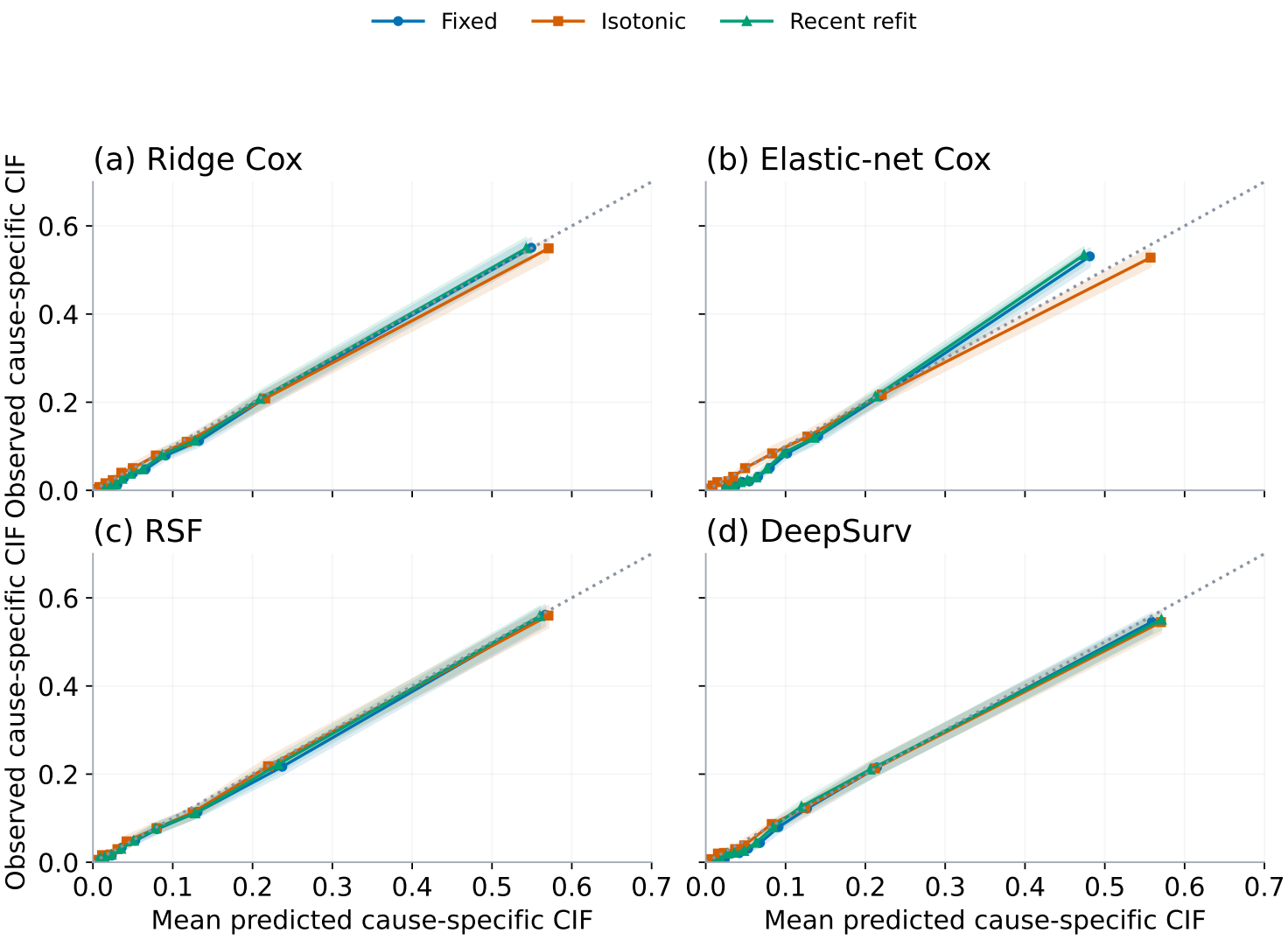
Supplementary Figure S1

Quantile-bin calibration curves for all four families, all five eras, both causes, and all three primary strategies. Points use original evaluation samples. Shaded regions show 95% pointwise binwise intervals from 200 evaluation-patient bootstrap replicates, conditional on fitted models and updates; these are not simultaneous bands. The ten figure pages follow.

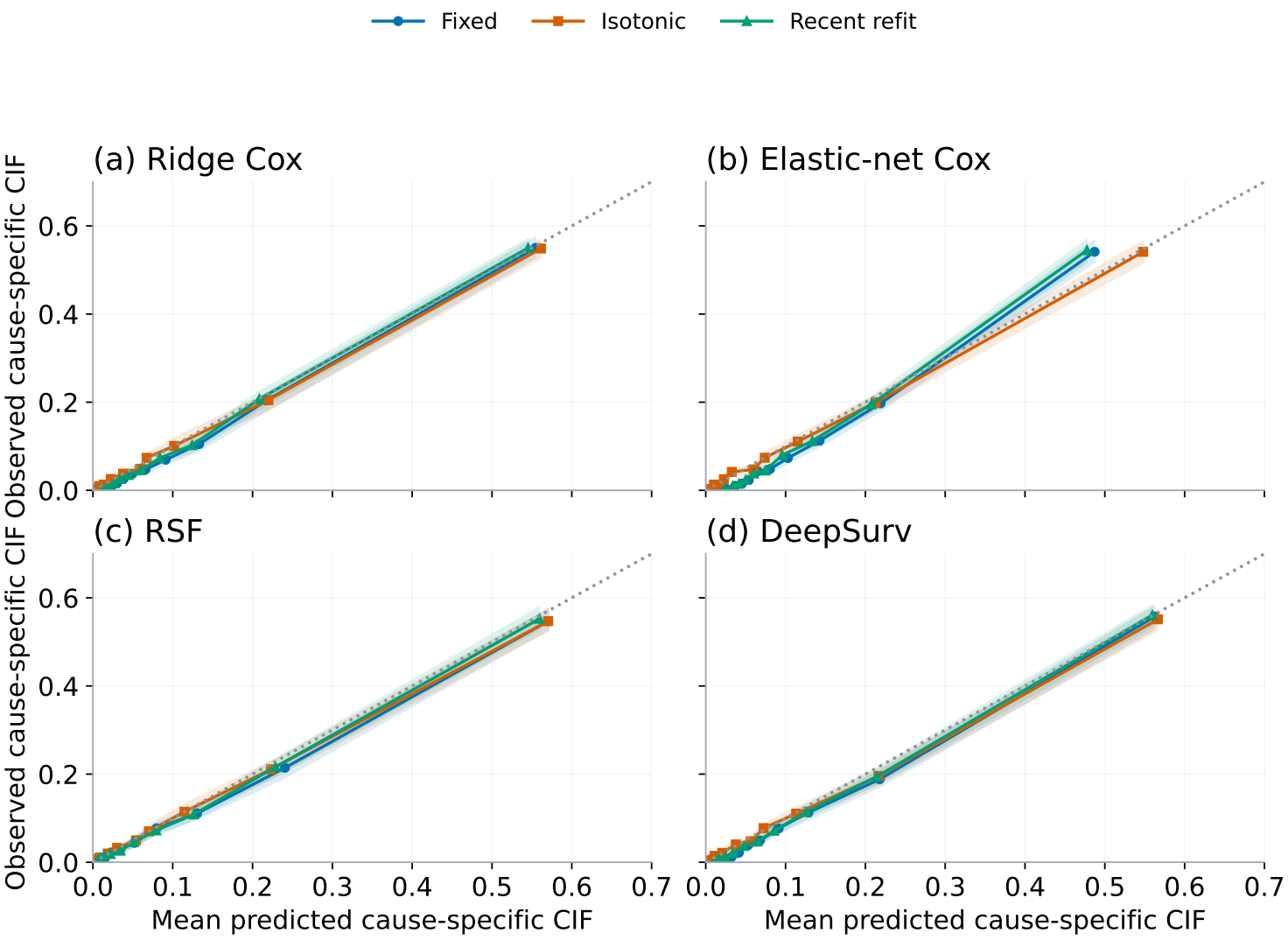
Supplementary Figure S1 — BCSM, 2008–2009
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



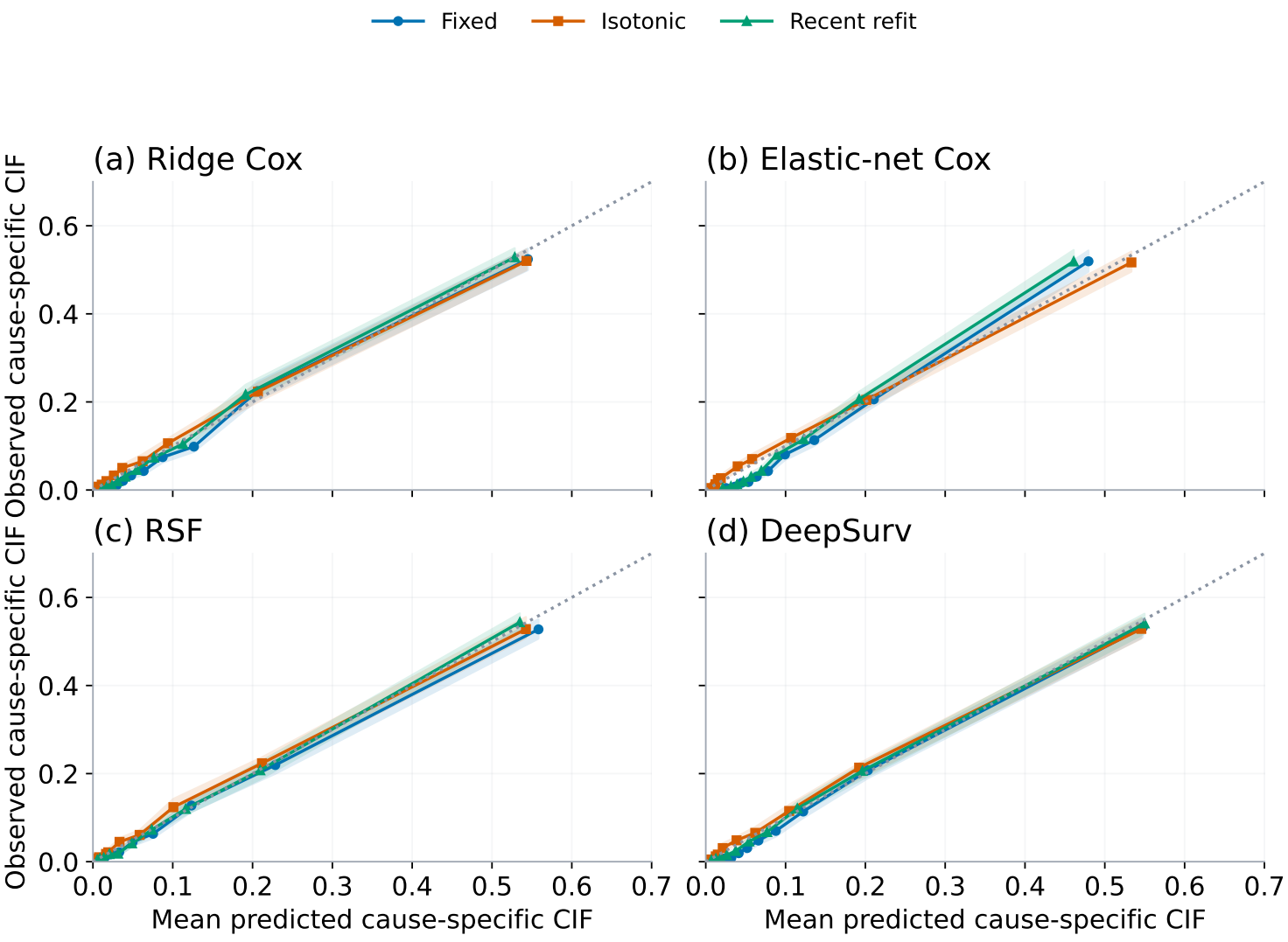
Supplementary Figure S1 — BCSM, 2010–2011
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



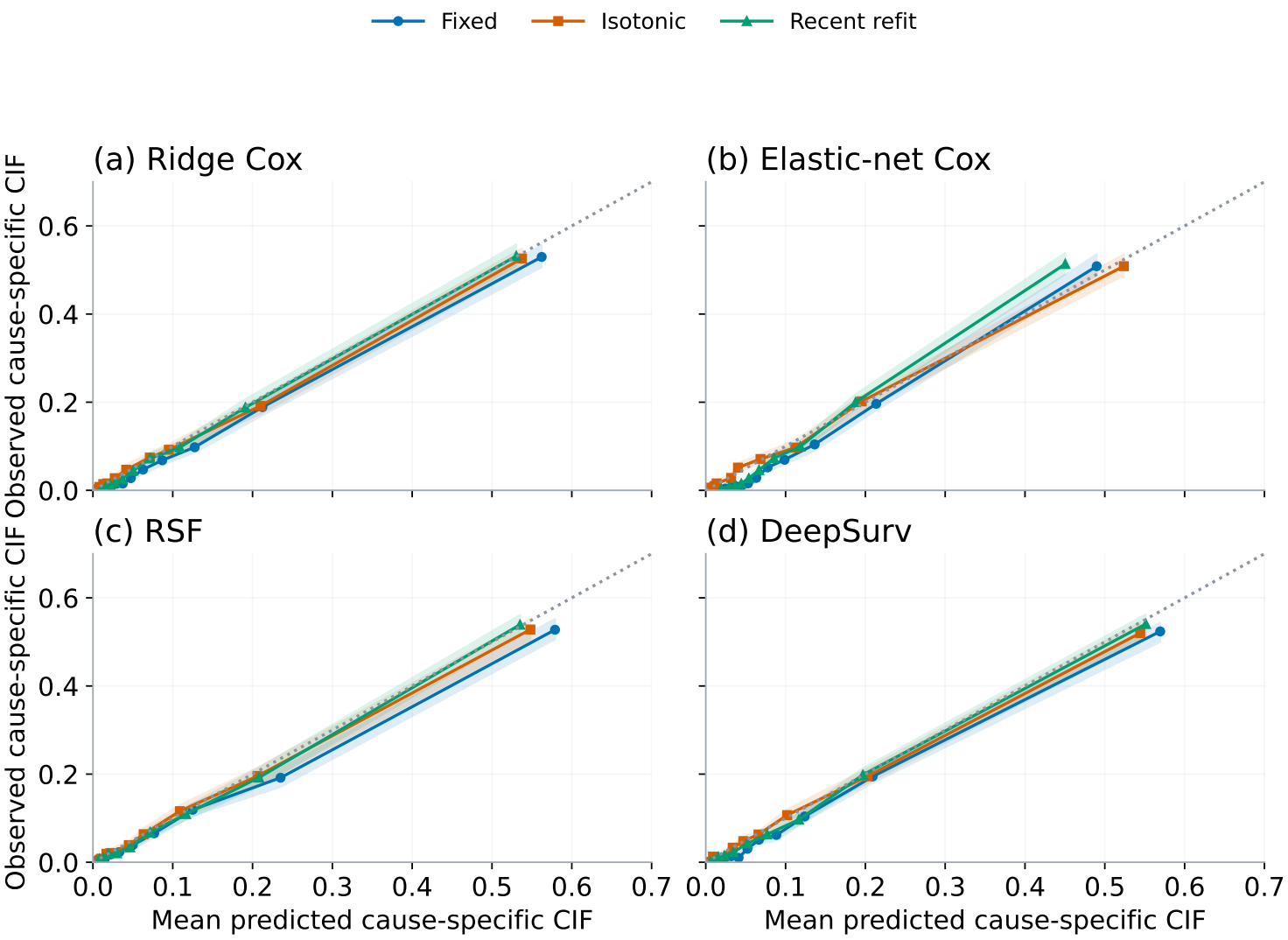
Supplementary Figure S1 — BCSM, 2012–2013
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



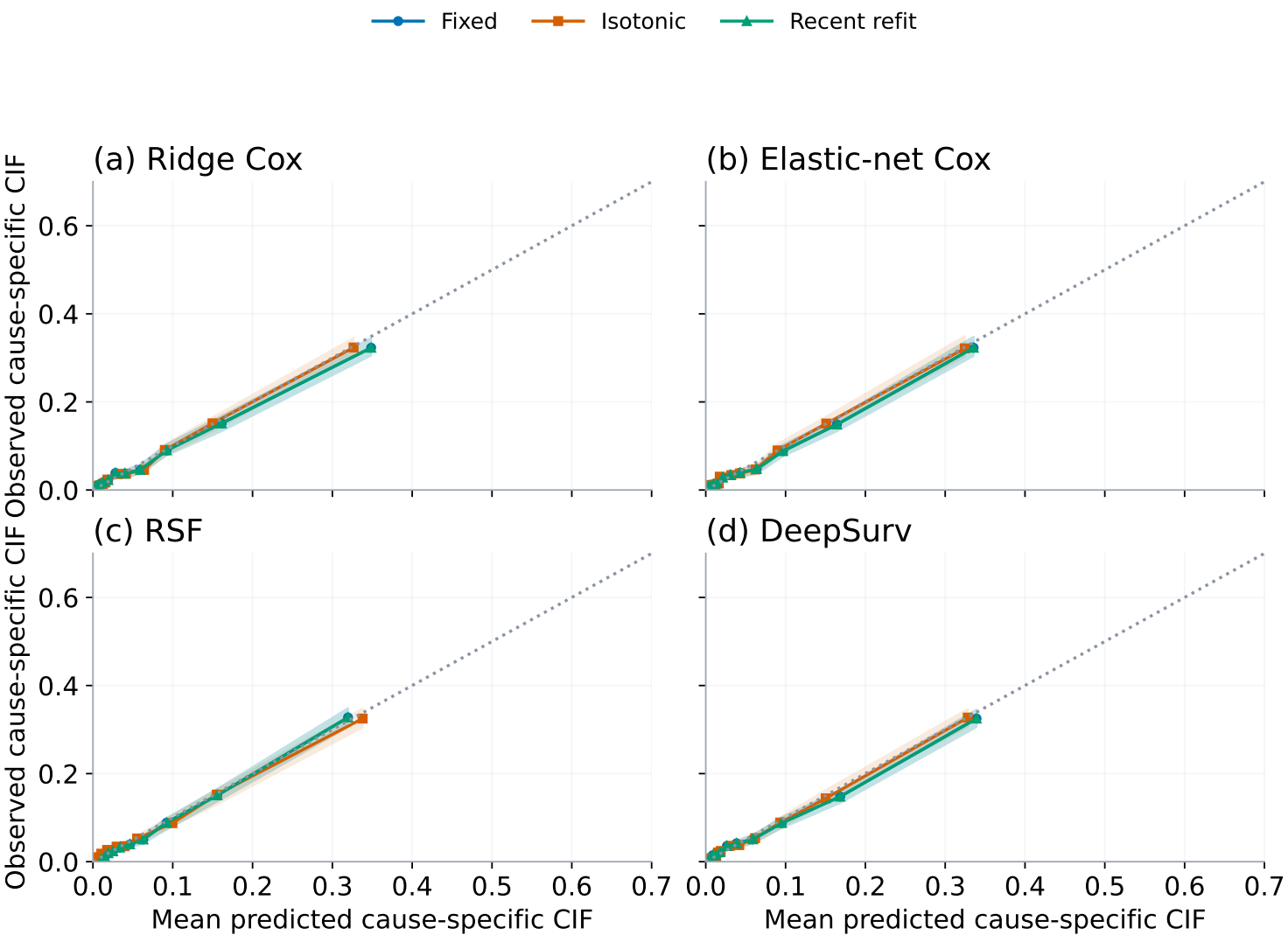
Supplementary Figure S1 — BCSM, 2014–2015
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



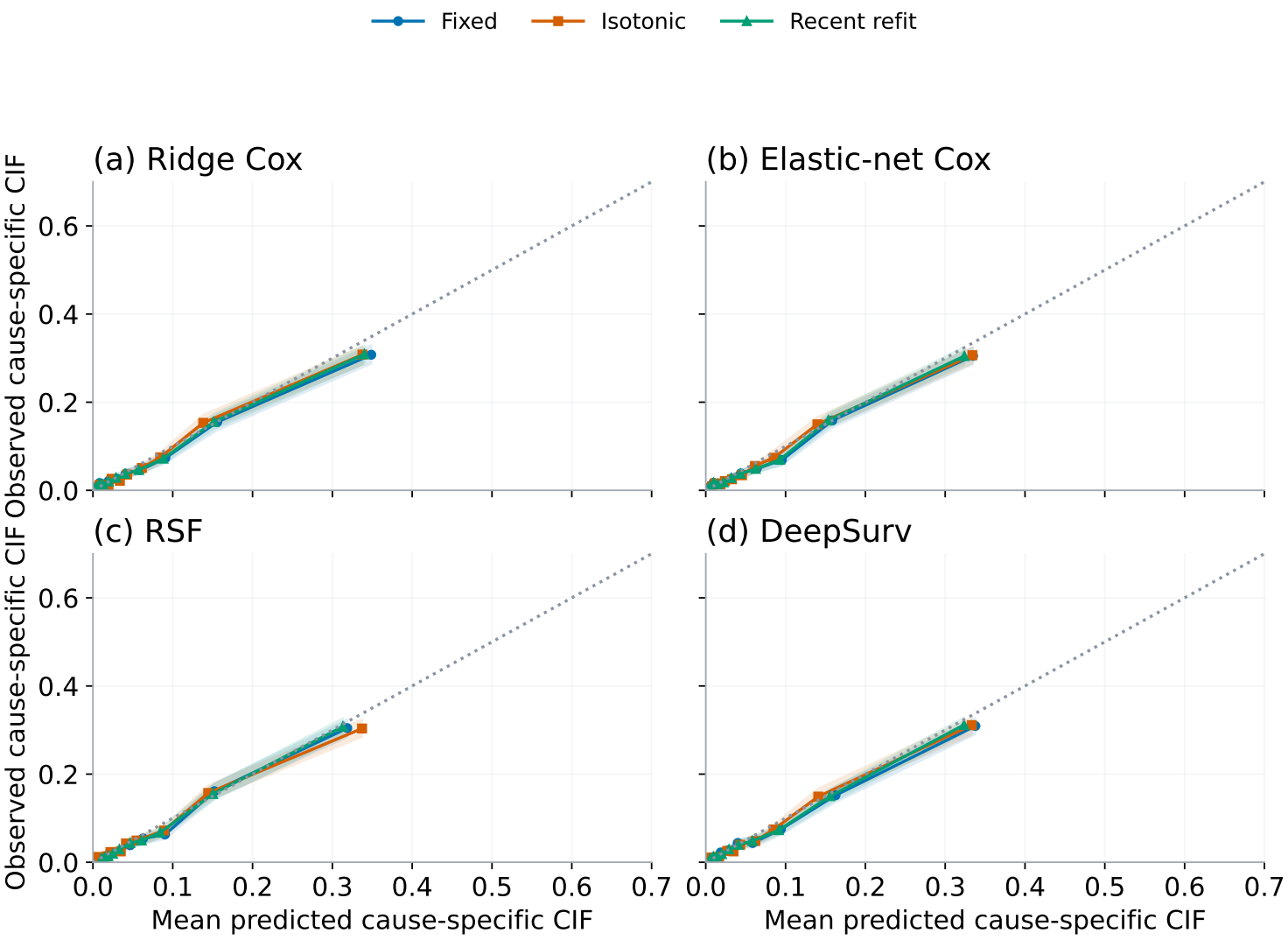
Supplementary Figure S1 — BCSM, 2016–2017
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



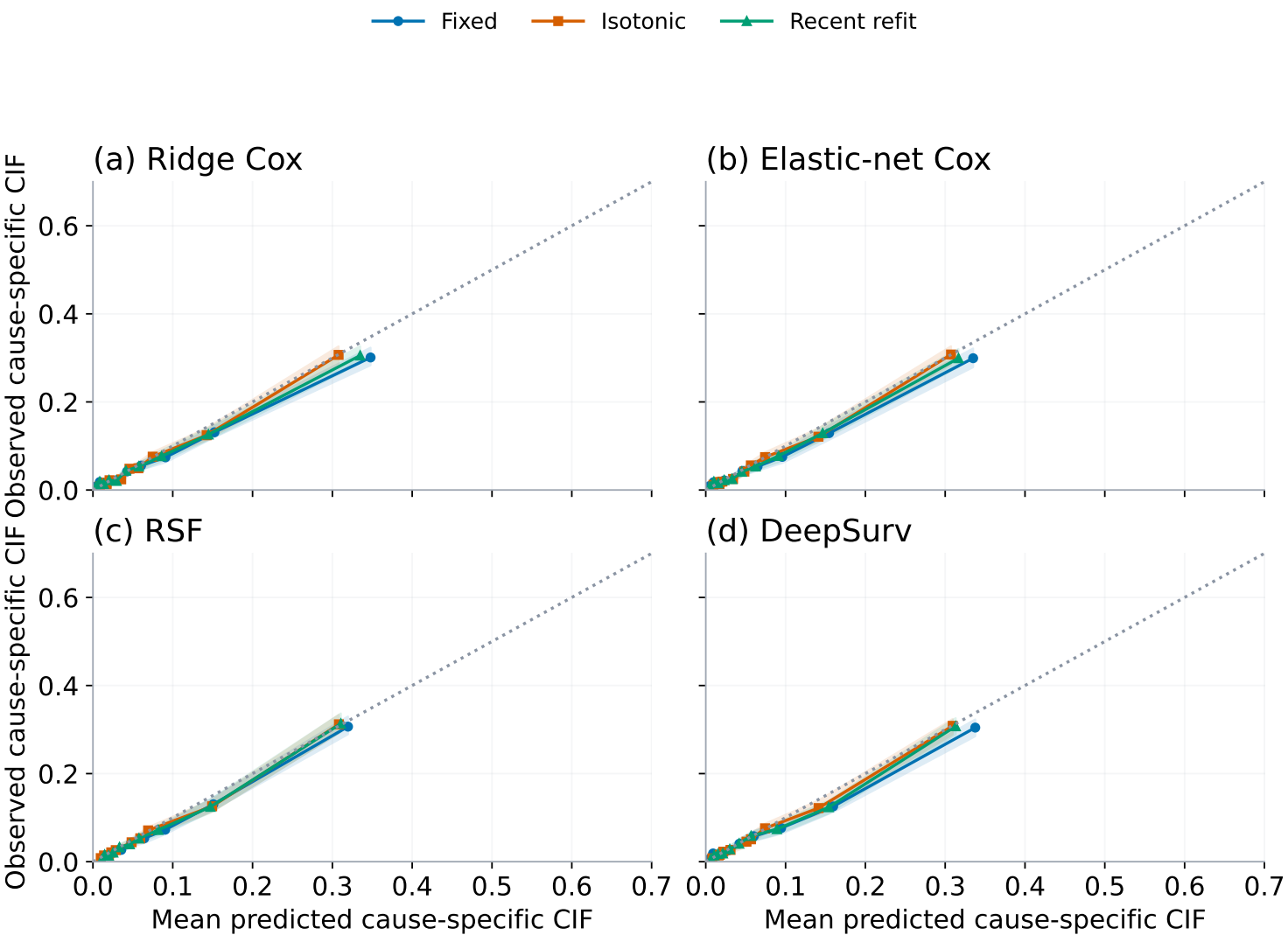
Supplementary Figure S1 — OCM, 2008–2009
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



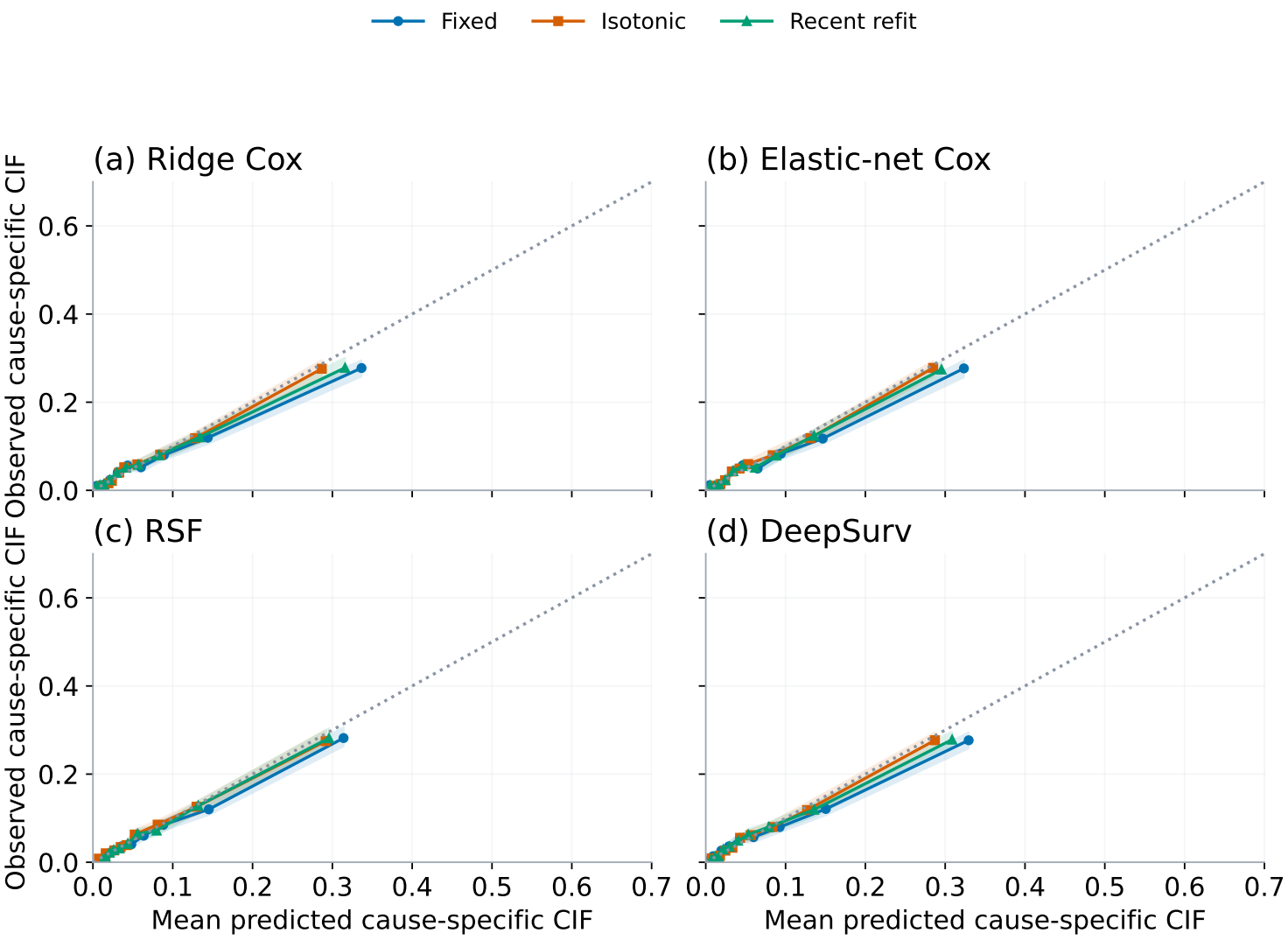
Supplementary Figure S1 — OCM, 2010–2011
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



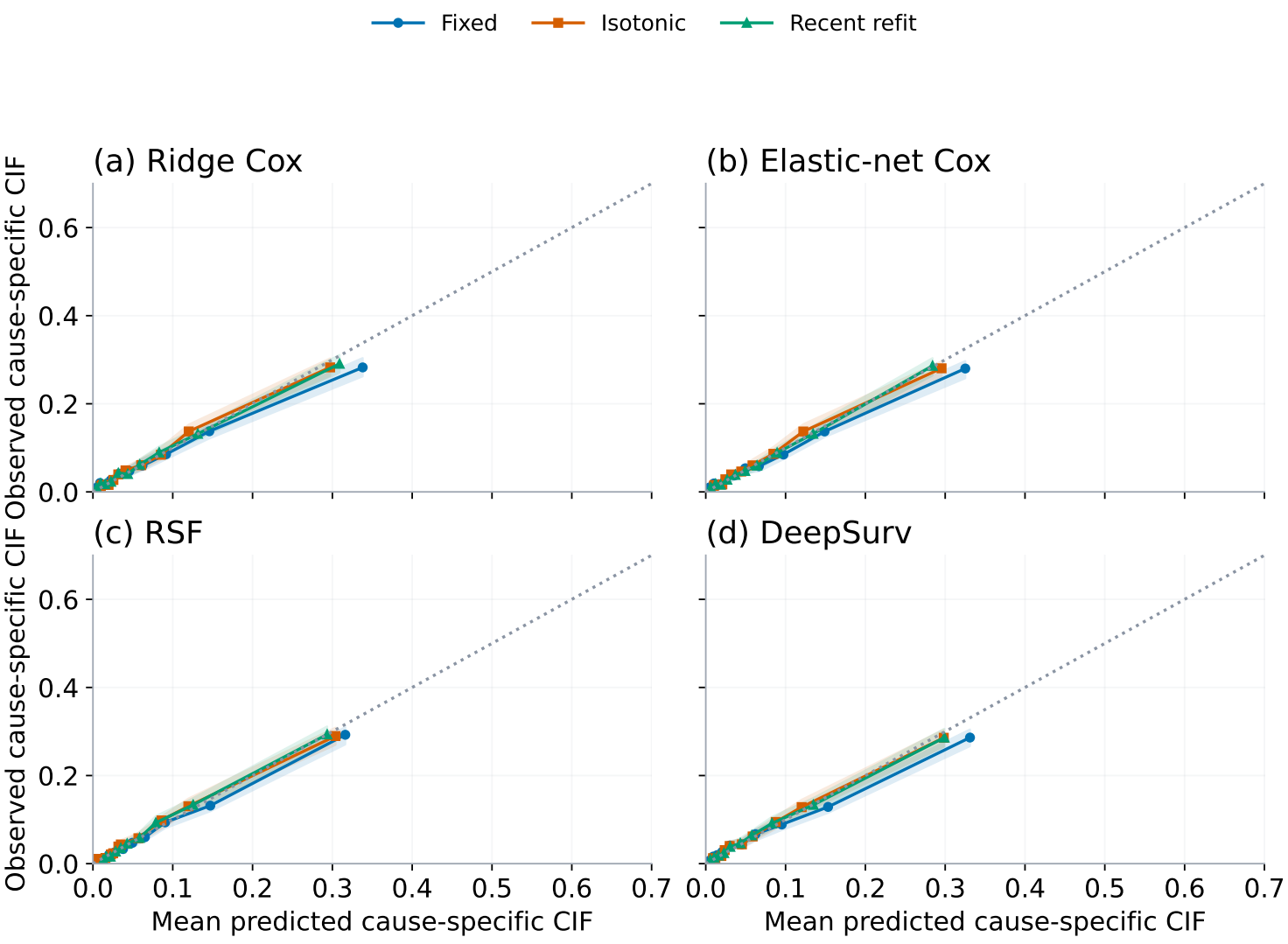
Supplementary Figure S1 — OCM, 2012–2013
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



Supplementary Figure S1 — OCM, 2014–2015
Original sample; pointwise binwise 95% evaluation-bootstrap intervals

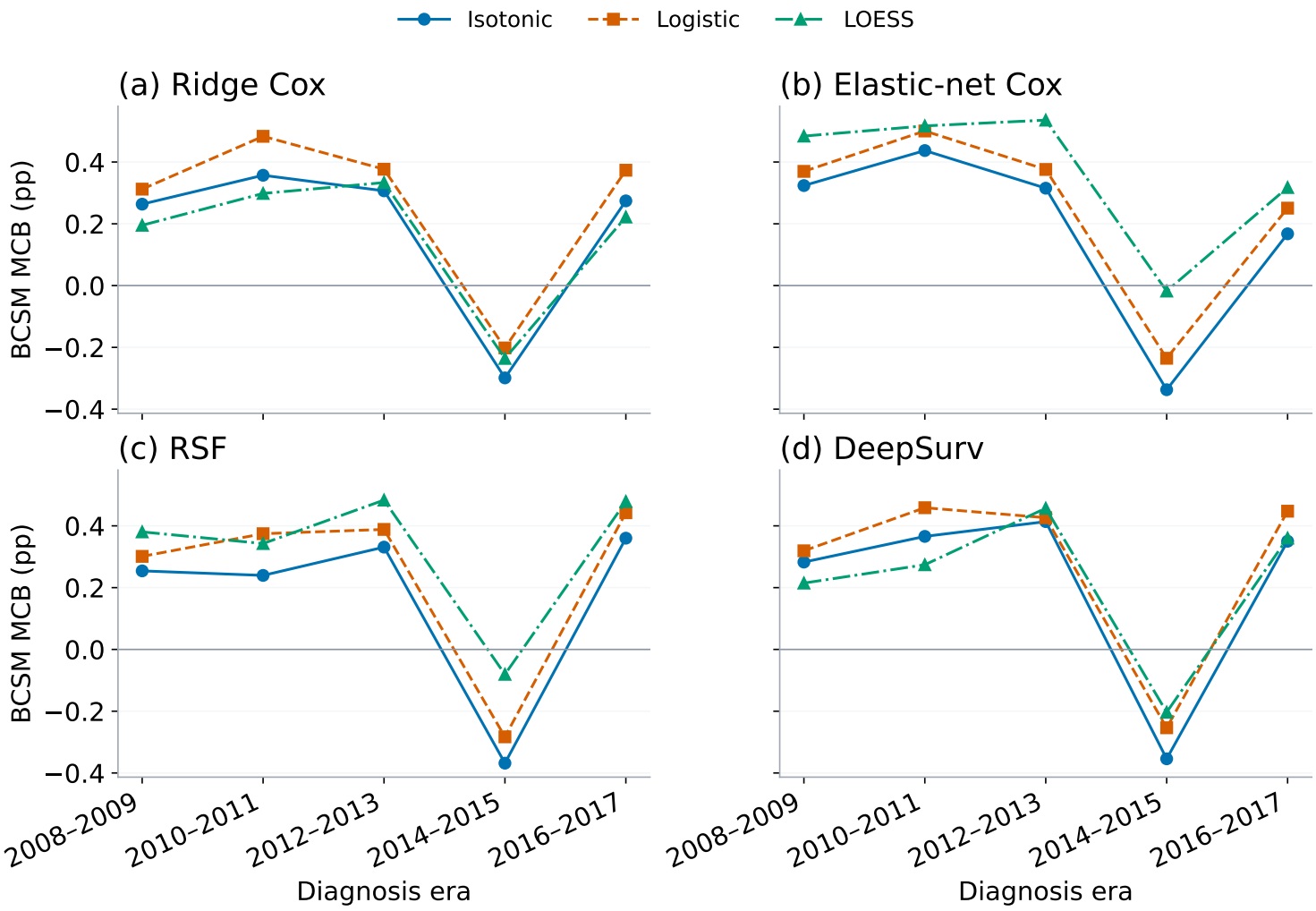


Supplementary Figure S1 — OCM, 2016–2017
Original sample; pointwise binwise 95% evaluation-bootstrap intervals



Supplementary Figure S2

Frozen-model and isotonic, logistic and LOESS mean calibration bias across eras. Calibrators use identical preceding-window patients and censoring weights; target-era outcomes are used only for evaluation. MCB is expressed in percentage points.


Isotonic
Logistic
LOESS
(a) Ridge Cox
(b) Elastic-net Cox
(c) RSF
(d) DeepSurv
BCSM MCB (pp)
0.4
0.2
0.0
−0.2
−0.4
2008–2009
2010–2011
2012–2013
2014–2015
2016–2017
Diagnosis era