

Schema-aware maintenance improves temporal calibration of prostate cancer survival predictions

Said Elshafae

Said.Elshafae@utah.edu

University of Utah

Sayed Youssef

Imam Mohammad Ibn Saud Islamic University

Rania Elshafae

Benha University

Maha Moussa

Utah State University

Article

Keywords: prostate cancer; temporal distribution shift; probability calibration; model maintenance; survival prediction; SEER

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10988852/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Schema-aware maintenance improves temporal calibration of prostate cancer survival predictions

Said Elshafae^{1,2,*}

Sayed Youssef³

Rania Elshafae⁴

Maha Moussa^{5,6}

¹School of Medicine & Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA

²Department of Pathology, Faculty of Veterinary Medicine, Benha University, Benha, Qalyobiya, Egypt 13511

³College of Languages and Translation, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

⁴Fellow, Department of Forensic Medicine and Clinical Toxicology, Benha University Hospitals, Faculty of Medicine, Benha University, Egypt

⁵Department of Mathematics & Statistics, Utah State University, Logan, Utah, USA

⁶Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Benha University, Benha, Qalyobiya, Egypt 13511

*Corresponding author: Said.Elshafae@utah.edu (Said Elshafae)

Abstract

Clinical risk models can remain discriminative while their predicted probabilities drift as data definitions and patient populations change. We evaluated a schema-aware rolling-origin maintenance strategy for a frozen prostate-cancer survival model using 1,334,775 Surveillance, Epidemiology, and End Results records. A discrete-time neural cause-specific hazard model was trained on 2010–2015 diagnoses, selected in 2016, frozen, and evaluated in 2017–2022. For 2019–2022, maintenance actions were selected from earlier diagnosis-year outcome cohorts among no update, global recalibration, and partially pooled local recalibration using a prespecified utility margin. Among 237,764 policy-evaluation records, the frozen model remained discriminative (AUROC 0.890) but overpredicted 12-month prostate-cancer-specific mortality. The utility-gated policy reduced Brier score from 0.019268 to 0.018006, a 6.55% improvement (clustered 95% CI for absolute gain, 0.001158–0.001366), and improved calibration slope from 0.799 to 0.913. Cause-specific IPCW evaluation showed a 6.83% improvement, a conservative two-year history-lag sensitivity showed 4.36%, and a dedicated competing-risk model showed a smaller 0.63% improvement. A well-calibrated histogram-gradient-boosting comparator was left unchanged. These results support probability-scale maintenance within SEER while requiring independent external validation before clinical deployment.

Keywords: prostate cancer; temporal distribution shift; probability calibration; model maintenance; survival prediction; SEER.

1 Introduction

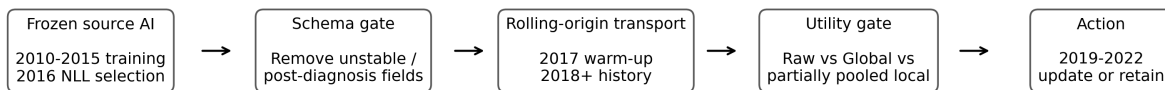
Artificial intelligence (AI) has become a major tool for cancer prognosis, image interpretation, treatment-support research, and population risk stratification. A model that performs well during development, however, is not guaranteed to remain reliable years later. Oncology is especially nonstationary: treatment standards change, diagnostic intensity evolves, coding systems are revised, patient populations shift,

and data fields can appear or disappear. In prostate cancer specifically, progression reflects substantial biological heterogeneity and complex tumor–microenvironment interactions [1]. Recent work in clinical machine learning therefore emphasizes temporally ordered evaluation rather than random train–test splitting [2, 3, 4], while deployment guidance increasingly treats monitoring and updating as lifecycle responsibilities rather than one-time validation tasks [5, 6, 7]. Temporal distribution shift can also degrade uncertainty estimation and calibration under realistic non-i.i.d. evolution [8].

Calibration deserves particular attention because a model may continue to rank patients well while its probability scale becomes systematically too high or too low. A stable AUROC does not imply stable absolute risk, and miscalibration can affect risk communication or threshold-based workflows even when discrimination remains useful [9, 10]. Temporal oncology studies have documented this pattern [11], and recalibration or lifelong updating has been proposed to address drift in prognostic probabilities [12, 13, 14].

Two deployment problems are easily conflated. First, apparent drift may reflect a change in data definition rather than a change in biology or care. Second, statistical evidence of calibration mismatch does not establish that an update is operationally worthwhile. In a large registry, tiny deviations can become statistically detectable long before they produce a meaningful reduction in predictive risk. More broadly, auditing under distribution shift depends on the target-label budget and design of the audit itself [15]. Maintenance rules should therefore separate evidence of change from evidence that a corrective action improves predictive utility.

We study a schema-aware, utility-gated lifecycle strategy for a frozen neural prostate-cancer model in SEER. The question is not which learner wins a static leaderboard; it is when a deployed probability scale should be left alone, recalibrated globally, or recalibrated locally. The workflow contains four safeguards (Figure 1): an outcome-blind schema audit, strict temporal separation of source and deployment eras, rolling-origin candidate-risk estimation from earlier diagnosis-year cohorts, and an operational utility margin that prevents automatic updating when improvement is too small. The contribution is an integration of established recalibration and partial-pooling ideas into an oncology-AI lifecycle, with population-scale temporal evaluation, censoring-aware and competing-risk sensitivities, source-model falsification, subgroup auditing, and explicit timing sensitivity. Reporting was guided by TRIPOD+AI and PROBAST+AI [16, 17].



Recalibration is authorized only when earlier diagnosis-year outcome cohorts show Brier improvement beyond a prespecified utility margin.

A two-year history-lag sensitivity tests a more conservative outcome-availability assumption.

Figure 1: Schema-aware, utility-gated model-maintenance workflow. The source neural survival model is frozen before the deployment era. Recalibration candidates are evaluated using earlier diagnosis-year outcome cohorts, and an update is authorized only when estimated Brier improvement exceeds the operational utility margin. A separate two-year history-lag analysis provides a conservative timing sensitivity.

2 Results

2.1 Temporal transport preserved ranking but shifted the probability scale

The supplied SEER prostate-cancer export contained 1,334,775 records. The corrected source model was fitted to 299,787 diagnoses from 2010–2015, selected on 51,272 diagnoses from 2016, and then frozen for 359,748 deployment-era records from 2017–2022. The primary maintenance evaluation comprised 237,764 records diagnosed in 2019–2022, including 4,746 prostate-cancer-specific deaths within 12 months.

Across 2017–2022, observed 12-month prostate-cancer-specific mortality (PCSM) ranged from 1.86% to 2.07%. Mean frozen-model risk was 1.94% in 2017 but rose to 2.97% in 2018 and approximately 2.80–3.10% in 2019–2022. Discrimination remained stable, with annual AUROC between 0.888 and 0.892 (Figure 2). The dominant transport failure was therefore probability-scale drift rather than collapse of ranking performance.

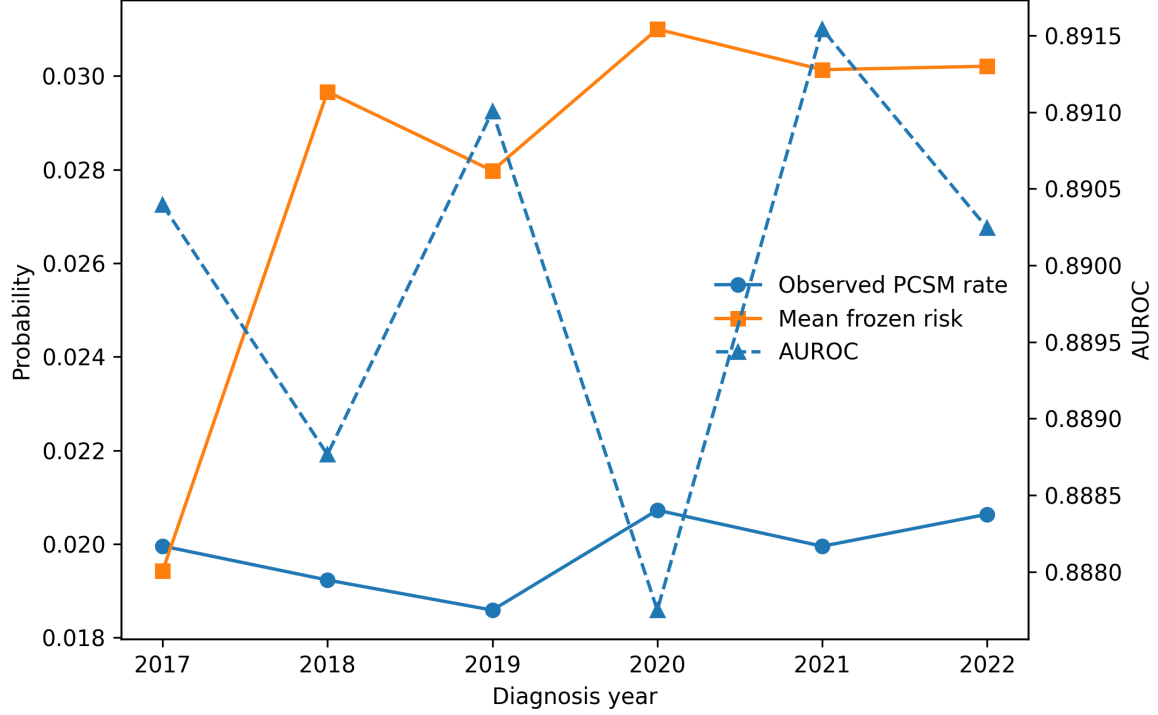


Figure 2: Temporal transport of the corrected frozen neural model. Left: observed 12-month PCSM and mean predicted risk by diagnosis year. Right: annual AUROC. Ranking performance remained stable while the probability scale became too high after 2017.

2.2 Rolling-origin maintenance reduced Brier risk by 6.55%

For 2019–2022, the frozen model had pooled Brier score 0.019268, AUROC 0.890260, calibration intercept -0.936 , calibration slope 0.799, and quantile-based ICI 0.01010 (Table 1). The utility-gated rolling-origin policy reduced Brier score to 0.018006, an absolute improvement of 0.001262 (6.55%). A Patient-ID-clustered paired analysis gave a 95% CI of 0.001158–0.001366 for the absolute Brier gain. Calibration slope improved to 0.913 and ICI to 0.00448. The small AUROC change is expected because recalibration primarily changes probabilities rather than ranking.

Table 1: Corrected pooled performance in the 2019–2022 policy-evaluation cohort.

Action	Brier	AUROC	Cal. intercept	Cal. slope	ICI	Mean risk
Frozen Raw AI	0.019268	0.890260	-0.936	0.799	0.01010	0.02981
Global recalibration	0.017641	0.890254	-0.162	0.977	0.00350	0.02176
Local EB candidate	0.017626	0.891992	-0.159	0.978	0.00344	0.02175
Risk-guided policy	0.018006	0.889741	-0.407	0.913	0.00448	0.02329

The rolling-origin action sequence was Raw/Global/Global/Global for 2019–2022 (Table 2). In 2019, earlier diagnosis-year evidence did not justify an update even though hindsight global recalibration would have improved that year’s Brier score. From 2020 onward, the accumulated earlier-cohort evidence crossed the utility threshold and global recalibration was selected. Local empirical-Bayes recalibration had slightly lower hindsight Brier risk than global recalibration, but its incremental improvement was

too small to justify the more complex action under the ordered gate.

Table 2: Year-specific corrected Brier scores and realized rolling-origin actions.

Year	n	Events	Raw	Global	Local EB	Policy	Action
2019	59,614	1,108	0.018157	0.016700	0.016699	0.018157	Raw
2020	52,889	1,096	0.020283	0.018434	0.018429	0.018434	Global
2021	62,550	1,248	0.019075	0.017437	0.017408	0.017437	Global
2022	62,711	1,294	0.019662	0.018069	0.018047	0.018069	Global

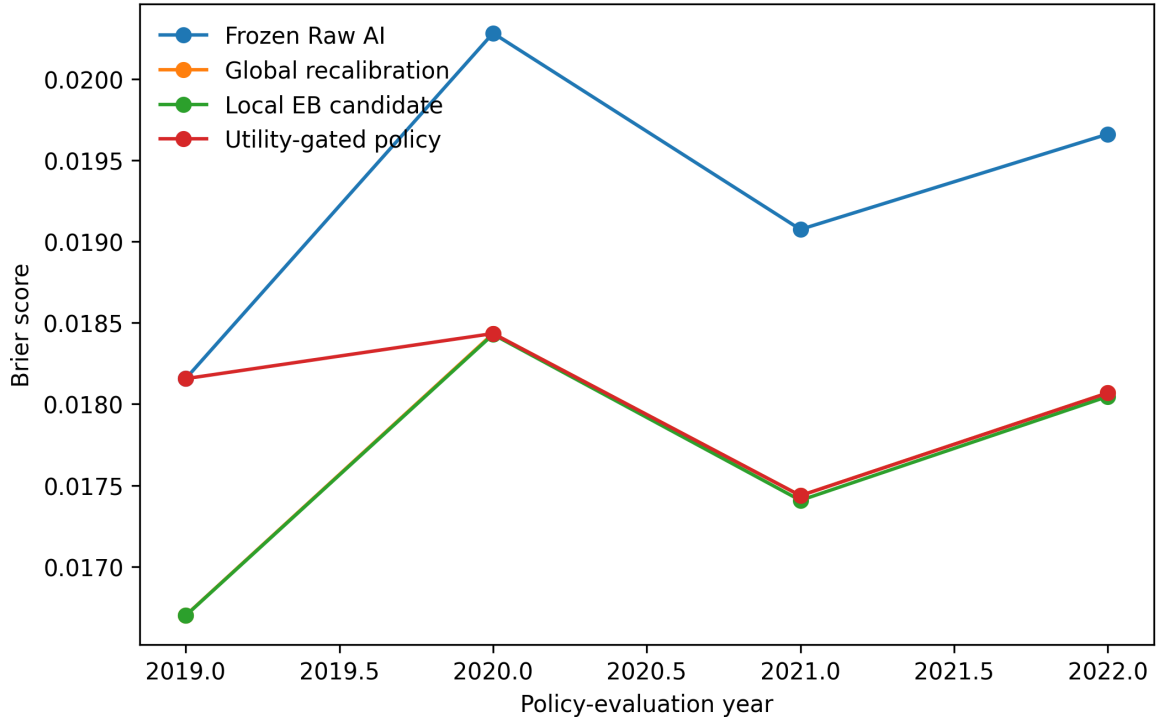


Figure 3: Corrected yearly Brier risk for the frozen model, global recalibration, local empirical-Bayes recalibration, and the realized utility-gated policy.

2.3 Calibration improved across the risk distribution

The frozen model systematically overpredicted 12-month PCSM in the policy era, whereas the selected policy moved predicted probabilities toward the observed event frequencies across risk bins (Figure 4). This visual pattern is consistent with the improvement in calibration intercept, slope, and ICI reported in Table 1.

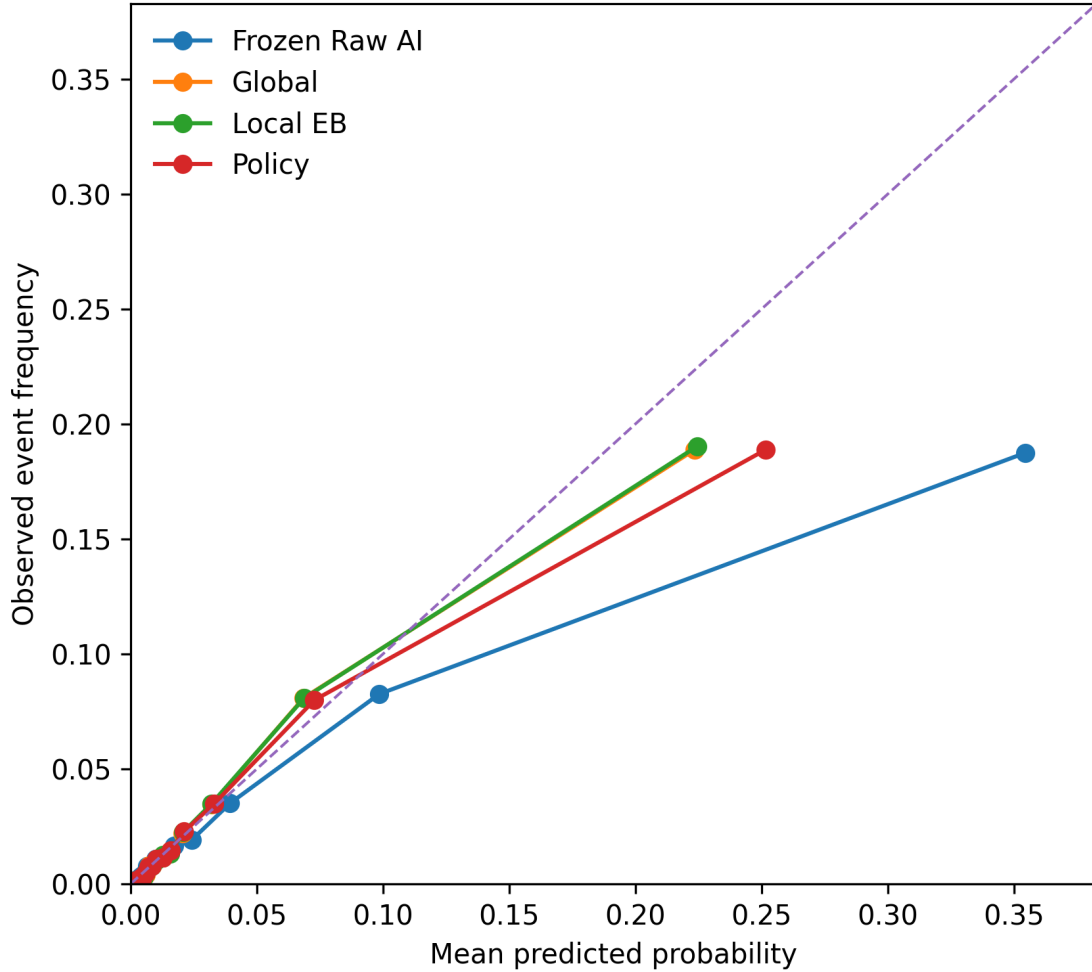


Figure 4: Reliability curves for corrected frozen and policy probabilities in the 2019–2022 evaluation cohort. Points summarize observed event frequencies within prediction bins.

2.4 The main conclusion survived a conservative timing sensitivity

The primary rolling-origin rule is ordered by diagnosis year, not by exact registry outcome-release date. To assess whether immediate use of the preceding diagnosis-year cohort was overly optimistic, we repeated the entire action-selection process with a two-year history lag. Under this stricter timing rule, the action sequence became Raw/Raw/Global/Global. Pooled policy Brier score was 0.018427 versus 0.019268 for the frozen model, an absolute gain of 0.000841 (4.36%; clustered 95% CI 0.000756–0.000925). For the final 2022 cohort, an action selected using diagnosis cohorts no later than 2020 remained Global and reduced Brier score from 0.019662 to 0.018076, an 8.06% relative improvement (95% CI for absolute gain 0.001351–0.001820). Thus, a more conservative outcome-availability assumption attenuated but did not reverse the maintenance signal.

2.5 Subgroup audit showed broadly positive point estimates

The pooled Raw-minus-policy Brier gain was positive in all six race/origin audit strata (Figure 5). The largest well-powered strata had clearly positive clustered intervals: Non-Hispanic White, 0.001314

(95% CI 0.001182–0.001446); Non-Hispanic Black, 0.001513 (0.001245–0.001780); Hispanic, 0.000644 (0.000343–0.000944); and Non-Hispanic Asian or Pacific Islander, 0.001399 (0.000981–0.001817). The American Indian/Alaska Native stratum was much smaller and its interval included zero; the Unknown Race stratum had only four events. These small strata are therefore descriptive rather than evidence of subgroup-specific effectiveness.

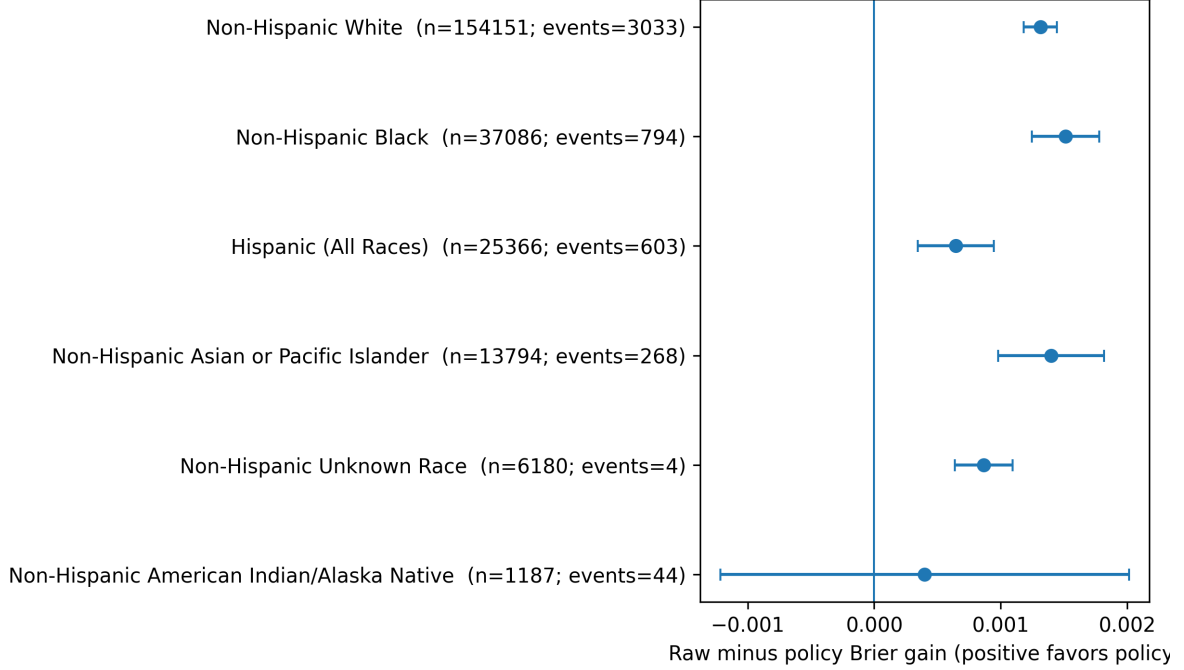


Figure 5: Patient-ID-clustered Raw-minus-policy Brier gain by race/origin audit stratum. Positive values favor the maintenance policy. Small strata are reported descriptively because of limited event counts.

2.6 Utility-threshold and source-model falsification analyses

The prespecified utility margin was $\delta = 0.0001956$. The corrected action sequence Raw/Global/Global/Global and aggregate gain 0.001262 were unchanged for δ values from 0 through 0.0005. At $\delta = 0.001$, the policy became Raw/Raw/Raw/Global and the aggregate gain decreased to approximately 0.000420 (Figure 6). Thus, the reported result is not an artifact of a narrow threshold choice, although sufficiently conservative thresholds defer updates.

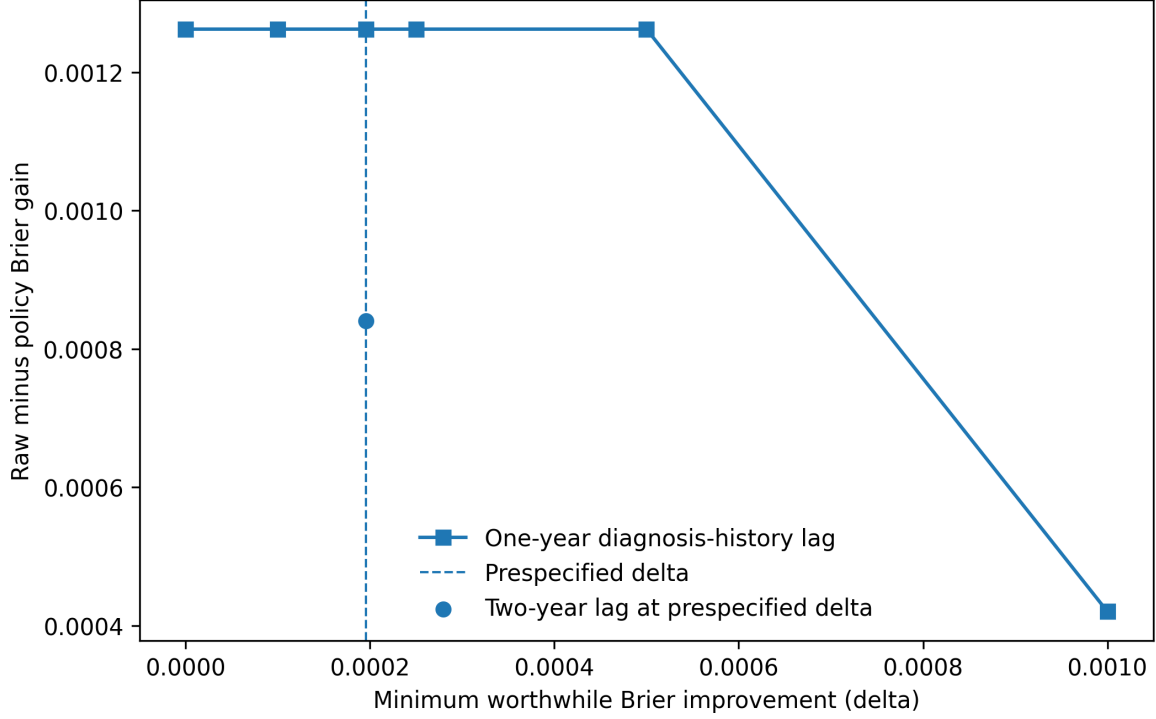


Figure 6: Sensitivity to the operational utility margin and to a two-year history lag. The primary action path remains unchanged across a broad range of utility margins; a two-year lag yields a smaller but still positive aggregate gain.

A source-model falsification analysis showed that the gate did not automatically recalibrate every learner (Table 3). The well-calibrated HGB model had raw Brier 0.017314 and was retained unchanged in every policy year. Race-blind HGB behaved similarly. Logistic and neural variants with larger calibration errors generally earned recalibration and showed positive maintenance gains. These post-hoc variants are sensitivity analyses, not a new model-selection exercise.

Table 3: Corrected source-model sensitivity and falsification results.

Variant	2016 AUROC	Raw Brier	Policy Brier	Relative gain	Actions
HGB	0.9083	0.017314	0.017314	0.00%	R/R/R/R
Logistic	0.8524	0.018753	0.018140	3.27%	G/G/G/G
Logistic + missingness flags	0.9181	0.018047	0.017153	4.95%	G/G/G/G
Neural + missingness flags	0.9261	0.017564	0.017194	2.11%	R/G/G/G
Race-blind HGB	0.9089	0.017328	0.017328	0.00%	R/R/R/R
Race-blind logistic	0.8523	0.018793	0.018159	3.37%	G/G/G/G
Race-blind neural	0.9063	0.020816	0.017725	14.85%	G/G/G/G

2.7 Censoring-aware and competing-risk analyses refined the interpretation

Cause-specific IPCW evaluation supported the primary conclusion: Brier risk decreased from 0.019004 to 0.017706, a 6.83% relative improvement, with action sequence Raw/Global/Global/Global. In contrast, the dedicated neural competing-risk model showed a much smaller but same-direction improvement, from 0.017412 to 0.017303 (0.63%), with action sequence Raw/Raw/Global/Global (Table 4). This

attenuation is scientifically important: the strongest evidence for maintenance is for the prespecified cause-specific probability scale, and the magnitude should not be generalized to cumulative-incidence prediction under competing risks.

Table 4: Corrected robustness and timing analyses.

Analysis	Raw Brier	Policy Brier	Relative gain	Action sequence
Primary known-status cause-specific	0.019268	0.018006	6.55%	R/G/G/G
Cause-specific IPCW	0.019004	0.017706	6.83%	R/G/G/G
Neural competing-risk IPCW	0.017412	0.017303	0.63%	R/R/G/G
Two-year history-lag primary	0.019268	0.018427	4.36%	R/R/G/G

3 Discussion

This population-scale study addresses a lifecycle problem that is different from ordinary model development. The corrected frozen neural model retained AUROC near 0.89 after temporal transport, yet its probability scale became too high. A schema-aware, utility-gated rolling-origin policy reduced pooled 12-month cause-specific Brier risk by 6.55%, improved calibration slope from 0.799 to 0.913, and produced a tightly positive Patient-ID-clustered interval for absolute gain. The result therefore represents probability-scale maintenance rather than recovery from a collapse in discrimination.

Several findings make the interpretation more specific. First, the action gate was selective rather than mechanically pro-update: HGB, which was already well calibrated, was retained unchanged in every year. Second, the 2019 Raw decision demonstrates the cost of non-oracle action selection. In hindsight, recalibration would have improved 2019, but the earlier diagnosis-year history did not yet justify an update. The framework trades immediate hindsight performance for a rule that can be specified before the target year’s outcomes are inspected. Third, a conservative two-year history lag still yielded a 4.36% aggregate improvement and an 8.06% gain in 2022, reducing concern that the result depends entirely on rapid outcome availability.

Censoring and competing risks require a more nuanced conclusion. The cause-specific IPCW sensitivity closely matched the primary result, with a 6.83% improvement, suggesting that fixed-horizon observability did not drive the principal finding. The dedicated competing-risk model, however, showed only a 0.63% improvement. The direction was preserved but the magnitude was substantially attenuated. The maintenance claim is therefore strongest for the cause-specific 12-month PCSM probability scale and should not be interpreted as evidence that every clinically relevant competing-risk estimand benefits equally from the same update rule.

The study also illustrates why schema auditing belongs upstream of drift diagnosis. SEER field definitions changed across calendar time: the legacy summary-stage field disappeared after 2017, grade coding changed at the 2017/2018 boundary, and the post-2018 PSA field contained a thousands-separator edge case. A frozen model that blindly consumes unstable definitions can manufacture apparent biological drift. We therefore excluded semantically unstable or post-diagnosis variables from the source model and corrected the PSA parser before the final rerun.

A final reproducibility audit identified a separate implementation issue in the original grouped-time neural likelihood: a censored partial interval could be counted as though it had been fully observed event-free. The corrected implementation includes only completed event-free intervals and the interval

containing an observed event; an incompletely observed censored interval is omitted. We refitted the source model and reran every numerical analysis reported in this version on the authorized SEER export. The primary conclusion was stable after correction (6.55% versus the earlier 6.61% estimate), but the competing-risk sensitivity changed materially. Reporting the corrected smaller effect is important because reproducibility should be used to refine claims, not merely confirm them.

These results are consistent with the broader literature emphasizing calibration, temporal validation, and lifecycle monitoring in clinical prediction [10, 2, 6]. The maintenance layer itself is intentionally simple. Logistic recalibration and empirical Bayes are established methods; the novelty is the operational composition: schema audit, frozen temporal transport, rolling-origin comparison of candidate actions, and a practical-risk threshold that can return “do nothing” when updating is not justified.

4 Limitations

This is a retrospective internal temporal study within SEER, not an independent external validation. Registry data are valuable for population-scale stress testing, but an independent health-system or cancer-registry cohort is required before clinical deployment. A SEER export containing a held-out registry identifier could provide a useful geographic transport test, but it would still be internal to the same surveillance system.

The primary rolling-origin chronology is based on diagnosis year rather than the exact date on which each 12-month outcome became available in the registry. We therefore avoid describing the primary sequence as a real-time prospective deployment. The two-year history-lag sensitivity is intentionally conservative and preserves a positive 4.36% aggregate gain, but exact release-date emulation would be preferable when such metadata are available.

The primary neural model estimates a cause-specific PCSM probability, censoring other-cause death. Cause-specific IPCW produced a similar 6.83% gain, but the dedicated competing-risk neural model yielded only 0.63%. This difference shows that the magnitude of the benefit is estimand-dependent. Longer horizons and clinically deployed risk thresholds would require additional competing-risk-aware validation.

The local empirical-Bayes candidate used administrative race/origin strata because the supplied case listing lacked the registry identifier needed for the planned geographic hierarchy. These strata should not be interpreted causally or biologically. Subgroup analyses were calibration audits rather than claims of differential treatment or benefit, and small-event strata are imprecise.

Finally, the source model was deliberately compact and used variables with stable meaning across eras. It was not designed to maximize a static leaderboard score. HGB achieved lower raw Brier risk and was correctly left unchanged by the gate, showing that maintenance need depends on the calibration state of the deployed learner. External validation should evaluate both the frozen source model and the maintenance layer without target-cohort retraining before the maintenance decision is assessed.

5 Conclusions

A frozen prostate-cancer model can remain discriminative while its probability scale drifts over time. After correcting grouped-time censoring and rerunning the full real-data pipeline, the schema-aware

utility-gated policy reduced 12-month cause-specific PCSM Brier risk by 6.55% relative to the frozen model. The signal remained strong under cause-specific IPCW evaluation (6.83%) and a conservative two-year history lag (4.36%), while a dedicated competing-risk model showed a smaller 0.63% improvement. A well-calibrated HGB source model was left unchanged. These findings support selective probability-scale maintenance under temporal shift while underscoring the need for explicit estimands, realistic outcome-availability assumptions, and independent external validation.

6 Methods

6.1 Study design and data source

We performed a retrospective temporal-transport model-maintenance study using a prostate-cancer case-listing export from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program [18]. The source export contained 1,334,775 records. Raw SEER records are governed by SEER data-use requirements and are not redistributed with this manuscript.

The model-development and evaluation sequence was fixed by diagnosis year. Diagnoses from 2010–2015 were used for source-model fitting ($n = 299,787$ after the positive-survival requirement), 2016 for neural-model selection ($n = 51,272$), and 2017–2022 for deployment-era evaluation ($n = 359,748$). The primary fixed-horizon calibration stream contained records whose 12-month cause-specific status was known. Diagnoses from 2019–2022 formed the primary policy-evaluation cohort ($n = 237,764$; 4,746 PCSM events). The corrected source-model selection epoch was 89, with 2016 masked negative log likelihood 0.02334.

The primary maintenance rule is *rolling-origin by diagnosis year*. For action year t , candidate risks are estimated from earlier diagnosis-year cohorts only. Because diagnosis year is not identical to outcome-release date, a separate two-year history-lag sensitivity excludes the immediately preceding diagnosis-year cohort from action selection.

6.2 Endpoint and fixed-horizon observability

The primary endpoint was PCSM within 12 months of diagnosis. The source neural model used a cause-specific convention: death from another cause was treated as censoring for the PCSM hazard likelihood. For fixed-horizon calibration, a label was considered known if PCSM occurred within 12 months or PCSM-free follow-up extended to at least 12 months. Cases without known fixed-horizon status were excluded only from the primary fixed-horizon calibration/maintenance evaluation; censoring masks were retained during survival-model fitting.

Because other-cause death is a genuine competing event for cumulative-incidence prediction, we separately evaluated cause-specific IPCW Brier risk and a dedicated competing-risk neural model. These analyses are sensitivity estimands rather than interchangeable versions of the primary endpoint [19, 20].

6.3 Outcome-blind schema-stability audit

Before interpreting temporal model behavior, we audited predictor availability and coding by diagnosis year without using outcome associations to decide whether a field was transportable. The supplied export contained two major structural discontinuities: the legacy “Summary stage 2000” field became

unavailable after 2017, and the numeric grade scale changed at the 2017/2018 boundary. We excluded stage and grade from the frozen source model rather than allowing coding transitions to masquerade as biological or clinical drift (Figure 7).

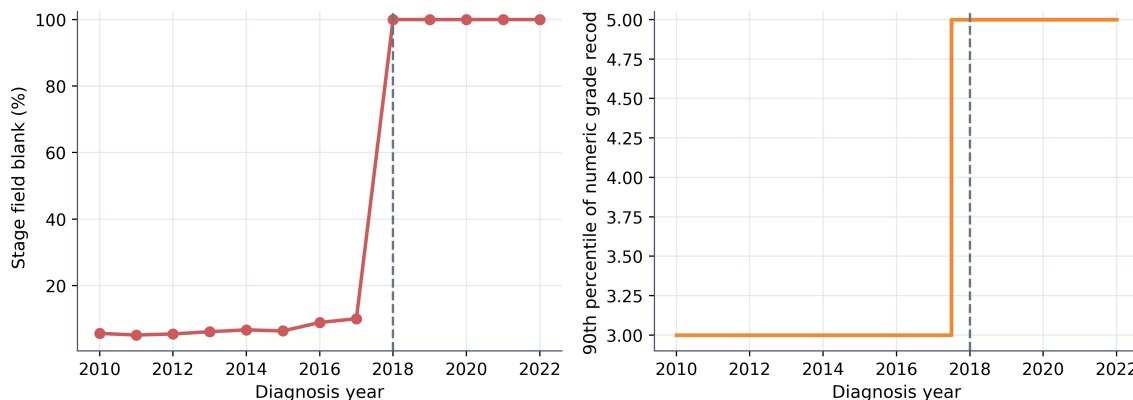


Figure 7: Schema-stability audit illustrating calendar-time discontinuities in stage and grade fields. These variables were excluded from the transport-stable source model.

We also excluded diagnosis year and post-diagnosis treatment variables (surgery, radiation, chemotherapy, and treatment delay). Hispanic origin was reconstructed from the exact SEER label “Hispanic (All Races)” rather than substring matching. The final transport-stable source predictors were age, $\log(\text{PSA} + 1)$, Gleason score, three race indicators, and corrected Hispanic origin.

The parser audit identified a numeric edge case in the 2018+ PSA field: the category “1,000 ng/ml or greater” contains a thousands separator. The final parser removes the separator before numeric extraction. Missing values were handled using source-era medians for continuous variables and explicit category/indicator logic documented in the reproducibility code (Figure 8).

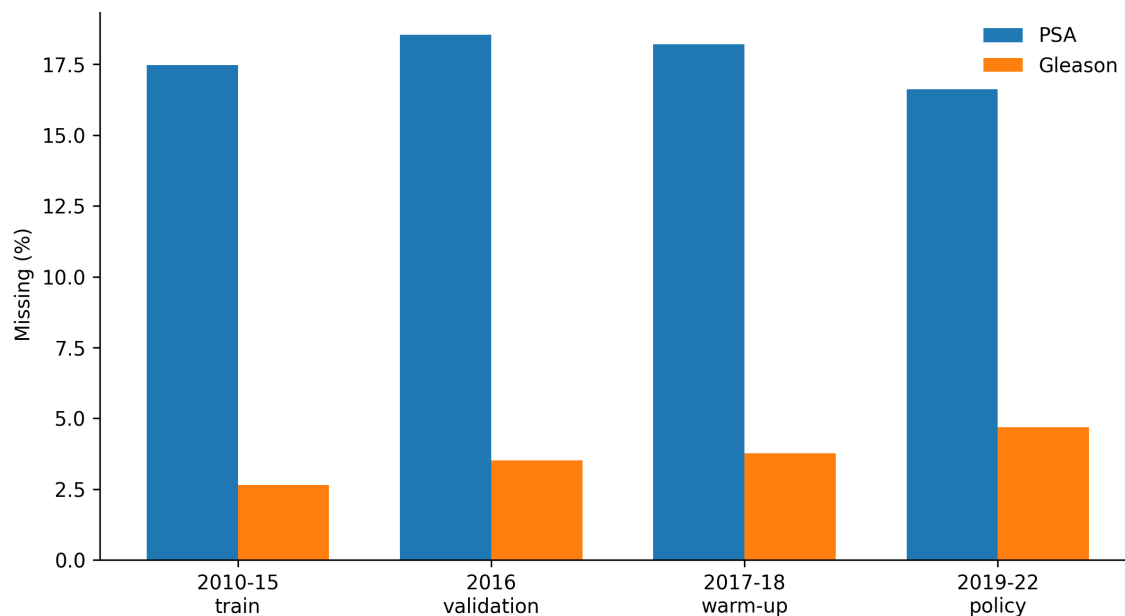


Figure 8: Predictor missingness audit across source and deployment eras.

6.4 Frozen discrete-time neural survival model

The source AI was a discrete-time neural cause-specific hazard model with four 3-month intervals. The final seven predictors were standardized or encoded using source-era transformations. The network used two 128-unit ReLU hidden layers, dropout 0.10, AdamW with learning rate 10^{-3} and weight decay 10^{-4} , mini-batches of 8,192, and gradient clipping at 5. Model selection used 2016 negative log likelihood; the selected model was then frozen before deployment-era maintenance analyses.

A reproducibility audit before this Scientific Reports submission identified a grouped-time censoring implementation issue in the prior code. The corrected likelihood mask follows an explicit interval-observability rule: an event-free interval contributes only when follow-up reaches the end of that interval; the interval containing an observed PCSM event contributes with event indicator 1; and an incompletely observed censored interval is omitted. The source model, deployment predictions, maintenance analyses, source-model sensitivities, IPCW analysis, competing-risk model, tables, and figures reported here were regenerated after this correction using the authorized SEER export.

6.5 Candidate recalibration actions

Let p_i denote the frozen model’s 12-month probability. Global recalibration used logistic calibration

$$\text{logit}(q_i) = \alpha + \beta \text{logit}(p_i),$$

with estimated intercept α and slope β [9, 10]. The local candidate used empirical-Bayes partial pooling of calibration intercepts/slopes across administrative race/origin strata. The local model was treated as a candidate maintenance action, not as a causal race model.

The Raw action retained p_i unchanged. All recalibration maps for an action year were fitted only to the diagnosis-year cohorts admitted to that year’s history. Candidate probability maps were then frozen before evaluation on the target diagnosis year.

6.6 Utility-gated rolling-origin action rule

Maintenance decisions used Brier risk, a strictly proper probability score [21, 22]. Let $\hat{R}_{\text{Raw},t}$, $\hat{R}_{\text{Global},t}$, and $\hat{R}_{\text{Local},t}$ denote candidate risks estimated from the permitted earlier diagnosis-year history for action year t . The operational utility margin δ was defined as 1% of the 2017 prevalence-reference Brier risk, yielding $\delta = 0.0001956$. The ordered action rule was

$$A_t = \begin{cases} \text{Local,} & \hat{R}_{\text{Local},t} + \delta < \min(\hat{R}_{\text{Raw},t}, \hat{R}_{\text{Global},t}), \\ \text{Global,} & \hat{R}_{\text{Global},t} + \delta < \hat{R}_{\text{Raw},t}, \\ \text{Raw,} & \text{otherwise.} \end{cases} \quad (1)$$

The margin is a governance criterion rather than a universal clinical-effect threshold. We varied δ from 0 to 0.001 in sensitivity analyses.

6.7 Performance metrics and uncertainty

The primary estimand was Raw-minus-policy Brier gain pooled across 2019–2022. Secondary measures included AUROC, calibration intercept and slope, mean predicted risk, and a 20-quantile approximation to the Integrated Calibration Index [23]. Because a small number of SEER Patient IDs appeared on more than one case record, paired Brier-gain uncertainty used a Patient-ID-clustered sandwich standard error. Subgroup analyses repeated paired clustered gain calculations within administrative race/origin strata and were interpreted as calibration audits.

6.8 Timing, source-model, censoring, and competing-risk sensitivities

The two-year history-lag sensitivity repeated the maintenance procedure while excluding the immediately preceding diagnosis-year cohort from action selection. We also isolated the 2022 cohort, for which the two-year-lag action used diagnosis cohorts no later than 2020.

Source-model sensitivity used fixed-horizon logistic regression and histogram gradient boosting, together with race-blind and missingness-indicator variants. These learners were fitted in the source era and then passed through the same maintenance gate. The goal was falsification: a well-calibrated source learner should be allowed to remain unchanged rather than being forced to recalibrate.

For cause-specific IPCW evaluation, censoring weights were estimated from the censoring distribution and applied to the policy-era deployment records. For the dedicated competing-risk sensitivity, each 3-month interval had mutually exclusive no-event, PCSM, and other-cause-death states. The neural model used a softmax output and masked multinomial negative log likelihood. Completed event-free intervals contributed to the likelihood; an observed event interval contributed its event state; and an incompletely observed censored interval was omitted. Twelve-month PCSM cumulative incidence was obtained by summing interval-specific PCSM probabilities weighted by survival through previous intervals. The competing-risk model used the same two-hidden-layer architecture and source/validation chronology as the primary neural model, with 2016 negative-log-likelihood model selection.

6.9 Use of AI-assisted tools

OpenAI ChatGPT was used during manuscript preparation for code review, reproducibility auditing, manuscript organization, figure refinement, and language editing. The corrected statistical analyses reported in this manuscript were executed using the accompanying code against the authorized SEER export, and numerical results were checked against machine-readable aggregate outputs. AI-assisted tools were not listed as authors; the human authors retain responsibility for the manuscript and for verification of the reported results.

Data availability

This study used a prostate-cancer case-listing export from the National Cancer Institute SEER Program. The original patient-level records, patient-level predictions, and fitted checkpoints are not redistributed because access and reuse are governed by SEER data-use requirements. Researchers can request SEER access independently and reproduce the protected stages using an authorized case-listing export containing the variables documented in the Methods and code. The accompanying reproducibility

package contains the corrected executable analysis scripts, environment files, synthetic smoke-test data, and shareable aggregate result tables underlying the values reported in this manuscript. Synthetic data are provided only for software testing and are not the data underlying the clinical results.

Code availability

The corrected analysis code and shareable aggregate outputs used for this Scientific Reports version are supplied in the review archive `ScientificReports_Reproducibility_Code_and_Aggregates.zip`. The archive covers source-model fitting, temporal maintenance, source-model sensitivities, cause-specific IPCW and competing-risk analyses, timing sensitivity, and result verification. Restricted patient-level SEER records are excluded. A permanent public repository identifier can be added upon public release if consistent with the authors' data-use obligations.

Ethics statement

This study is a secondary analysis of de-identified registry data obtained from the Surveillance, Epidemiology, and End Results (SEER) Program under applicable SEER data-use terms. No new participant contact, intervention, treatment assignment, or animal experimentation was performed as part of this study.

Author contributions

S.E. contributed to conceptualization, methodology, supervision, project administration, and writing – review and editing. S.Y. contributed to funding acquisition and writing – review and editing. R.E. contributed to validation and writing – review and editing. M.M. contributed to conceptualization, methodology, software, formal analysis, data curation, visualization, writing – original draft, and writing – review and editing. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Funding

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2604).

Acknowledgements

The authors acknowledge the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program for providing access to the registry data used in this study.

References

- [1] Elshafae, S. M. *et al.* The role of osteomimicry factors in prostate cancer progression and metastasis. *American Journal of Physiology-Cell Physiology* **330**, C605–C618 (2026).
- [2] Schuessler, M., Fleming, S., Meyer, S., Seto, T. *et al.* Diagnostic framework to validate clinical machine learning models locally on temporally stamped data. *Communications Medicine* **5**, 261 (2025).
- [3] Koch, L. M., Baumgartner, C. F. & Berens, P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. *npj Digital Medicine* **7**, 120 (2024).
- [4] Silva, G. F. S., Barcellos Filho, F. N., Wichmann, R. M., Silva Junior, F. C. & Chiavegatto Filho, A. D. P. Strategies for detecting and mitigating dataset shift in machine learning for health predictions: A systematic review. *Journal of Biomedical Informatics* **170**, 104902 (2025).
- [5] Davis, S. E., Walsh, C. G. & Matheny, M. E. Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings. *Frontiers in Digital Health* **4**, 958284 (2022).
- [6] Poddar, M., Marwaha, J. S., Yuan, W. *et al.* An operational guide to translational clinical machine learning in academic medical centers. *npj Digital Medicine* **7**, 129 (2024).
- [7] Finlayson, S. G. *et al.* The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine* **385**, 283–286 (2021).
- [8] Friesacher, H. R. *et al.* Temporal distribution shift in real-world pharmaceutical data: Implications for uncertainty quantification in QSAR models. *Artificial Intelligence in the Life Sciences* **8**, 100132 (2025).
- [9] Van Calster, B. *et al.* A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology* **74**, 167–176 (2016).
- [10] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the achilles heel of predictive analytics. *BMC Medicine* **17**, 230 (2019).
- [11] de Groot, T. M., Ramsey, D., Groot, O. Q. *et al.* Does the SORG machine-learning algorithm for extremity metastases generalize to a contemporary cohort of patients? temporal validation from 2016 to 2020. *Clinical Orthopaedics and Related Research* **481**, 2419–2430 (2023).
- [12] Booth, S., Riley, R. D., Ensor, J., Lambert, P. C. & Rutherford, M. J. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *International Journal of Epidemiology* **49**, 1316–1325 (2020).
- [13] Chi, S. *et al.* A novel lifelong machine learning-based method to eliminate calibration drift in clinical prediction models. *Artificial Intelligence in Medicine* **125**, 102256 (2022).
- [14] Feng, J., Gossmann, A., Sahiner, B. & Pirracchio, R. Bayesian logistic regression for online recalibration and revision of risk prediction models with performance guarantees. *Journal of the American Medical Informatics Association* **29**, 841–852 (2022).

- [15] Moussa, M. Auditing conformal prediction under distribution shift: A detectability boundary, exact label-budget design, and repair. Research Square preprint (2026). Preprint, posted August 7, 2026.
- [16] Collins, G. S. *et al.* TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
- [17] Moons, K. G. M. *et al.* PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025).
- [18] National Cancer Institute. Surveillance, epidemiology, and end results (SEER) program. <https://seer.cancer.gov/>. Accessed September 2026.
- [19] Aalen, O. O. & Johansen, S. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics* **5**, 141–150 (1978).
- [20] Gerds, T. A. & Schumacher, M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* **48**, 1029–1040 (2006).
- [21] Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3 (1950).
- [22] Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378 (2007).
- [23] Austin, P. C. & Steyerberg, E. W. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine* **38**, 4051–4065 (2019).