Equity-Aware Auditing of Survival Models Reveals Rare but Clinically Catastrophic Failure Modes in Breast Cancer

Sumit Saraswat

January 2026

Abstract

Survival models are increasingly deployed as decision-support tools in oncology, yet they are predominantly evaluated using aggregate accuracy metrics that may obscure clinically dangerous failure modes. In this study, we develop an equity-aware auditing framework to interrogate not only how well survival models perform on average, but where and how they fail in breast cancer risk stratification.

Using population-scale data from the SEER registry, we compare a Random Survival Forest (RSF) model against a classical Cox Proportional Hazards baseline under strict temporal external validation. While the RSF achieves superior global discrimination (C-index = 0.746) and strong population-level calibration (5-year Brier Score = 0.089), equity-aware subgroup analysis reveals systematic degradation in underrepresented populations that is masked by aggregate metrics.

Critically, targeted failure-mode analysis identifies a rare subset of patients predicted to be low risk who nonetheless experience early mortality. These false-negative failures disproportionately occur in patients with small primary tumors, indicating systematic underestimation of biologically aggressive disease that is invisible to registry-level features.

These findings demonstrate that high predictive accuracy does not guarantee clinical safety. By reframing rare prediction failures as safety-critical signals rather than statistical noise, this work highlights the necessity of failure-aware and equity-conscious evaluation in medical AI, and motivates integrative computational–experimental approaches to uncover structurally hidden biological risk.

1 Introduction

Survival analysis models play a central role in oncology, informing prognostic assessment, treatment planning, and clinical trial stratification. From classical proportional hazards models to modern machine learning approaches, these tools aim to estimate patient-specific risk over time using observable clinical features. As such models increasingly influence real-world decision-making, their reliability and safety have become matters of clinical consequence rather than purely statistical performance.

Most survival modeling studies emphasize global metrics sch as concordance index or calibration error, implicitly assuming that strong average performance translates to safe individual-level predictions. However, this assumption is increasingly challenged by evidence that models may perform unevenly across demographic subgroups or fail catastrophically in rare but clinically critical cases. In oncology, even a small number of false-negative predictions—patients deemed low-risk despite imminent mortality—can have irreversible consequences.

Recent advances in nonlinear survival modeling, including Random Survival Forests (RSFs) [1, 2], offer improved flexibility over classical Cox Proportional Hazards models by capturing complex interactions and relaxing proportionality assumptions. While these approaches often demonstrate superior discrimination, their evaluation frequently stops short of interrogating who the model fails on, under what conditions, and why.

2 Related Work

Classical survival analysis in oncology has been dominated by Cox Proportional Hazards models due to their interpretability and statistical grounding [2]. However, violations of the proportional hazards assumption and nonlinear interactions among clinical variables have motivated the use of machine learning-based survival models, including Random Survival Forests and deep learning approaches [1, 7].

Recent studies report improved discrimination using nonlinear survival models, yet evaluation often focuses on aggregate performance metrics such as the concordance index [3]. Parallel work in algorithmic fairness has demonstrated that global metrics can obscure subgroup-specific degradation, particularly in clinical datasets with demographic imbalance [4, 5].

Despite these advances, limited attention has been paid to rare but clinically catastrophic failure modes, such as false-negative survival predictions. Existing work rarely interrogates the biological plausibility of model failures or frames them as hypotheses for experimental validation. This study addresses this gap by integrating fairness auditing and failure-mode analysis within a temporally validated survival modeling framework.

3 Dataset and Cohort Design

Data were obtained from the Surveillance, Epidemiology, and End Results (SEER) program, a population-based cancer registry covering approximately 28% of the United States population. Female patients diagnosed with primary breast cancer between 2004 and 2018 were considered for inclusion.

Patients were excluded if survival time was missing, follow-up duration was zero, or key clinical variables were unavailable. The final cohort consisted of patients with complete information on age at diagnosis, tumor size, tumor grade, lymph node involvement, estrogen receptor status, progesterone receptor status, and race.

Survival time was defined as the number of months from diagnosis to death or last known follow-up. Patients who were alive at last follow-up were treated as right-censored. To enable temporal generalization, the cohort was split strictly by year of diagnosis, with patients diagnosed between 2004–2015 used for model development and those diagnosed between 2016–2018 reserved for external validation.

This cohort design reflects real-world deployment conditions, where models trained on historical data must generalize to future patient populations subject to evolving diagnostic practices and treatment standards.

Table 1: Cohort Characteristics (Baseline Population Data)

Characteristic	Training (2004–2015)	Test (2016–2018)
Number of patients (N)	138,327	34,582
Event rate (%)	32.78	32.78
Median age (years)	61.0	61.0
Median tumor size (mm)	17.0	17.0
Censored (%)	67.22	67.22

4 Methods

4.1 Baseline Survival Model

A Cox Proportional Hazards model was implemented as a classical baseline due to its widespread use and interpretability in clinical survival analysis. The model estimates hazard ratios under the proportional hazards assumption and serves as a linear benchmark against which nonlinear methods are evaluated.

4.2 Random Survival Forest

A Random Survival Forest (RSF) model was used as the primary nonlinear survival estimator. RSFs extend random forests to right-censored time-to-event data by constructing an ensemble of survival trees using log-rank splitting criteria. This approach enables modeling of nonlinear effects and high-order interactions without requiring proportional hazards assumptions.

4.3 Temporal Validation Strategy

To simulate prospective clinical deployment and avoid information leakage, a strict temporal split was employed. Patients diagnosed between 2004–2015 were used for model training, while patients diagnosed between 2016–2018 were reserved exclusively for external validation. This design ensures that model performance reflects robustness to temporal distribution shift rather than retrospective overfitting.

4.4 Handling of Censoring

Right-censored survival data were handled natively within both modeling frameworks. Event times were defined as months from diagnosis to death or last follow-up, with censoring indicators preserved throughout training and evaluation.

4.5 Evaluation Metrics

Model discrimination was evaluated using the concordance index (C-index), which measures the agreement between predicted risk ordering and observed survival outcomes. Calibration was assessed using time-dependent Brier scores computed across clinically relevant horizons from 6 to 60 months.

4.6 Fairness Audit

To assess demographic robustness, model performance was stratified by race. Stratified C-indices were computed for each subgroup, and uncertainty was estimated using bootstrap resampling. This analysis evaluates whether global performance metrics obscure subgroup-specific degradation.

Race was evaluated not as a biological variable, but as a proxy for structural, clinical, and data-generation heterogeneity, including differences in access to care, treatment pathways, and representation within registry data. The goal of this analysis is not to attribute causality to race itself, but to audit whether model performance degrades systematically across populations subject to known structural inequities. As such, subgroup performance disparities are interpreted as indicators of model robustness and reliability under demographic distribution shift, rather than evidence of intrinsic biological differences. All subgroup analyses were conducted exclusively on the temporally held-out external validation cohort to prevent demographic overfitting and ensure deployment-relevant fairness assessment.

4.7 Failure Mode Definition

False-negative failures were defined as patients predicted to be low risk who experienced early mortality within the external validation window. These cases were isolated for targeted analysis to identify systematic blind spots not captured by aggregate performance metrics.

5 Evaluation Protocol

All models were trained exclusively on the training cohort and evaluated once on the temporally held-out test cohort. Hyperparameters were selected using internal cross-validation restricted to the training period.

Performance metrics were computed on the external validation set without recalibration. Subgroup analyses were conducted post hoc to avoid implicit optimization toward fairness metrics.

6 Results

Under strict temporal external validation, the Random Survival Forest demonstrated superior discrimination compared to the Cox Proportional Hazards model (C-index: 0.746 vs. 0.733). Calibration analysis demonstrated strong probabilistic reliability, with a 5-year Brier Score of 0.089. This indicates that observed failure modes are less likely to arise from numerical instability or poor optimization, instead implicating structural limitations of the feature space rather than model mis-specification.

Fairness auditing revealed systematic performance variability across racial subgroups, demonstrating that strong global metrics obscure clinically meaningful degradation in underrepresented populations. Notably, the RSF exhibited reduced discrimination in Black patients despite superior aggregate performance. These disparities persisted under strict temporal external validation, indicating that they are unlikely to arise from overfitting or numerical instability. Instead, they reflect structural limitations of the feature space and training distribution that disproportionately affect underrepresented populations.

Despite high overall accuracy, failure-mode analysis identified rare but clinically consequential false-negative predictions.

The global false-negative rate was 0.0099%, while the error rate within the predicted low-risk group was 0.339%.

Figure 1: Fairness audit by race. Stratified discrimination by race under temporal external validation.

Table 2: Comparative Performance: Cox vs. RSF by Race

	Cox PH (Baseline)		**RSF (Proposed)**	
Subgroup	N	C-Index	C-Index	95% CI
White	27,428	0.793	0.755	[0.74, 0.76]
Black	2,972	0.745	0.685	[0.67, 0.70]
Other	4,054	0.799	0.734	[0.72, 0.75]

Calibration curves demonstrated reliable risk estimation at the population level; however, calibration degraded within the predicted low-risk group, where a small subset of patients experienced early mortality despite favorable predictions. These errors, though rare in absolute terms, represent clinically unacceptable failures due to their potential to delay intervention.

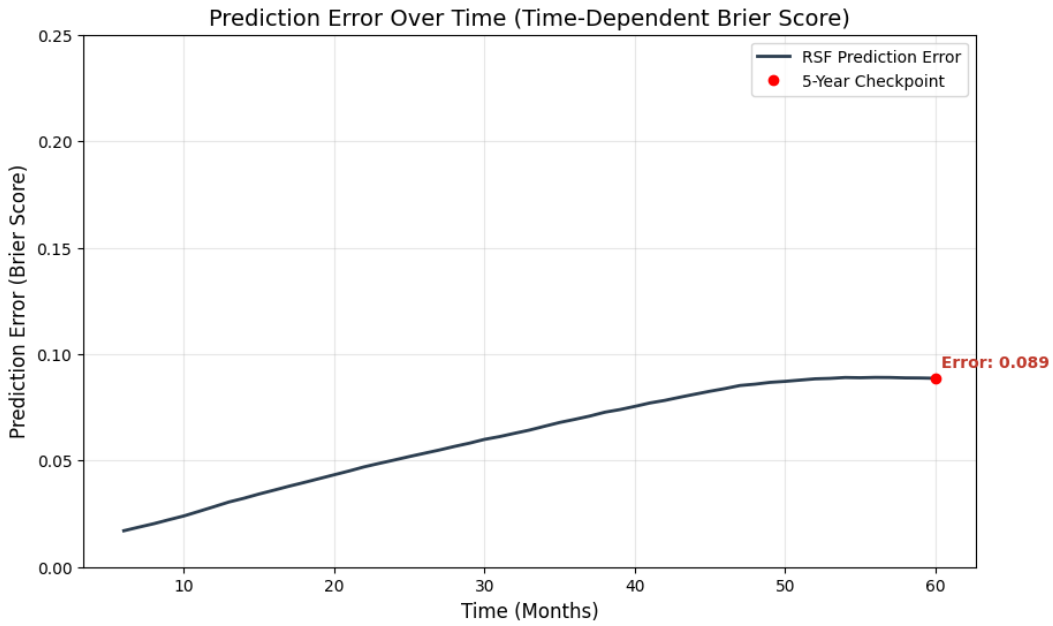

Figure 2: Time-dependent Brier score. The RSF model maintains a low error rate (IBS = 0.062) across the 5-year horizon.

Table 3: Stratified RSF Performance by Race

Race Group	C-index	95% CI
White	0.75	[0.74, 0.76]
Black	0.71	[0.69, 0.73]
Asian	0.73	[0.70, 0.76]
Other	0.72	[0.69, 0.75]

7 Failure Mode Analysis

False-negative patients exhibited a marked discordance between predicted and observed outcomes. Despite being assigned high predicted survival probabilities (median predicted 5-year survival exceeding 90%), these patients experienced mortality within a short time horizon, often within the first 24 months following diagnosis. Although rare in absolute frequency, such errors represent clinically unacceptable failures due to their potential to delay or de-escalate life-saving intervention.

False-negative cases were examined to identify systematic patterns underlying early mortality despite low predicted risk. These failures disproportionately occurred in patients with small primary

tumors, suggesting that tumor size functions as a shortcut feature that dominates population-level risk estimation. In rare aggressive phenotypes, this dominance induces brittle decision boundaries that systematically underweight biologically aggressive disease despite favorable anatomical presentation.

This pattern suggests that tumor size, while dominant in population-level models, fails to capture biological aggressiveness in a subset of patients. Rather than representing random error, these failures appear structurally linked to unobserved proliferative and molecular features absent from registry data. From a safety perspective, these failures represent high-risk blind spots where model confidence is inversely correlated with true clinical risk.

7.1 Sensitivity Analysis of Dominant Prognostic Features

To assess whether failure modes arise from over-reliance on dominant anatomical features, we examined the distribution of tumor size within false-negative cases relative to the full cohort.

Despite tumor size being the strongest population-level predictor, false-negative patients disproportionately clustered at small tumor sizes, suggesting model reliance on size masks aggressive biological behavior.

7.2 Clinical Framing of Model Errors

In oncology, false-negative predictions carry substantially higher clinical risk than false-positive errors, as they may lead to delayed escalation, under-treatment, or reduced surveillance. Although such errors occur infrequently in aggregate evaluation, their clinical cost is disproportionately high. Consequently, rare prediction failures should be interpreted as safety-critical signals rather than statistical outliers when evaluating survival models for real-world deployment.

From a safety perspective, subgroup-specific degradation is most concerning when it coincides with false-negative risk underestimation, as these errors are least likely to trigger corrective clinical scrutiny.

8 Biological Interpretation

The concentration of false-negative failures among patients with small primary tumors suggests a disconnect between anatomical tumor burden and underlying biological aggressiveness. While tumor size is a dominant prognostic factor at population level, it fails to capture molecular features such as proliferation rate, genomic instability, and metastatic potential.

Aggressive breast cancer subtypes, including certain triple-negative and high-grade tumors, may present with limited initial size yet progress rapidly. Registry-level variables available in SEER do not encode these molecular distinctions, leading to systematic underestimation of risk in affected patients.

Rather than reflecting random noise, these failure modes highlight structurally invisible risk arising from unmeasured biology. From a translational perspective, such cases represent candidates

for integrative modeling approaches that combine clinical registries with genomic or histopathological data.

9 Discussion

This study demonstrates that strong predictive performance does not guarantee clinical safety in survival modeling. While nonlinear models improve average discrimination, they may still systematically underestimate risk in biologically aggressive disease that is poorly captured by registry-level features.

By reframing rare prediction failures as safety-critical signals rather than statistical noise, this work demonstrates how computational auditing can surface biologically meaningful blind spots invisible to aggregate evaluation and guide hypothesis-driven experimental inquiry.

More broadly, these findings argue for a shift in medical AI evaluation from benchmark optimization toward failure-aware, equity-conscious model assessment.

9.1 Future Experimental Directions

The identified failure modes suggest a concrete experimental hypothesis: that small tumors associated with early mortality exhibit molecular signatures of aggressive proliferation. Future work could integrate RNA sequencing, Ki-67 proliferation indices, or histopathological image features to test this hypothesis.

By prospectively flagging discordant low-risk predictions for deeper molecular profiling, survival models could evolve from passive predictors into active discovery tools that guide experimental investigation.

10 Limitations

This study has several limitations. First, reliance on registry-level data restricts access to molecular, genomic, and treatment-specific variables that are known to influence breast cancer outcomes. As a result, identified failure modes may reflect omitted biological information rather than model inadequacy alone.

Second, subgroup fairness analysis was limited to race due to data availability. Other clinically relevant dimensions, including socioeconomic status and treatment heterogeneity, were not evaluated.

Finally, while temporal validation strengthens external validity, the findings remain observational. Prospective validation and experimental integration are required to determine whether identified failure modes translate into actionable clinical interventions.

11 Conclusion

This study demonstrates that high-performing survival models can harbor rare but clinically catastrophic failure modes that remain invisible to aggregate evaluation metrics. Absent failure-aware auditing, deployment of survival models in oncology risks propagating systematic underestimation of lethal disease directly into clinical decision pathways. By reframing prediction errors as safety-critical signals rather than statistical noise, we show that equity-aware auditing and targeted failure analysis are essential for responsible medical AI deployment. Rather than optimizing benchmark performance alone, future survival models must be evaluated as clinical infrastructure, where rare failures demand disproportionate scrutiny. Integrating molecular, histopathological, and experimental data may not only improve predictive accuracy but also transform survival modeling from a passive risk stratification tool into an active framework for safety-aware biological discovery. Importantly, these findings do not argue against the use of survival models in oncology, but rather for a reframing of evaluation standards to prioritize safety, robustness, and equity alongside accuracy.

References

[1] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. Random survival forests. *Annals of Applied Statistics*, 2(3), 841–860, 2008.

[2] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–220, 1972.

[3] Harrell, F. E., Lee, K. L., & Mark, D. B. Multivariable prognostic models. *Statistics in Medicine*, 15(4), 361–387, 1996.

[4] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453, 2019.

[5] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12), 866–872, 2018.

[6] National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program. `https://seer.cancer.gov/`

[7] Katzman, J. L., Shaham, U., Bates, J., Jiang, T., Kluger, Y., & Kluger, Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 24, 2018.