IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 April 2026, accepted 26 May 2026, date of publication 1 June 2026, date of current version 4 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3698751

RESEARCH ARTICLE

Regularized Cox Models Versus Deep Survival Networks in Low Events-per-Variable Regimes: A Benchmark Across Censoring Levels

ERSIN YILMAZ
Department of Statistics, Muğla Sıtkı Koçman University, 48000 Muğla, Türkiye
e-mail: ersinyilmaz@mu.edu.tr

ABSTRACT Survival analysis on biomedical data has seen quick methodological growth, but practitioners face limited guidance on which family of methods to prefer when censoring rates are high or the ratio of predictors to events is unsuitable. This paper presents a systematic benchmark of six survival models—the classical Cox model, lasso-penalized Cox, the Bayesian elastic net Cox model (BEN-Cox), random survival forests, DeepSurv, and Cox-Time—across three publicly available datasets (METABRIC, SUPPORT, and TCGA-BRCA) under controlled censoring regimes. The experimental grid scans events-per-variable (EPV) ratios from 0.6 to 7.6 and censoring rates from 33% to 86%. Performance is evaluated using Harrell's concordance index, time-dependent concordance, the integrated Brier score, and the calibration slope, with statistical comparisons carried out by paired Wilcoxon signed-rank tests with Holm correction. Three main findings are obtained. First, BEN-Cox, lasso-Cox, and random survival forests form a narrow top tier in discrimination, with pooled C-index values within 0.003 of each other, while DeepSurv and Cox-Time do not reach this tier in any configuration. Second, calibration separates models that discrimination does not: BEN-Cox achieves the lowest integrated Brier score in all six configurations, while lasso-Cox produces calibration slopes closest to one on the high-dimensional datasets; deep models and the unpenalized Cox model are systematically overconfident. Third, increasing censoring widens the gap between the regularized top tier and the deep architectures without changing which model family is preferred. An empirical partition of the (c, EPV) plane summarizes these observations. Because the entire grid falls below the classical EPV threshold of ten, the partition documents the low-EPV regime in which regularized models dominate but does not identify the boundary at which deep models may become competitive.

INDEX TERMS Bayesian elastic net, censoring, Cox proportional hazards model, deep learning, events per variable, survival analysis.

I. INTRODUCTION

Survival analysis remains one of the main statistical tools for modeling time-to-event outcomes in biomedical research, where censoring is intrinsic to the data structure and absolute risk estimation is often as important as discrimination. Since the seminal work of Cox [1], the proportional hazards model has occupied a central place in this literature because it combines a semiparametric formulation with relatively clear interpretation and practical estimation through the partial likelihood. In many applied areas, including oncology and biomarker-driven clinical studies, it continues to serve as the main reference model against which newer approaches are evaluated.

However, the methodological landscape of survival modeling has expanded substantially over the last two decades.

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Bellan.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

As biomedical datasets became larger, more heterogeneous, and often higher dimensional, penalized Cox extensions such as the lasso and elastic net became standard tools for handling settings in which the number of candidate predictors is not negligible relative to the amount of event information [2], [3]. In parallel, Bayesian regularized Cox formulations were developed to place shrinkage directly in a probabilistic framework, thereby allowing posterior uncertainty quantification together with regularized estimation [4], [5]. More recently, Bayesian elastic-net type Cox formulations have also been revisited in benchmark-oriented biomedical settings, where they have shown competitive predictive performance and promising calibration behavior in difficult regimes [6]. At the same time, machine learning and deep learning approaches, including random survival forests, DeepSurv, Cox-Time, and related neural survival models, have been proposed to capture nonlinear effects, interactions, and possible departures from proportional hazards [7], [8], [9], [10]. Taken together, these developments have created a rich but practically difficult model space for applied researchers.

Despite this methodological progress, one practical question remains insufficiently clarified: under which data conditions should one prefer a classical or regularized Cox model over a more flexible machine-learning or deep-learning alternative? Existing benchmark studies have improved the empirical picture [11], [12], [13], but many of them summarize performance across datasets and methods in a way that does not isolate the specific roles of censoring and effective event information which can be interpreted as an important omission. In biomedical survival data, censoring may be substantial, and the number of observed events may be small relative to the predictor dimension. Under such conditions, model flexibility alone is not necessarily an advantage, because the effective amount of information available for fitting and evaluation becomes limited.

This issue is closely related to the classical events-per-variable (EPV) discussion. Earlier work showed that low EPV may lead to instability, bias, and weak precision in Cox regression, while later studies made clear that the well-known EPV heuristics should not be interpreted as universal fixed thresholds [14], [15], [16]. Even so, the interaction between EPV and censoring has not been systematically mapped for modern survival model families, especially when Bayesian shrinkage methods and deep survival models are compared within the same benchmark. The literature also suggests that calibration should not be treated as secondary in this setting. A model may rank individuals reasonably well while still producing risk estimates that are not sufficiently reliable for practical use. For this reason, a comparison that considers discrimination and calibration together is more informative than one based only on ranking performance.

The present study is motivated by this gap. We consider a benchmark framework that compares six representative survival models drawn from the main methodological families currently used in practice: the classical Cox proportional hazards model, lasso-penalized Cox regression, a Bayesian elastic net Cox model (BEN-Cox) [6], random survival forests, DeepSurv, and Cox-Time. The comparison is conducted on three publicly available biomedical datasets, namely METABRIC, SUPPORT, and TCGA-BRCA, chosen to reflect different sample sizes, predictor structures, and baseline censoring levels [17], [18], [19]. Rather than comparing methods only at their native censoring levels, we evaluate them under controlled censoring regimes so that the effect of increasing information loss can be examined more directly. We then assess performance using both discrimination-oriented and calibration-sensitive measures, with particular attention to how model ranking changes as a function of censoring and events per variable.

The contribution of this paper is primarily empirical and comparative. First, it provides a focused benchmark in which Bayesian regularized Cox models and deep survival models are examined within the same evaluation framework, which is still less common than comparisons confined either to penalized Cox variants or to machine-learning methods alone. Second, it studies censoring and events per variable jointly, rather than treating them as background characteristics. The experimental grid spans EPV values from approximately 0.6 to 7.6 and censoring rates from 33% to 86%, a range that falls entirely below the classical EPV $=$ 10 guideline and is representative of many applied biomedical survival settings, particularly in oncology and molecular cohorts. Third, it summarizes the observed model behavior over the (c, EPV) space in a descriptive and interpretable way, documenting that regularized models—both Bayesian and frequentist—form a competitive top tier throughout the low-EPV regime covered here, while deep survival architectures do not reach that tier in any configuration of the present grid. We emphasize that this summary should be viewed as an empirical characterization of the regimes observed in the present benchmark, not as a universal conclusion about deep survival models, whose advantage may materialize at higher EPV values not covered by the present experimental design. In this sense, the paper offers an empirical complement to the classical EPV discussion by making censoring an explicit part of the comparison.

The remainder of the paper is organized as follows. Section II briefly reviews the methodological streams most relevant to the benchmark and clarifies the specific gap addressed here. Section III introduces the survival setting and the models under comparison. Section IV describes the datasets, censoring protocol, preprocessing, and evaluation strategy. Section V reports the empirical findings. Section VI describes the empirical partition of the (c, EPV) plane. Section VII discusses the interpretation, limitations, and implications for practical model choice. Section VIII concludes the paper.

II. RELATED WORK

The literature most relevant to the present study can be grouped into four closely connected streams: penalized Cox

regression, Bayesian regularized Cox modeling, machine-learning and deep-learning survival methods, and recent benchmark studies.

Penalized Cox regression remains the most established extension of the classical proportional hazards model in high-dimensional settings. The lasso formulation of Tibshirani [2] introduced sparsity into Cox regression, and later elastic-net implementations made regularized Cox fitting computationally practical at scale [3]. These methods remain strong baselines because they preserve the Cox structure while stabilizing estimation when the amount of event information is limited. In applied work, they are often preferred for their combination of interpretability, mature software, and relatively stable optimization.

A second strand concerns Bayesian regularized Cox models. In this line of work, shrinkage is imposed through prior distributions rather than only through penalized optimization, which allows regularization and uncertainty quantification to be handled within the same inferential framework. This literature includes Bayesian variable-selection and Bayesian elastic-net type constructions for semiparametric proportional hazards models [4], [5]. From the perspective of the present paper, this direction is important because it provides a principled alternative to purely optimization-based regularization, especially in regimes where event information is limited and calibration may be sensitive to overfitting. It is also directly relevant here because BEN-Cox-type formulations have recently been considered in biomedical benchmark settings and have shown that Bayesian shrinkage may remain competitive even when more flexible survival learners are included in the comparison [6]. At the same time, Bayesian Cox variants are still represented less often in broad empirical benchmarks, partly because of their computational cost and their less standardized software ecosystem.

The third strand is the machine-learning and deep-learning survival literature. Random survival forests provided an early nonparametric alternative to Cox-type models by relaxing linearity and proportional-hazards structure while naturally accommodating interactions and nonlinear effects [7]. Later, neural survival models expanded in two main directions. The first direction retains the Cox partial likelihood while replacing the linear predictor with a nonlinear risk function, as in DeepSurv [8]. The second direction moves further away from the proportional hazards structure, either by allowing time-dependent effects, as in Cox-Time [9], or by modeling the event-time distribution more directly in discrete time, as in DeepHit [10]. These methods have shown promising results, especially in larger and structurally richer datasets, but their practical advantage is not uniform across data regimes.

Recent reviews and benchmark studies help clarify this point. Herrmann et al. [11] showed in a large-scale multi-omics benchmark that more complex methods do not automatically dominate Cox-based baselines. The conclusions also depend on which metric is emphasized and on the structure of the underlying dataset. More recent review work has similarly noted that deep survival learning is expanding rapidly, but that reproducibility, task coverage, and careful regime-specific comparison remain open issues [12]. These observations are important for the present study because they suggest that the practical value of a method cannot be judged only by its flexibility or novelty.

Against this background, the specific gap addressed here is narrower and more targeted than a general ''which method is best'' benchmark. Our interest is in how the relative performance of Bayesian regularized Cox models and deep survival models changes when censoring and effective event information are considered jointly. Existing benchmarks contain important pieces of this picture, but they do not usually isolate censoring through controlled regimes while simultaneously interpreting model behavior through the EPV perspective. The present study is designed to address that narrower question, while keeping both discrimination and calibration visible in the comparison.

III. METHODS

This section formalizes the survival setup used throughout the benchmark, describes the six models under comparison, defines the effective covariate dimension and the role of the proportional hazards assumption, and lists the evaluation metrics.

A. SURVIVAL SETUP AND NOTATION

Let $(T_i, C_i, \mathbf{x}_i)$ denote independent observations for subjects $i = 1, \ldots, n$, where $T_i \in \mathbb{R}^+$ is the true event time, $C_i \in \mathbb{R}^+$ is the right-censoring time, and $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of baseline covariates. We observe

$$\widetilde{T}_i = \min(T_i, C_i), \qquad \delta_i = \mathbf{1}\{T_i \le C_i\},$$

that is, the minimum of the event and censoring times together with the event indicator. The survival and hazard functions conditional on covariates are

$$S(t \mid \mathbf{x}) = \Pr(T > t \mid \mathbf{x}), \qquad h(t \mid \mathbf{x}) = -\frac{d}{dt} \log S(t \mid \mathbf{x}).$$

Throughout the paper we assume non-informative right-censoring, $T_i \perp C_i \mid \mathbf{x}_i$. This assumption is preserved by construction under the censoring augmentation protocol of Section IV-C, since the auxiliary censoring times are drawn independently of T_i and $\mathbf{x}_i$.

The Cox proportional hazards model [1] specifies

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\beta^\top \mathbf{x}), \tag{1}$$

where $h_0(t)$ is an unspecified baseline hazard and $\beta \in \mathbb{R}^p$ is a vector of log hazard ratios. Estimation of β is based on the partial likelihood

$$L(\beta) = \prod_{i:\, \delta_i = 1} \frac{\exp(\beta^\top \mathbf{x}_i)}{\sum_{j \in \mathcal{R}(\widetilde{T}_i)} \exp(\beta^\top \mathbf{x}_j)}, \tag{2}$$

where $\mathcal{R}(t) = \{j : \widetilde{T}_j \ge t\}$ is the risk set at time t. All models considered in this paper either maximize, or approximate, a penalized or regularized version of (2), or replace the linear predictor $\beta^\top \mathbf{x}$ with a nonlinear function $f(\mathbf{x})$.

B. MODELS COMPARED

Six models are included in the benchmark. The selection is intended to span the methodological families currently in active use through mature software, rather than to enumerate every published variant. For each model we record whether it retains the proportional hazards structure, since this distinction interacts with the interpretation of the results reported in Section V and with the proportional hazards diagnostics described below.

1) COX PROPORTIONAL HAZARDS MODEL

The unpenalized model (1) serves as the reference baseline. When p/n is non-trivial or the number of observed events is small, this estimator is known to suffer from instability and inflated coefficient magnitudes, and its behavior in such regimes is itself informative for the comparison. The proportional hazards structure is retained.

2) LASSO-PENALIZED COX MODEL

The lasso-Cox model [2] estimates β by maximizing the penalized log partial likelihood

$$\ell(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

where $\ell(\beta) = \log L(\beta)$ and $\lambda \geq 0$ is a tuning parameter selected by inner cross-validation. We use the coordinate descent implementation of [3] available in the `glmnet` package. The proportional hazards structure is retained.

3) BAYESIAN ELASTIC NET COX MODEL (BEN-COX)

BEN-Cox places a hierarchical global-local shrinkage prior on β, constructed so that the marginal prior density corresponds to the elastic net penalty in the sense of [20]. Let $\lambda_1, \lambda_2 > 0$ denote the global shrinkage hyperparameters, and let $z_j > 0, j = 1, \ldots, p$, be latent scale variables. The prior is specified hierarchically as

$$\beta_j \mid z_j, \lambda_2 \sim \mathcal{N}\Big(0,\ (1/z_j + \lambda_2)^{-1}\Big), \qquad z_j \sim \mathrm{Exp}(\lambda_1^2/2),$$

with half-Cauchy hyperpriors on λ_1 and λ_2. In this parameterization, λ_2 controls the quadratic shrinkage component shared across coefficients, while the latent scale z_j induces coefficient-specific adaptive shrinkage in the spirit of the Bayesian lasso. The posterior of β given the observed data is obtained by combining this prior with the partial likelihood (2), and inference is performed using Hamiltonian Monte Carlo. The construction is designed to retain the grouping behavior associated with the frequentist elastic net at the posterior mode, while also yielding full posterior distributions for hazard ratios and patient-level survival curves. The proportional hazards structure is retained. A complete description of the model and its computational implementation is given in [6].

4) RANDOM SURVIVAL FORESTS (RSF)

Random survival forests [7] extend the random forest paradigm to censored outcomes by using a log-rank splitting rule and estimating the cumulative hazard within terminal nodes via the Nelson-Aalen estimator. The method is fully nonparametric and makes no proportional hazards assumption, and it naturally captures nonlinear and interaction effects at the cost of less direct interpretability of individual covariate contributions. Tuning parameters include the number of trees, the number of variables sampled per split (`mtry`), and the minimum terminal node size.

5) DeepSurv

DeepSurv [8] replaces the linear predictor in (1) with a feedforward neural network $f_\theta(\mathbf{x})$, yielding the hazard model

$$h(t \mid \mathbf{x}) = h_0(t) \exp\big(f_\theta(\mathbf{x})\big).$$

The network parameters θ are estimated by minimizing the negative log partial likelihood with standard regularization techniques, including dropout and weight decay. Since the time dependence remains confined to the baseline hazard, the proportional hazards structure is retained.

6) COX-TIME

Cox-Time [9] generalizes DeepSurv by letting the relative risk depend on both covariates and time through a network $g_\theta(t, \mathbf{x})$, so that

$$h(t \mid \mathbf{x}) = h_0(t) \exp\big(g_\theta(t, \mathbf{x})\big).$$

This relaxation allows non-proportional hazards and has been reported to improve predictive performance when the proportionality assumption is substantively violated, at the cost of additional computational effort and less interpretable individual effect estimates. Of the six models, Cox-Time is the only one that does not assume proportional hazards.

C. EFFECTIVE DIMENSION AND THE PROPORTIONAL HAZARDS ASSUMPTION

Two structural quantities recur in the interpretation of the results and deserve to be defined explicitly at the methods stage, since both bear directly on how the events-per-variable ratio of Section IV-E should be read.

1) EFFECTIVE COVARIATE DIMENSION

Although the raw predictor count p is fixed once preprocessing has been carried out (see Section IV-B), the *effective* covariate dimension actually used by each model can be substantially smaller, in particular after regularization. We denote this quantity by p_{eff} and define it on a model-specific basis. For lasso-Cox we take p_{eff} to be the number of nonzero coefficients at the cross-validated penalty λ, which corresponds to the standard lasso degrees-of-freedom estimate [21]. For BEN-Cox we use the posterior mean of the

implied number of unshrunk coefficients,

$$p_{\text{eff}} = \mathbb{E}\left[\sum_{j=1}^{p}(1-\kappa_j)\,\middle|\,\text{data}\right],$$

where $\kappa_j \in (0,1)$ denotes the posterior shrinkage factor for coefficient j, with values close to one corresponding to strong shrinkage toward zero and values close to zero indicating that the coefficient is left essentially unshrunk. This construction follows the Bayesian analog of model size used in related shrinkage formulations [22] and provides a comparable counterpart to the lasso degrees-of-freedom estimate. For the unpenalized Cox model, RSF, DeepSurv, and Cox-Time, an effective dimension is less straightforward to define, since none of these models applies an explicit shrinkage operator on the covariate representation in the same sense as the penalized Cox variants. For these models we therefore set $p_{\text{eff}} = p$ as a conservative proxy rather than a formal degrees-of-freedom estimate, which has the consequence that any EPV-based comparison should be read as nominally favorable to shrinkage models; this asymmetry is acknowledged again as a limitation in Section VII. The events-per-variable ratio reported throughout the paper, defined formally in Section IV-E as Equation (5), is computed using p_{eff} rather than the raw p.

2) PROPORTIONAL HAZARDS

Five of the six models assume proportional hazards; only Cox-Time relaxes this assumption. To document the extent to which the assumption is consistent with the data, the score test of Grambsch and Therneau [23] is applied to the unpenalized Cox model fitted on each training fold of each dataset-regime combination, and the global p-values are summarized alongside the predictive results in Section V. The tests are not used to gate the inclusion of any model, since the benchmark is descriptive in character. Their role is to indicate whether the relative performance of Cox-Time in a given regime can plausibly be related to departures from proportional hazards rather than to model flexibility alone, and to provide context for the interpretation of the empirical partition described in Section VI.

D. EVALUATION METRICS

Predictive performance is assessed along two complementary dimensions, discrimination and calibration, using four metrics that are standard in the survival literature. Consistent with the position taken in Section I, the two dimensions are treated as equally important: ranking quality and absolute risk quality jointly determine the practical value of a survival model, and a comparison restricted to either dimension alone would obscure precisely the trade-offs that the present benchmark is designed to expose.

1) HARRELL'S CONCORDANCE INDEX

The concordance index [24] estimates the probability that, for a randomly chosen pair of comparable subjects, the subject with the higher predicted risk experiences the event first. Values range from 0.5 (no discrimination) to 1 (perfect discrimination) and are insensitive to monotone transformations of the risk score.

2) TIME-DEPENDENT CONCORDANCE

As a complement to Harrell's statistic, we report the inverse-probability-of-censoring-weighted concordance measure of [25], which mitigates the dependence of Harrell's index on the censoring distribution. This adjustment is particularly relevant in the high-censoring regimes studied here.

3) INTEGRATED BRIER SCORE

The Brier score at time t is the expected squared error between the predicted survival probability $\widehat{S}(t \mid \mathbf{x})$ and the true event status, and can be estimated consistently in the presence of censoring using inverse-probability-of-censoring weights [26]. Integrating over a clinically meaningful time horizon yields the integrated Brier score (IBS), a proper scoring rule that rewards both discrimination and calibration. Lower values indicate better performance.

4) CALIBRATION SLOPE

To assess calibration directly, we regress the observed event indicator on the log cumulative hazard implied by each model at a pre-specified evaluation horizon defined consistently within each dataset-regime combination, and report the resulting slope. The horizon is fixed in advance and shared across models so that the calibration comparison within a given regime is not affected by model-specific choices of evaluation time. A slope of one indicates perfect calibration; slopes below one indicate overconfident risk predictions, a common failure mode in small-event regimes.

5) STATISTICAL COMPARISON

Differences between models are assessed using paired Wilcoxon signed-rank tests across cross-validation folds, with Holm correction for multiple comparisons. In addition to p-values, we report median differences and rank-based effect sizes to facilitate substantive interpretation, in line with the recommendations of recent benchmark studies in the survival literature [11].

IV. EXPERIMENTAL DESIGN

This section describes the datasets, preprocessing steps, censoring augmentation protocol, cross-validation scheme, hyperparameter tuning strategy, and computational environment used in the benchmark. The design choices are intended to ensure that the six models of Section III-B are compared under fair and consistent conditions, so that the empirical contrasts reported in Section V reflect differences in model behavior rather than asymmetries in tuning or preprocessing.

A. DATASETS

Three publicly available biomedical datasets are used, selected to span a range of sample sizes, covariate

dimensionalities, and baseline censoring proportions. Together, they provide a useful range of settings for the joint (c, EPV) analysis from large clinical cohorts with moderate censoring to smaller high-dimensional cohorts with very high censoring.

1) METABRIC

The Molecular Taxonomy of Breast Cancer International Consortium cohort [17] contains $n = 1{,}903$ patients with breast cancer, for whom clinical variables and gene expression measurements are available. After an unsupervised low-variance filter, $p = 440$ gene expression features are retained. The primary endpoint is overall survival, with baseline censoring of approximately 58%.

2) SUPPORT

The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments [18] is used here in its publicly distributed subset of $n = 1{,}000$ seriously ill hospitalized adults, as commonly adopted in recent survival benchmarks [8], [9]. This subset is used to maintain consistency with prior methodological comparisons and to keep the analysis within the low-EPV regime that is the focus of the present study.

3) TCGA-BRCA

The Breast Invasive Carcinoma cohort of The Cancer Genome Atlas [19] contributes $n = 1{,}071$ patients with clinical annotations. The present benchmark uses the clinical covariate block ($p = 27$ raw features, expanded to $p = 91$ after one-hot encoding of categorical covariates); molecular features are not included. The baseline censoring proportion is approximately 86%, representative of early-stage oncology cohorts in which follow-up is limited relative to the typical time to event.

Taken together with the censoring augmentation protocol described in Section IV-C, the three datasets contribute six distinct (dataset, regime) combinations to the experimental grid, which is the set of combinations over which Sections V and VI report results.

B. PREPROCESSING

Continuous covariates are standardized to zero mean and unit variance using statistics computed on the training folds only. Categorical covariates are one-hot encoded. Missing values in continuous variables are imputed with the training-fold median and in categorical variables with the training-fold mode; this simple strategy is used consistently across all models in order to isolate model-level effects from imputation-level effects. For TCGA-BRCA, the present benchmark uses the clinical covariate block only; no molecular feature block is included, and no supervised or unsupervised dimensionality-reduction step is applied beyond the encoding of categorical covariates. For METABRIC, an unsupervised low-variance filter is applied to the gene expression block before modeling, retaining the 440 most variable genes. This unsupervised reduction is applied only to make the molecular feature space computationally comparable to the rest of the benchmark setting and does not use outcome information; the other two datasets are left at their native dimensionalities. No further supervised feature selection is performed outside each model's own regularization, so that the effective dimension p_{eff} introduced in Section III-C reflects only what each model itself shrinks away rather than an externally imposed preselection.

C. CENSORING AUGMENTATION PROTOCOL

To study performance under controlled censoring regimes, we augment each dataset with additional independent non-informative right-censoring. Recall from Section III-A that the observed data are $(\widetilde{T}_i, \delta_i, \mathbf{x}_i)$, with $\widetilde{T}_i = \min(T_i, C_i)$ and $\delta_i = \mathbf{1}\{T_i \le C_i\}$. For each target censoring rate $c \in \{0.30, 0.50, 0.70\}$, we draw auxiliary censoring times $C_i^\star \sim \text{Uniform}(0, \tau_c)$ independently of T_i and $\mathbf{x}_i$, and define the augmented observations

$$\widetilde{T}_i^{(c)} = \min(\widetilde{T}_i, C_i^\star), \qquad \delta_i^{(c)} = \delta_i \cdot \mathbf{1}\{\widetilde{T}_i \le C_i^\star\}. \tag{4}$$

The threshold τ_c is chosen by bisection so that the empirical censoring rate of $\{\delta_i^{(c)}\}_{i=1}^n$ matches the target c up to a tolerance of 0.01. This construction augments censoring without modifying any observed event time, and follows the general framework of [27] for simulation studies with predefined censoring rates.

Two properties of the protocol deserve emphasis. First, the procedure can only *increase* the censoring proportion relative to the baseline. Consequently, for TCGA-BRCA, whose baseline censoring of approximately 0.86 exceeds all three target rates, only the baseline regime is considered. For METABRIC, whose baseline censoring of approximately 0.58 exceeds the 0.30 and 0.50 targets, only the baseline and $c = 0.70$ regimes are evaluated. For SUPPORT, whose baseline censoring of approximately 0.33 is below 0.50, the baseline, $c = 0.50$, and $c = 0.70$ regimes are evaluated. The $c = 0.30$ target is never binding across the three datasets. This asymmetry restricts the number of (c, EPV) combinations available for the partition analysis of Section VI and is acknowledged again as a limitation in Section VII. Second, non-informative censoring is preserved by construction since $C_i^\star$ is independent of both T_i and $\mathbf{x}_i$. The conclusions drawn from the augmented data should therefore be read as applying to the non-informative censoring case, and any extension to informative censoring would require either a different augmentation mechanism or datasets in which informative censoring is explicitly documented.

D. CROSS-VALIDATION AND TUNING

Predictive performance is estimated using 5×2 repeated cross-validation, that is, five independent repetitions of two-fold cross-validation with different random seeds. This scheme provides ten train-test pairs per model-dataset-regime combination while reducing the risk of the optimistic bias

that can arise with conventional k-fold cross-validation in small-event settings [11]. Within each training fold, hyperparameters are selected by inner five-fold cross-validation on the training portion only, so that no information from the outer test fold leaks into model selection.

The tuning procedures are kept deliberately modest and approximately matched in computational budget across methods, so that no single family is favored by a substantially more aggressive search. For lasso-Cox, the penalty parameter λ is searched over a log-spaced grid, with the final value selected by inner-fold partial-likelihood deviance. For BEN-Cox, the global shrinkage hyperparameters λ_1 and λ_2 introduced in Section III-B are equipped with their half-Cauchy hyperpriors and are sampled jointly with β rather than tuned by grid search; posterior inference uses two Hamiltonian Monte Carlo chains of 2,000 iterations each, discarding the first half as warm-up, with convergence assessed by the potential scale reduction factor $\widehat{R}$ and effective sample size diagnostics. For random survival forests, `mtry` and the minimum terminal node size are searched over small grids with the number of trees fixed at 500. For DeepSurv and Cox-Time, a random search of fixed size is performed over the learning rate, dropout rate, hidden layer width, and number of hidden layers, using early stopping on the inner validation fold to limit overfitting. The size of the random search is kept comparable, in wall-clock terms, to the inner cross-validation cost of lasso-Cox so that the deep models do not benefit from a substantially larger effective tuning budget.

This budget-matched design is a deliberate choice. It may understate the performance of the deep survival models, which are known to benefit from more extensive architecture search, but it arguably better reflects the conditions under which applied researchers typically operate, and it protects against the optimistic bias that arises when one family of models is tuned more aggressively than another. The implications of this choice are revisited in Section VII as a threat to validity.

E. EFFECTIVE DIMENSION AND EVENTS-PER-VARIABLE RATIO

For each dataset-regime combination and each model, we report the events-per-variable ratio

$$\mathrm{EPV} = \frac{\sum_{i=1}^{n} \delta_i^{(c)}}{p_{\mathrm{eff}}}, \tag{5}$$

where $\delta_i^{(c)}$ is the augmented event indicator defined in Equation (4) and p_{eff} is the effective covariate dimension introduced in Section III-C. As discussed there, p_{eff} is computed as the number of nonzero coefficients at the cross-validated penalty for lasso-Cox, as the posterior expected number of unshrunk coefficients for BEN-Cox, and as the raw post-preprocessing predictor count p for the unpenalized Cox model, RSF, DeepSurv, and Cox-Time. Equation (5) therefore yields a model-specific EPV that is strictly equal to or larger than the raw events-to-p ratio for the regularized Cox variants, and equal to it for the remaining four models.

To make this asymmetry transparent, both the model-specific EPV based on p_{eff} and the raw events-to-p ratio are reported alongside the discrimination and calibration results in Section V. The empirical partition described in Section VI is constructed using the raw events-to-p ratio, and the implications of this choice are discussed in Section VII-D.

F. REPRODUCIBILITY AND COMPUTATIONAL ENVIRONMENT

All experiments are conducted with fixed random seeds at each level of the cross-validation hierarchy to support close computational reproducibility. Classical and penalized Cox models are fit using `glmnet`; random survival forests are fit using `randomForestSRC`; DeepSurv and Cox-Time are fit using `pycox`; the Grambsch-Therneau score test of Section III-C is computed using `survival::cox.zph`; and BEN-Cox is fit using a Stan implementation developed by the author. All computations are performed on a single workstation with standard specifications. Complete source code, package versions, and seeds are made available in a public repository accompanying this paper to facilitate independent replication.

V. RESULTS

This section reports the empirical performance of the six models described in Section III-B across the datasets and censoring regimes defined in Section IV. Results are organized in four parts: overall performance aggregated across datasets, the effect of the censoring rate c on discrimination and calibration, the interaction between the events-per-variable ratio (Eq. 5) and model ranking, and computational cost. Unless otherwise stated, reported values are medians over the 5×2 cross-validation folds, with interquartile ranges in parentheses.

A. OVERALL PERFORMANCE

Table 1 reports the four evaluation metrics of Section III-D pooled across the six dataset–regime combinations of the experimental grid. Three performance groups are visible.

In discrimination, BEN-Cox, lasso-Cox, and random survival forests form a narrow top tier, with pooled Harrell's C-index between 0.723 and 0.726 and pooled Uno's C-statistic between 0.722 and 0.724. The three models are not separated at the pooled level: their medians lie within a range smaller than the typical cross-fold interquartile range of any single combination. The unpenalized Cox model and the two deep survival architectures sit roughly 0.02 below this top tier on both concordance measures. The gap is largest on datasets in which the predictor dimension is comparable to the number of events: on METABRIC ($p = 440$) the unpenalized Cox model drops to $C \approx 0.56$, consistent with the known instability of the partial likelihood estimator in that regime.

In integrated Brier score, BEN-Cox is first at the pooled level, with an IBS of 0.146 against 0.150 for lasso-Cox and 0.153 for RSF. The separation between the regularized linear models and the deep architectures is larger here than for discrimination: DeepSurv, Cox-Time, and the unpenalized Cox model cluster around 0.165–0.169, and the unpenalized Cox model shows a pronounced IBS penalty on high-dimensional combinations (for example 0.32 on METABRIC at baseline censoring).

The calibration-slope picture is partly distinct from the IBS picture, and the distinction is informative. Lasso-Cox attains the slope closest to one at the pooled level (1.09), followed by RSF (1.02). BEN-Cox is slightly overconfident at the pooled level with a slope of 0.87. DeepSurv, Cox-Time, and the unpenalized Cox model are substantially overconfident, with pooled slopes near 0.5 and, for the unpenalized Cox on high-dimensional configurations, collapsing to values close to zero. The apparent contrast between BEN-Cox's IBS advantage and its sub-unit calibration slope reflects the two quantities measuring different aspects of the predictive distribution: the IBS rewards both sharpness and calibration jointly, while the calibration slope isolates the width of the predicted risk spread.

Paired Wilcoxon signed-rank tests with Holm correction, applied within each (dataset, regime, metric) stratum as described in Section III-D, yield 15 paired comparisons per stratum and 90 comparisons per metric across the six configurations. After Holm correction, 64 of 90 C-index comparisons, 65 of 90 Uno C comparisons, 66 of 90 IBS comparisons, and 82 of 90 calibration-slope comparisons remain significant at the 0.05 level. The statistical tests confirm the grouping visible in Table 1: the ordering within the top discrimination tier (BEN-Cox, lasso-Cox, RSF) is not resolved in a majority of strata, while the separation of those three from the deep architectures and from the unpenalized Cox model is resolved in a majority of strata, and the calibration-slope separation is sharpest of all.

TABLE 1. Overall predictive performance pooled across the six dataset–regime combinations of the experimental grid. Values are medians over the 5 × 2 cross-validation folds within each combination, then pooled across combinations. Higher is better for C-index and Uno's C-statistic; lower is better for IBS; closer to one is better for the calibration slope. Training time is the median wall-clock time per fold.

Model	C-index	Uno C	IBS	Calib. slope	Time (s)
Cox	0.710	0.688	0.166	0.51	0.07
Lasso-Cox	0.724	0.722	0.150	1.09	1.2
BEN-Cox	0.726	0.722	0.146	0.87	114.0
RSF	0.723	0.724	0.153	1.02	22.5
DeepSurv	0.706	0.698	0.169	0.44	15.1
Cox-Time	0.702	0.682	0.165	0.50	6.2

B. EFFECT OF THE CENSORING RATE

The censoring augmentation protocol of Section IV-C produces two datasets with multiple regimes: SUPPORT (baseline $c \approx 0.33$, $c = 0.50$, $c = 0.70$) and METABRIC (baseline $c \approx 0.58$, $c = 0.70$). TCGA-BRCA enters the grid only at its baseline censoring of $c \approx 0.86$, as discussed in Section IV-C.

A methodological remark is necessary before model-level patterns are described. Because the augmentation protocol of Eq. (4) is applied to the full dataset prior to the cross-validation split, both the training and the test folds carry the augmented censoring structure. The set of comparable pairs entering the concordance computation therefore changes with c: at higher censoring, fewer events are observed, and the evaluation is restricted to a different subset of the original follow-up. Consequently, absolute values of the C-index and the IBS should not be compared across censoring regimes; the meaningful comparison is the relative ranking of models within each regime.

1) DISCRIMINATION

Table 2 and Figure 1 report median Harrell's C-index and interquartile range for each model within each dataset–regime combination. Three observations emerge.

First, the relative ordering of the six models is broadly stable across censoring regimes. BEN-Cox, lasso-Cox, and RSF occupy the top three ranks in every configuration except SUPPORT at $c = 0.70$, where RSF takes the lead by a margin of 0.005 over BEN-Cox. DeepSurv and Cox-Time occupy the lower tier in every configuration, regardless of censoring level.

Second, the gap between the top tier and the deep architectures does not narrow as censoring decreases. On SUPPORT at baseline ($c \approx 0.33$), the lowest censoring level in the grid, the gap between BEN-Cox and DeepSurv is 0.021 in C-index; at $c = 0.70$ the corresponding gap between RSF and DeepSurv is 0.028. If deep architectures benefited primarily from larger effective sample sizes at low censoring, the gap should close in low-c configurations; the data do not show this pattern in the EPV range covered by the experimental grid.

Third, the unpenalized Cox model is a consistent outlier on METABRIC, with $C \approx 0.56$ at both censoring levels, roughly 0.12 below the regularized models. On SUPPORT, where the post-preprocessing predictor count ($p = 44$) is small relative to the event count, the unpenalized Cox model is competitive with the top tier, confirming that its poor performance on METABRIC is driven by the high-dimensional covariate structure rather than by censoring alone.

2) INTEGRATED BRIER SCORE

Table 3 and Figure 2 report median IBS per configuration. BEN-Cox achieves the lowest IBS in all six configurations of the experimental grid. The margin over lasso-Cox is small (typically 0.002–0.004) but consistent, and the separation from the deep architectures is larger (typically 0.015–0.030). On METABRIC, the unpenalized Cox model produces IBS values approximately double those of the regularized linear models (for example 0.322 versus 0.157 at baseline), an expected consequence of variance inflation when p approaches the number of events.

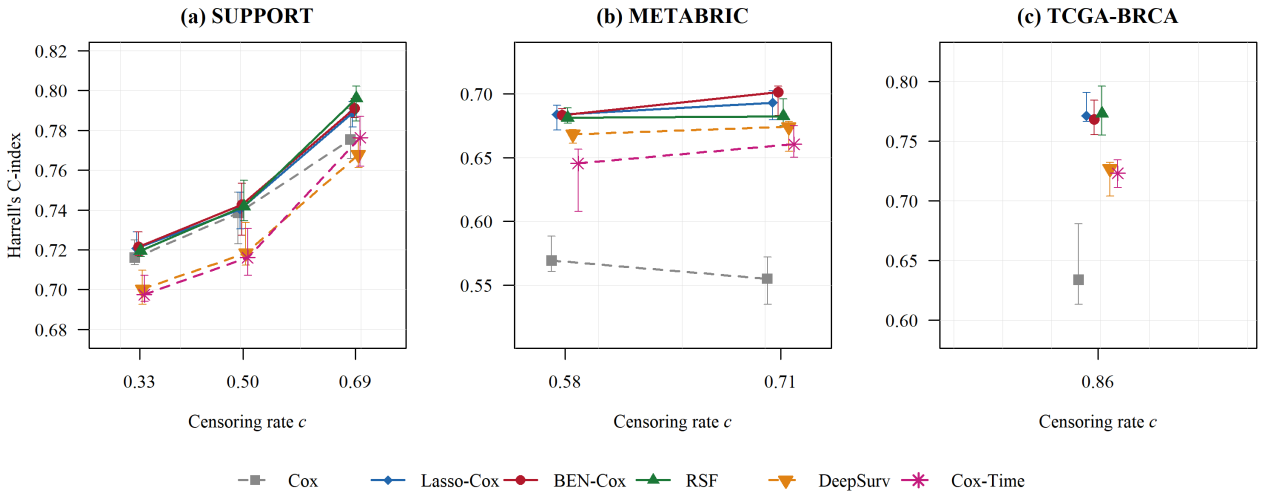


FIGURE 1. Harrell's *C*-index as a function of the censoring rate *c*. Each panel corresponds to one dataset; each curve corresponds to one model. Error bars indicate the interquartile range over the 5 × 2 cross-validation folds. TCGA-BRCA is shown at its baseline censoring level only, as explained in Section IV-C.

TABLE 2. Median Harrell's *C*-index (interquartile range) per dataset–regime combination. Bold indicates the highest median within each combination.

Dataset	c	Cox	Lasso-Cox	BEN-Cox	RSF	DeepSurv	Cox-Time
METABRIC	0.58	.569 (.561–.588)	.684 (.672–.691)	.684 (.681–.689)	.681 (.677–.689)	.668 (.661–.672)	.646 (.608–.657)
METABRIC	0.70	.555 (.535–.572)	.693 (.680–.702)	**.701** (.683–.706)	.682 (.675–.696)	.674 (.655–.679)	.661 (.650–.675)
SUPPORT	0.33	.716 (.712–.725)	.721 (.718–.729)	**.722** (.717–.729)	.720 (.717–.720)	.700 (.693–.710)	.697 (.694–.707)
SUPPORT	0.50	.738 (.723–.749)	.740 (.730–.749)	**.743** (.727–.754)	.742 (.734–.755)	.718 (.712–.734)	.716 (.707–.731)
SUPPORT	0.70	.775 (.766–.791)	.789 (.782–.794)	.791 (.786–.796)	**.796** (.785–.802)	.768 (.761–.776)	.776 (.762–.787)
TCGA-BRCA	0.86	.634 (.613–.681)	.771 (.766–.791)	.768 (.755–.784)	**.773** (.755–.796)	.727 (.704–.732)	.723 (.711–.734)

3) CALIBRATION SLOPE

Table 4 and Figure 3 report median calibration slopes at the per-fold evaluation horizon defined in Section IV-D. The most informative contrast is across censoring regimes on SUPPORT, where three levels of c are available. Lasso-Cox and BEN-Cox maintain slopes within [0.99, 1.18] across all three regimes, indicating stable calibration as censoring increases. RSF degrades from 1.08 at baseline to 0.86 at $c = 0.70$. DeepSurv remains overconfident with slopes near 0.50 at all three censoring levels. Cox-Time improves from 0.54 at baseline to 0.78 at $c = 0.70$, the opposite of the pattern shown by DeepSurv, but remains below the calibrated range.

On METABRIC, the unpenalized Cox model collapses to a calibration slope of 0.03 at both censoring levels, indicating that its predicted risk spread bears essentially no relation to observed outcomes. On TCGA-BRCA, all models except lasso-Cox (slope = 0.88) fall below 0.65, reflecting the difficulty of calibrating absolute risk predictions when 86% of subjects are censored.

4) PROPORTIONAL HAZARDS DIAGNOSTICS

The Grambsch–Therneau global test described in Section III-C was computed on each training fold. On METABRIC and TCGA-BRCA the test could not be evaluated reliably because the unpenalized Cox model failed to converge or produced numerically unstable Schoenfeld residuals. On SUPPORT the global p-value is below 0.01 in all three censoring regimes (median $p < 10^{-3}$ at baseline and $c = 0.50$; median $p \approx 0.004$ at $c = 0.70$), indicating that the proportional hazards assumption is violated in this dataset across all regimes considered.

This finding has a direct bearing on the interpretation of the relative performance of Cox-Time, the only model in the benchmark that relaxes the proportional hazards assumption. Despite the documented PH violation on SUPPORT, Cox-Time does not achieve the highest discrimination or the best calibration in any SUPPORT configuration. This suggests that, in the EPV range covered by the present benchmark, the sample-efficiency advantage of regularized proportional-hazards models outweighs the specification advantage of a non-proportional model, even when the PH assumption is substantively violated.

C. CALIBRATION AND INTEGRATED BRIER SCORE

The per-configuration results of Section V-B show that BEN-Cox achieves the lowest IBS in every configuration while its calibration slope is mildly below one (pooled median 0.87), whereas lasso-Cox has slopes closer to one (pooled median 1.09) but higher IBS. This apparent contrast reflects the fact that the two quantities measure different aspects of the predictive distribution.

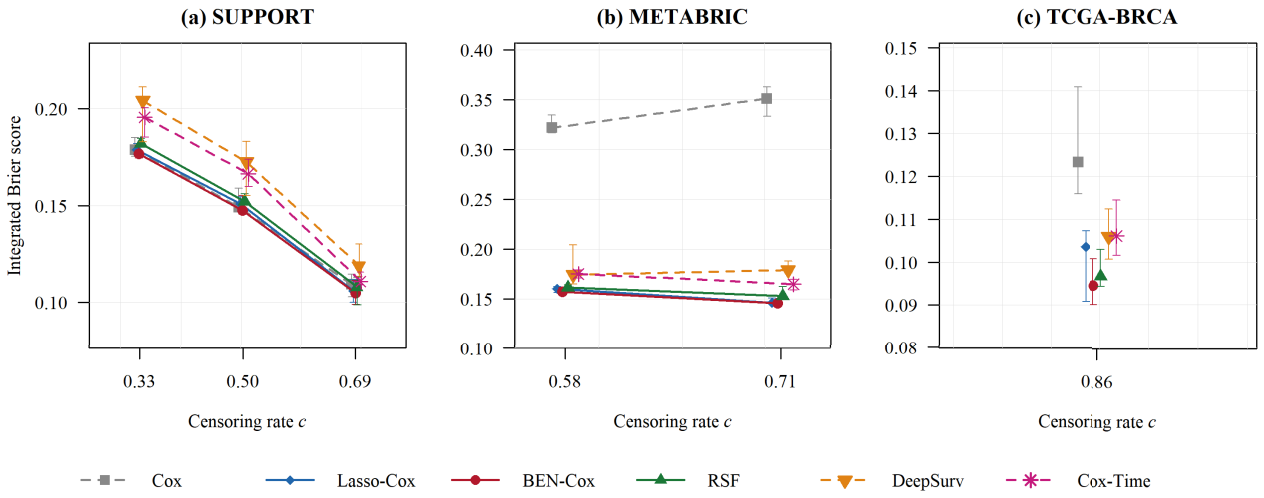


FIGURE 2. Integrated Brier score as a function of the censoring rate c. Lower values indicate better predictive accuracy. Layout and error bars as in Fig. 1.

TABLE 3. Median integrated Brier score per dataset–regime combination. Bold indicates the lowest (best) median within each combination.

Dataset	c	Cox	Lasso-Cox	BEN-Cox	RSF	DeepSurv	Cox-Time
METABRIC	0.58	.322	.160	**.157**	.162	.174	.175
METABRIC	0.70	.351	.146	**.146**	.153	.179	.165
SUPPORT	0.33	.179	.179	**.177**	.182	.204	.196
SUPPORT	0.50	.149	.151	**.148**	.152	.173	.166
SUPPORT	0.70	.109	.106	**.105**	.108	.119	.111
TCGA-BRCA	0.86	.123	.104	**.095**	.097	.106	.106

TABLE 4. Median calibration slope per dataset–regime combination. Values closer to one indicate better calibration. Bold indicates the value closest to one within each combination.

Dataset	c	Cox	Lasso-Cox	BEN-Cox	RSF	DeepSurv	Cox-Time
METABRIC	0.58	.03	**.99**	.81	1.29	.37	.47
METABRIC	0.70	.03	**1.07**	.80	1.12	.27	.37
SUPPORT	0.33	.82	1.15	**1.01**	1.08	.47	.54
SUPPORT	0.50	.84	1.10	**1.04**	.98	.50	.51
SUPPORT	0.70	.77	1.18	**1.01**	.86	.51	.78
TCGA-BRCA	0.86	.07	**.88**	.62	.60	.56	.50

The integrated Brier score is a proper scoring rule that rewards both sharpness and calibration jointly: a model that produces well-separated predicted probabilities with slight overconfidence can achieve a lower IBS than a model with a slope closer to one but wider predicted intervals. BEN-Cox's posterior predictive distribution is sharper than that of lasso-Cox and RSF on most configurations, and this sharpness advantage compensates for its mildly sub-unit slope in the IBS composite. Lasso-Cox, conversely, achieves near-unit slopes by producing a less concentrated risk spread. The two measures are therefore complementary rather than contradictory, and both are needed to characterize predictive performance: the IBS for overall predictive accuracy, and the calibration slope for the reliability of absolute risk estimates.

Neither BEN-Cox nor lasso-Cox can be declared uniformly best-calibrated across the grid. On SUPPORT, BEN-Cox is closer to one at all three censoring levels. On METABRIC, lasso-Cox is closer to one at both levels. On TCGA-BRCA, lasso-Cox again leads with a slope of 0.88 against 0.62 for BEN-Cox. This dataset dependence suggests that the calibration advantage of Bayesian shrinkage is most pronounced in the moderate-p, moderate-c regime (SUPPORT), while frequentist penalization provides more stable calibration in the high-p and very-high-c regimes (METABRIC, TCGA-BRCA). DeepSurv, Cox-Time, and the unpenalized Cox model remain substantially overconfident across the full grid, with pooled slopes of 0.44, 0.50, and 0.51 respectively.

The TCGA-BRCA configuration calls for a bit closer reading, since it is the only point in the experimental grid with $c \approx 0.86$ and therefore the closest available approximation to the extreme-censoring regime. At this level, the linear calibration estimator is itself fitted on a small effective sample, and the slope is correspondingly noisier than in the moderate-censoring configurations. Even so, the qualitative ordering remains interpretable. Lasso-Cox attains

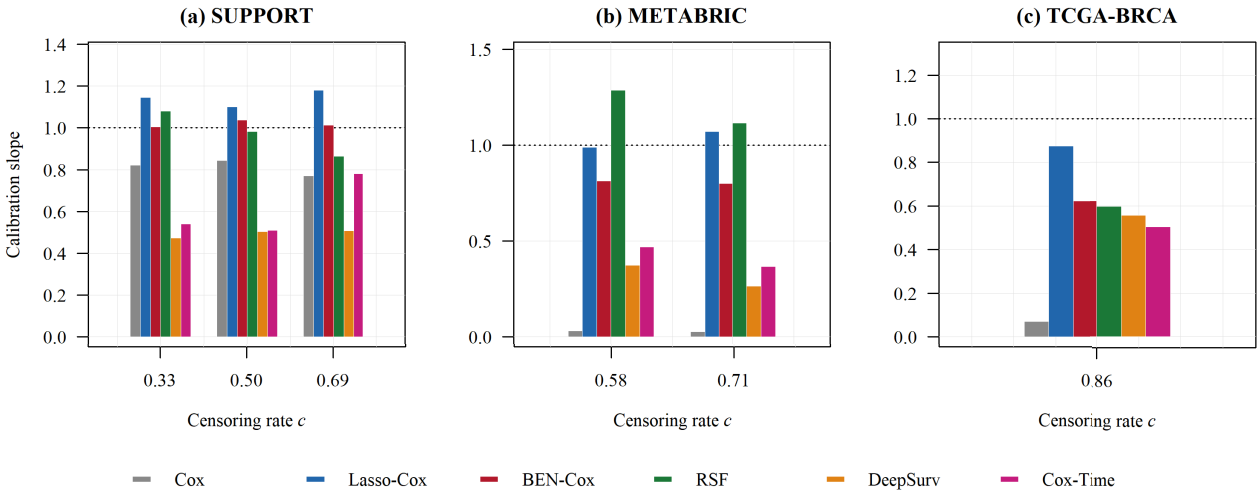


FIGURE 3. Calibration slope per dataset–regime combination. The dotted horizontal line marks slope = 1 (perfect calibration). Slopes below one indicate overconfident risk predictions; slopes above one indicate underconfident predictions.

a slope of 0.88, BEN-Cox and RSF settle near 0.60, and the unpenalized Cox model collapses to 0.07, while DeepSurv and Cox-Time remain near 0.50. The pattern suggests that strong regularization helps maintain a credible risk spread even when the event count is seriously reduced, whereas unregularized and flexible estimators degrade sharply in this part of the grid. A more complete characterization of calibration behavior at extreme censoring would require datasets that natively sit in this regime, and this extension is identified as a direction for future work in Section VII-E.

D. INTERACTION BETWEEN CENSORING AND EVENTS PER VARIABLE

The empirical findings of Sections V-A–V-C are further clarified by examining the joint structure of the censoring rate c and the events-per-variable ratio defined in Eq. (5). Figure 4 displays the six dataset–regime configurations of the experimental grid in the (c, EPV) plane, with each configuration labeled by the model that achieved the highest median C-index across the 5×2 cross-validation folds.

The most striking feature of Figure 4 is that the entire experimental grid lies below the classical EPV $=$ 10 guideline of [14]. The observed EPV values range from approximately 0.6 (METABRIC at $c = 0.70$) to 7.6 (SUPPORT at baseline), a range that is representative of many applied biomedical survival settings, particularly in oncology and molecular cohorts where the number of candidate predictors is large relative to the number of observed events.

Within this low-EPV range, two models share the top rank. BEN-Cox achieves the highest median C-index in four of six configurations (METABRIC at $c = 0.70$, SUPPORT at baseline, SUPPORT at $c = 0.50$, and—effectively tied with lasso-Cox—METABRIC at baseline). RSF achieves the highest median C-index in the remaining two configurations (SUPPORT at $c = 0.70$ and TCGA-BRCA at baseline). The margins separating the top two models are small in most configurations (Table 2), with only the METABRIC $c = 0.70$ configuration showing a gap exceeding 0.005. These margins are comparable to or smaller than the cross-fold interquartile ranges reported in Figure 1, so the partition should be read as indicating which model families are *competitive* in each region rather than which model is definitively superior.

Deep survival models do not achieve the top rank in any configuration of the present grid. This finding does not imply that deep architectures are uniformly inferior; rather, it indicates that in the EPV range covered here (EPV < 8), the sample-efficiency advantage of regularized linear models and ensemble methods outweighs the flexibility advantage of neural survival architectures. Whether deep models would gain an advantage at higher EPV values, for example on larger clinical cohorts with lower predictor dimensionality, cannot be determined from the present data and is identified as a direction for future work in Section VII-E.

E. COMPUTATIONAL COST

Figure 5 reports median wall-clock training time per fold, pooled across all dataset–regime combinations. The six models span roughly four orders of magnitude in training cost. The unpenalized Cox model is fastest at 0.07 seconds per fold, followed by lasso-Cox at 1.2 seconds. Cox-Time (6.2 s) and DeepSurv (15.1 s) occupy an intermediate range, with their cost driven primarily by the random hyperparameter search described in Section IV-D. RSF requires 22.5 seconds per fold with 500 trees. BEN-Cox is the most expensive at 114 seconds per fold due to Hamiltonian Monte Carlo sampling, roughly twice the cost of DeepSurv and five times that of Cox-Time.

From a practical standpoint, all six models remain feasible on a single workstation for the dataset sizes considered. The 5×2 cross-validation scheme produces ten folds per dataset–regime combination, so the total per-configuration cost ranges from under one second (Cox) to approximately

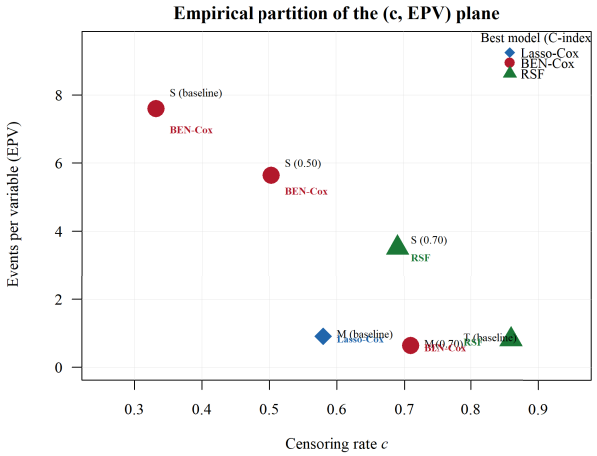


FIGURE 4. Empirical partition of the (c, EPV) plane. Each point corresponds to one dataset–regime combination, positioned at its censoring rate c and events-per-variable ratio (Eq. 5), and labeled with the model that achieved the highest median C-index. The dashed line marks EPV = 10, the classical guideline of [14]. All six configurations fall below this threshold.

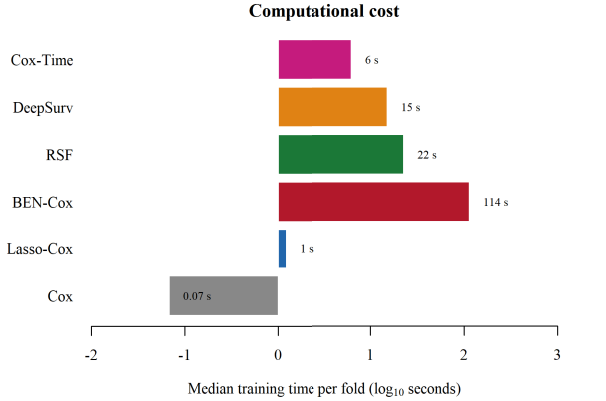


FIGURE 5. Median wall-clock training time per fold for each model, pooled across all dataset–regime configurations. The horizontal axis is on a $\log_{10}$ scale.

twenty minutes (BEN-Cox). The full benchmark, comprising six configurations and six models, completed in approximately eight hours of wall-clock time. Per-model training times are also reported in the last column of Table 1.

Training cost, however, should not be interpreted as the whole computational burden of the methods. After model fitting is completed, prediction becomes almost constant time per subject for the partial-likelihood-based models, namely Cox, lasso-Cox, and BEN-Cox. This is because prediction only requires evaluating a linear risk score and using the already estimated baseline hazard. For BEN-Cox, the posterior summary is also obtained once during model fitting, and it is not recomputed at the prediction stage. In contrast, RSF needs to traverse all 500 trees for each subject, and therefore it is slower at prediction, although the required time is still only in the millisecond range on a standard workstation. DeepSurv and Cox-Time only require a single forward pass through a neural network with at most three hidden layers, and in terms of wall-clock time this is comparable to RSF traversal. The peak resident memory usage was modest for all six methods, including the METABRIC case with $p = 440$ predictors. In our setting, the main memory requirement came from storing cross-validation predictions rather than from any individual fitted model. Therefore, for the cohort sizes considered in this study, none of the six methods can be considered computationally prohibitive. The main differences shown in Fig. 5 are mostly related to model fitting rather than to the deployment or prediction stage.

VI. EMPIRICAL PARTITION OF THE (C, EPV) PLANE

The empirical findings of Section V, and in particular the interaction between censoring and the events-per-variable ratio examined in Section V-D, suggest that two quantities computable directly from the data, the censoring rate c and the events-per-variable ratio defined in Eq. (5), capture much of the variation in relative model performance across the experimental grid. This section describes the partition observed in the benchmark, compares it with the classical events-per-variable guideline, and discusses its interpretation and limitations. The partition is reported as a descriptive summary of the present experiments, not as a prescriptive rule for new datasets.

A. DESCRIPTION OF THE PARTITION

For each dataset–regime combination in the experimental grid, we record the model that achieved the highest median C-index across the 5×2 cross-validation folds, together with the pair (c, EPV) computed from the training folds using Eq. (5). The resulting labeled points are displayed in Figure 4, together with the winning model in each configuration.

The most prominent feature of the partition is that all six configurations fall below the classical EPV $=$ 10 threshold of [14], with observed EPV values ranging from approximately 0.6 (METABRIC at $c = 0.70$) to 7.6 (SUPPORT at baseline). This range is representative of many applied biomedical survival settings, particularly in oncology and molecular cohorts where the number of candidate predictors is large relative to the number of observed events. No configuration in the present grid reaches the regime in which deep survival models have been reported to show their strongest advantage.

Within this low-EPV range, two model families share the top rank. BEN-Cox achieves the highest median C-index in four of six configurations: METABRIC at $c = 0.70$, SUPPORT at baseline, SUPPORT at $c = 0.50$, and METABRIC at baseline (where lasso-Cox is within 0.0003). RSF achieves the highest median C-index in the remaining two configurations: SUPPORT at $c = 0.70$ and TCGA-BRCA at baseline. The deep survival models DeepSurv and Cox-Time do not achieve the top rank in any configuration of the present grid. Table 5 summarizes the observed partition.

The margins separating the top two models are small in most configurations (Table 2), with only the METABRIC $c =$

TABLE 5. Observed partition of the (c, EPV) plane in the experimental grid. The entries summarize the best-performing model family in each region and should be interpreted as descriptive of the present benchmark rather than prescriptive for new datasets.

EPV range	c range	Best-performing family
< 1 (high-p)	0.58–0.86	BEN-Cox / Lasso-Cox
1–4 (moderate)	0.50–0.70	BEN-Cox / RSF
4–8 (moderate)	0.33	BEN-Cox / RSF

Deep models (DeepSurv, Cox-Time): not top-ranked in any observed configuration.

0.70 configuration showing a gap exceeding 0.005. These margins are comparable to or smaller than the cross-fold interquartile ranges reported in Figure 1, so the partition should be read as indicating which model families are *competitive* in each region rather than which model is definitively superior.

B. RELATION TO THE CLASSICAL EVENTS-PER-VARIABLE GUIDELINE

The classical recommendation of [14] that unpenalized Cox regression requires approximately ten events per variable to yield stable estimates is contextualized by the present findings in two ways.

First, the threshold is not sharp: performance degrades gradually as the events-per-variable ratio decreases, and the model that best handles the low-EPV regime is not the unpenalized Cox model but BEN-Cox and, to a comparable extent, lasso-Cox and RSF. The hierarchical shrinkage prior of BEN-Cox absorbs much of the instability that motivated the original guideline, while lasso-Cox achieves similar discrimination through frequentist penalization. The unpenalized Cox model, by contrast, collapses on the high-dimensional datasets (calibration slopes near zero on METABRIC and TCGA-BRCA), confirming that the EPV $=$ 10 guideline retains practical relevance specifically for unregularized estimators.

Second, censoring plays a role that the original guideline did not consider. Within the EPV range covered by the present benchmark (EPV < 8), increasing censoring does not change which model family is preferred—BEN-Cox and RSF remain competitive throughout—but it does widen the gap between the regularized top tier and the deep architectures. At baseline censoring on SUPPORT ($c \approx 0.33$), the gap between BEN-Cox and DeepSurv is 0.021 in C-index; at $c = 0.70$ the corresponding gap between RSF and DeepSurv is 0.028. This asymmetric degradation suggests that the classical EPV rule is best understood as one coordinate of a joint criterion in (c, EPV), with the censoring rate modulating the severity of the sample-efficiency constraint rather than the identity of the preferred model family.

C. INTERPRETATION AND LIMITATIONS

Three caveats merit emphasis before the partition is used to inform practice.

First, the partition is constructed from only six dataset–regime combinations: two regimes from METABRIC, three from SUPPORT, and one from TCGA-BRCA. This coverage is constrained by the baseline censoring rates of the three datasets and by the one-sided nature of the censoring augmentation protocol described in Section IV-C, which can only increase censoring relative to the baseline. The boundaries between regions should therefore be understood qualitatively and not as sharp numerical thresholds; a larger experimental grid would be required to estimate such thresholds with statistical confidence, and this extension is identified as a priority in Section VII-E.

Second, the partition depends only on (c, EPV) and ignores other structural features such as covariate correlation, the plausibility of the proportional hazards assumption, and the availability of informative prior information. These factors are known to influence model performance and would need to be incorporated into any richer analysis. For these reasons, the partition is presented as an empirical characterization of the regimes observed in the benchmark, and as a starting point for discussion rather than as a replacement for careful model selection in any individual study.

Third, the experimental grid does not include any configuration with EPV $\geq$ 10, which is precisely the regime in which deep survival models are most likely to show a competitive advantage. The absence of deep models from the top rank in Table 5 should therefore not be read as evidence that deep architectures are generally inferior, but rather as evidence that their advantage does not materialize in the low-EPV conditions that characterize the present benchmark. Extending the grid to include larger cohorts with lower predictor dimensionality, such as the full SUPPORT cohort or the Rotterdam breast cancer dataset, would be necessary to map the boundary at which deep models become competitive, and this extension is identified as future work in Section VII-E.

VII. DISCUSSION

The benchmark reported in Sections V and VI was designed to address a specific question: how does the joint structure of the censoring rate c and the events-per-variable ratio EPV (Eq. 5) shape the relative performance of Bayesian regularized and deep learning survival models? This section interprets the findings, relates them to the existing literature, and discusses threats to validity, limitations, and directions for future work.

A. INTERPRETATION OF THE MAIN FINDINGS

Three principal observations emerge from the benchmark.

First, model ranking in discrimination is not determined by model complexity. BEN-Cox, lasso-Cox, and RSF, three methodologically distinct models spanning Bayesian shrinkage, frequentist penalization, and nonparametric ensembles, form a narrow top tier with pooled C-index values within 0.003 of each other (Table 1). The deep architectures DeepSurv and Cox-Time do not reach this tier in any of the

six dataset–regime configurations examined, despite having the capacity to capture nonlinear effects and, in the case of Cox-Time, non-proportional hazards. This ordering is stable across censoring regimes: the gap between the top tier and the deep models does not narrow as censoring decreases (Section V-B).

Second, calibration separates models that discrimination does not. While BEN-Cox, lasso-Cox, and RSF are nearly indistinguishable in C-index, they differ meaningfully in calibration behavior. BEN-Cox achieves the lowest integrated Brier score in all six configurations, reflecting the joint contribution of sharpness and calibration to the IBS composite (Section V-C). Lasso-Cox produces calibration slopes closest to one on the high-dimensional datasets (METABRIC and TCGA-BRCA), while BEN-Cox is closest to one on SUPPORT. DeepSurv, Cox-Time, and the unpenalized Cox model are systematically overconfident, with pooled slopes near 0.5 or below. In applied settings where absolute risk estimates inform patient-level decisions, this calibration gap may matter more than the small discrimination differences within the top tier.

Third, the entire experimental grid falls below the classical EPV $=$ 10 guideline (Figure 4), with observed EPV values between 0.6 and 7.6. In this low-EPV regime, the sample-efficiency advantage of regularized models dominates the flexibility advantage of neural architectures. Whether deep models would become competitive at higher EPV cannot be determined from the present data.

Two mechanisms plausibly explain these patterns. The hierarchical shrinkage prior of BEN-Cox, described in Section III-B, encodes a structural assumption of approximate sparsity that is particularly valuable when the effective sample size is small relative to p. In such regimes, the variance reduction attributable to shrinkage outweighs the bias introduced by the prior, in a manner reminiscent of the classical bias–variance trade-off for ridge-type estimators. Concurrently, the Cox partial likelihood of Eq. (2) loses information with each additional censored observation, and deep architectures with many parameters have less information with which to regularize themselves in high-censoring regimes. Lasso-Cox benefits from an analogous mechanism through its ℓ_1 penalty, which explains its comparable discrimination to BEN-Cox at a fraction of the computational cost (Section V-E). RSF, while not applying explicit coefficient shrinkage, achieves implicit regularization through ensemble averaging and the restriction of splits to random covariate subsets, which yields similar sample efficiency in the regimes studied here.

A remark on the proportional hazards assumption is warranted. The Grambsch–Therneau test indicates PH violation on all SUPPORT configurations where it could be evaluated (Section V-B). Despite this, Cox-Time, the only model that relaxes the PH assumption, does not outperform the PH-based models on SUPPORT in either discrimination or calibration. This suggests that, at the sample sizes and EPV values covered by the present benchmark, correct specification of the time-dependence structure is less important than sample-efficient estimation of the covariate effects. This finding should not be generalized to larger datasets or to settings with stronger departures from proportionality.

A practical implication of these findings is related to how predicted survival probabilities are used in later clinical decision-making. In some clinical settings, absolute risk estimates directly support individual-level decisions, such as adjuvant therapy planning, risk-stratified surveillance, or selecting patients who may need more intensive follow-up. Therefore, the calibration difference observed between the regularized linear models and the deep learning models is not only a technical issue, but it also affects the reliability of these decisions. A model may rank patients reasonably well, but if it gives systematically overconfident absolute risk estimates, it may still lead to inappropriate decisions when these probabilities are interpreted directly rather than only used as risk scores. For the EPV ranges considered in this benchmark, BEN-Cox and lasso-Cox therefore seem to be more defensible default choices when the predicted probabilities, and not only the relative ranking of patients, are used in subsequent clinical reasoning.

B. RELATION TO PRIOR BENCHMARKS

The results of Section V are broadly consistent with the findings of [11], who reported that classical Cox-based methods remain competitive with more complex alternatives on multi-omics cancer data. The present study extends that observation by showing that the competitiveness of regularized Cox methods is concentrated in the low-EPV regime and persists across the full range of censoring levels examined, while the deep architectures do not reach the top tier in any configuration of the present grid. A recent systematic review by O'Donnell et al. [28] across 90 cancer survival predictors reached a consistent conclusion, noting that regularized and tree-based methods frequently perform comparably to or better than deep learning alternatives.

The systematic review of [12] noted that benchmarks of deep survival methods often lack clear guidance on the regimes in which these methods are preferable. The (c, EPV) partition of Figure 4 and the descriptive summary of Section VI-A represent one empirical step toward such guidance, with the important caveat that the present grid covers only the low-EPV regime and therefore cannot identify the boundary at which deep models become competitive.

The classical rule of ten events per variable originally proposed by [14] for unpenalized Cox regression is contextualized here as one coordinate of a joint criterion in (c, EPV). The finding that regularized models—both Bayesian and frequentist—perform well below EPV $=$ 10 is consistent with the relaxations suggested by [15], and the present benchmark adds the observation that censoring modulates the severity of the low-EPV penalty differently across model families.

C. THREATS TO VALIDITY

Three threats to validity warrant careful discussion.

First, the censoring augmentation protocol of Section IV-C generates non-informative censoring by construction, since the auxiliary times $C_i^\star$ are drawn independently of T_i and $\mathbf{x}_i$. In real-world data, censoring is often informative, for example when patients with deteriorating prognosis withdraw from follow-up. The conclusions of this paper should therefore be understood as applying to the non-informative censoring case, and methods explicitly designed to handle informative censoring may alter the relative ranking of models in informative settings.

Second, the experimental grid consists of three datasets and a discrete set of target censoring rates, producing only six dataset–regime configurations. Although the three datasets differ substantively in sample size, dimensionality, and clinical context, they do not exhaust the heterogeneity of biomedical survival data. In particular, no configuration reaches EPV $\geq$ 10, so the regime in which deep models are most likely to be competitive is not represented. The qualitative regions reported in Table 5 should accordingly be interpreted as a descriptive summary of the present experiments rather than as sharp or universal boundaries.

Third, the tuning budgets allocated to the six models are deliberately modest and matched across methods, as described in Section IV-D. This design choice favors methods that are less sensitive to hyperparameter configuration and may understate the performance of deep survival models, which are known to benefit from more extensive architecture search. While this asymmetry cannot be fully eliminated without committing substantially more compute, a budget-matched comparison arguably better reflects the conditions under which applied researchers typically operate, and it protects against the optimistic bias that arises when one family of models is tuned more aggressively than another. The deep models in the present benchmark used a random search of fixed size with early stopping; a larger search budget or more sophisticated tuning strategies such as Bayesian optimization could improve their performance, and this possibility should be considered when interpreting the results.

D. LIMITATIONS

Beyond the threats to validity discussed above, several limitations of the present study should be acknowledged. The benchmark is confined to right-censored single-event data, and competing risks, interval censoring, and recurrent events are not considered. Time-dependent covariates are likewise not used, even where available, in order to keep the comparison focused on the role of the baseline covariate vector $\mathbf{x}_i \in \mathbb{R}^p$ introduced in Section III-A. The empirical partition described in Section VI depends on only two quantities, (c, EPV), and intentionally ignores other relevant structural features such as covariate correlation, the validity of the proportional hazards assumption, and the availability of domain-informed priors. These factors are known to affect model performance and could, in principle, be incorporated into a richer analysis at the cost of reduced interpretability.

Additionally, the EPV values reported throughout the paper use the raw post-preprocessing predictor count p for RSF, DeepSurv, and Cox-Time, and the effective dimension p_{eff} for lasso-Cox and BEN-Cox, as discussed in Section III-C. This asymmetry means that the EPV for the regularized Cox variants is nominally higher than the EPV for the remaining models at the same configuration, which should be kept in mind when reading the partition of Section VI-A.

E. DIRECTIONS FOR FUTURE WORK

Several extensions of the present study would be informative.

A first natural direction is to enlarge the dataset grid with additional public cohorts spanning a wider range of baseline censoring levels and events-per-variable ratios, in particular cohorts with EPV $\geq$ 10. Candidates include the full SUPPORT cohort ($n \approx 9{,}000$), the Rotterdam breast cancer dataset ($n \approx 3{,}000$), and the GBSG trial ($n \approx 2{,}200$), each of which would contribute configurations in the moderate-to-high EPV range that is absent from the present grid. Such an expansion would make it possible to determine whether the deep survival models become competitive at higher EPV, as the existing literature suggests but the present benchmark cannot confirm.

A second direction is to extend the benchmark to informative censoring settings, either through simulation or by using datasets for which informative censoring is documented, in order to assess the robustness of the empirical partition in that more demanding regime.

A third direction concerns the role of covariate correlation: a controlled simulation study in which the correlation structure of $\mathbf{x}$ is varied while holding (c, EPV) fixed would help disentangle the contribution of correlation from that of dimensionality.

Finally, the empirical partition could be extended beyond (c, EPV) to incorporate formal tests of the proportional hazards assumption, which would allow Cox-Time and related non-proportional methods to be characterized on principled grounds rather than purely empirically. The finding that PH violation on SUPPORT did not translate into a Cox-Time advantage (Section V-B) suggests that the relationship between PH violation severity, sample size, and model ranking is itself a research question deserving systematic study.

VIII. CONCLUSION

This paper presented a systematic benchmark of six representative survival models, spanning the classical Cox proportional hazards model, its lasso-penalized and Bayesian elastic net extensions, random survival forests, and two deep survival architectures, evaluated across three public biomedical datasets under controlled censoring regimes. The benchmark was designed to isolate the joint effect of the

censoring rate c and the events-per-variable ratio EPV (Eq. 5) on model ranking, a question that has received limited systematic attention in prior benchmark studies.

Three findings emerged. First, in the low-EPV regime covered by the experimental grid (EPV $<$ 8), BEN-Cox, lasso-Cox, and random survival forests form a narrow top tier in discrimination, with pooled C-index values within 0.003 of each other, while the deep survival architectures DeepSurv and Cox-Time do not reach this tier in any dataset–regime configuration examined. Second, calibration separates models that discrimination does not: BEN-Cox achieves the lowest integrated Brier score in all six configurations, reflecting its sharper posterior predictive distribution, while lasso-Cox produces calibration slopes closest to one on the high-dimensional datasets. Deep models and the unpenalized Cox model are systematically overconfident across the full grid. Third, the entire experimental grid falls below the classical EPV $=$ 10 guideline, and within this range, increasing censoring widens the gap between the regularized top tier and the deep architectures without changing which model family is preferred.

These findings were summarized descriptively through an empirical partition of the (c, EPV) plane, presented in Section VI. The partition complements the classical events-per-variable guideline by explicitly accounting for censoring and by documenting that regularized models, both Bayesian and frequentist, perform well below EPV $=$ 10. It is offered as a compact summary of the present experiments rather than as a prescriptive rule, and its qualitative nature reflects the limited number of dataset–regime combinations that were available.

An important limitation of the present benchmark is that no configuration reaches the EPV $\geq$ 10 regime in which deep survival models are most likely to show a competitive advantage. The absence of deep models from the top rank should therefore be read as specific to the low-EPV conditions studied here, not as a general verdict on their utility. The principal directions for future work identified in Section VII-E concern the extension of the benchmark to larger cohorts with higher EPV, informative censoring settings, and the incorporation of covariate correlation and proportional hazards diagnostics into the partition analysis. Together, these extensions would clarify the boundary at which deep survival models become competitive and provide a more complete account of when each family of methods is preferable in practice.

DATA AVAILABILITY

All datasets used in this study are publicly available. The METABRIC dataset [17] is accessible through the cBioPortal for Cancer Genomics (https://www.cbioportal.org). The SUPPORT dataset [18] is available through the Vanderbilt Biostatistics dataset archive (https://hbiostat.org/data). The TCGA-BRCA dataset [19] is accessible through the Genomic Data Commons portal of the National Cancer Institute (https://portal.gdc.cancer.gov). The benchmark-ready versions of all three datasets, including the preprocessing and censoring augmentation scripts used in this study, are provided in the accompanying code repository.

CODE AVAILABILITY

Complete source code for all experiments, including data loading, preprocessing, censoring, augmentation, cross-validation scaffolding, model fitting, evaluation, and statistical comparison is publicly available at https://github.com/yilmazersin13/survival-model-benchmark-r. The repository includes all R scripts, the Stan implementation of BEN-Cox, Python wrappers for the `pycox` models (DeepSurv and Cox-Time), fixed random seeds, and package version information sufficient for independent replication. The BEN-Cox Stan model code is also available as part of the repository accompanying the original BEN-Cox publication [6].

ACKNOWLEDGMENT

The author acknowledges the data contributors of the METABRIC, SUPPORT, and TCGA-BRCA cohorts for making these datasets publicly available for research.

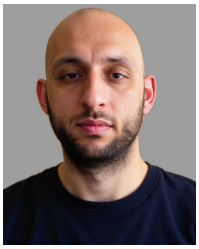
DECLARATION OF COMPETING INTEREST

The author declares no competing interests.

REFERENCES

[1] D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 34, no. 2, pp. 187–220, Jan. 1972.
[2] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statist. Med.*, vol. 16, no. 4, pp. 385–395, Feb. 1997.
[3] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *J. Stat. Softw.*, vol. 39, no. 5, pp. 1–13, 2011.
[4] K. H. Lee, S. Chakraborty, and J. Sun, "Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data," *Int. J. Biostatistics*, vol. 7, no. 1, pp. 1–32, Apr. 2011.
[5] K. H. Lee, S. Chakraborty, and J. Sun, "Survival prediction and variable selection with simultaneous shrinkage and grouping priors," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 8, no. 2, pp. 114–127, Apr. 2015.
[6] E. Yılmaz, S. E. Ahmed, and D. Aydın, "Bayesian elastic net cox models for time-to-event prediction: Application to a breast cancer cohort," *Entropy*, vol. 28, no. 3, p. 264, Feb. 2026.
[7] H. Ishwaran, U. B. Kogalur, E. H. Blackwell, and M. S. Lauer, "Random survival forests," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 841–860, 2008.
[8] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 24, Dec. 2018.
[9] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and cox regression," *J. Mach. Learn. Res.*, vol. 20, no. 129, pp. 1–30, 2019.
[10] C. Lee, W. R. Zame, J. Yoon, and M. V. D. Schaar, "DeepHit: A deep learning approach to survival analysis with competing risks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 2314–2321.
[11] M. Herrmann, P. Probst, R. Hornung, V. Jurinovic, and A.-L. Boulesteix, "Large-scale benchmark study of survival prediction methods using multi-omics data," *Briefings Bioinf.*, vol. 22, no. 3, p. 167, May 2021.
[12] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, and A. Bender, "Deep learning for survival analysis: A review," *Artif. Intell. Rev.*, vol. 57, no. 3, p. 65, Feb. 2024.
[13] L. Burk, J. Zobolas, R. Bender, B. Bischl, and A. Bender, "A large-scale neutral comparison study of survival models on low-dimensional data," 2024, *arXiv:2406.04098*.

[14] P. Peduzzi, J. Concato, A. R. Feinstein, and T. R. Holford, "Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates," *J. Clin. Epidemiol.*, vol. 48, no. 12, pp. 1503–1510, Dec. 1995.
[15] E. Vittinghoff and C. E. McCulloch, "Relaxing the rule of ten events per variable in logistic and cox regression," *Amer. J. Epidemiology*, vol. 165, no. 6, pp. 710–718, Jan. 2007.
[16] R. D. Riley, K. I. E. Snell, J. Ensor, D. Burke, F. E. Harrell, K. G. M. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes," *Statist. Med.*, vol. 38, no. 7, pp. 1276–1296, 2018.
[17] M. Group et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, Jun. 2012.
[18] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, and D. P. Wagner, "The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults," *Ann. Internal Med.*, vol. 122, no. 3, pp. 191–203, Feb. 1995.
[19] T. C. G. A. Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.
[20] Q. Li and N. Lin, "The Bayesian elastic net," *Bayesian Anal.*, vol. 5, no. 1, pp. 151–170, Mar. 2010.
[21] H. Zou, T. Hastie, and R. Tibshirani, "On the degrees of freedom of the lasso," *Ann. Statist.*, vol. 35, no. 5, pp. 2173–2192, 2007.
[22] J. Piironen and A. Vehtari, "Sparsity information and regularization in the horseshoe and other shrinkage priors," *Electron. J. Statist.*, vol. 11, no. 2, pp. 5018–5051, Jan. 2017.
[23] P. M. Grambsch and T. M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, vol. 81, no. 3, pp. 515–526, 1994.
[24] F. E. Harrell, "Evaluating the yield of medical tests," *JAMA, J. Amer. Med. Assoc.*, vol. 247, no. 18, pp. 2543–2546, May 1982.
[25] H. Uno, T. Cai, M. Pencina, R. B. D'Agostino, and L. J. Wei, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statist. Med.*, vol. 30, no. 10, pp. 1105–1117, 2011.
[26] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data," *Statist. Med.*, vol. 18, nos. 17–18, pp. 2529–2545, Sep. 1999.
[27] F. Wan, "Simulating survival data with predefined censoring rates for proportional hazards models," *Statist. Med.*, vol. 36, no. 5, pp. 838–854, Feb. 2017.
[28] A. O'Donnell, M. Cronin, S. Moghaddam, and E. Wolsztynski, "A systematic review on machine learning techniques for survival analysis in cancer," *Cancer Med.*, vol. 14, no. 22, p. 71375, Nov. 2025.



ERSIN YILMAZ received the B.Sc. degree in statistics from Ege University, Türkiye, in 2013, and the M.Sc. and Ph.D. degrees in statistics from Muğla Sıtkı Koçman University, Türkiye, in 2018 and 2023, respectively. His Ph.D. thesis focused on post-shrinkage estimation in high-dimensional partially linear models. He is currently an Assistant Professor with the Department of Statistics, Muğla Sıtkı Koçman University, and a Visiting Researcher with the Probabilistic Machine Learning Group, Aalto University, Finland. He previously worked as a Postdoctoral Researcher at Aalto University within the EU Horizon AI-Mind Project, where he contributed to EEG-based neural fingerprinting and cognitive decline prediction using Bayesian reduced-rank regression. His research interests include high-dimensional data analysis, semiparametric regression, shrinkage estimation, survival analysis, Bayesian hierarchical modeling, and EEG/MEG connectivity analysis.