Check for updates

Received: 18 January 2023 | Revised: 11 September 2023 | Accepted: 21 September 2023

DOI: 10.1002/sim.9931

RESEARCH ARTICLE

Statistics in Medicine WILEY

Sample size and predictive performance of machine learning methods with survival data: A simulation study

Gabriele Infante[1,2] | **Rosalba Miceli**[2] | **Federico Ambrogi**[1,3]

[1]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy

[2]Unit of Biostatistics for Clinical Research, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

[3]Scientific Directorate, IRCCS Policlinico San Donato, San Donato Milanese, Italy

Correspondence
Federico Ambrogi, Department of Clinical Sciences and Community Health, University of Milan, Via Celoria 22, 20133, Milan, Italy.
Email: federico.ambrogi@unimi.it

Funding information
Italian Ministry of Education, University and Research, Grant/Award Numbers: PRIN 2017, prot. 20178S4EK9_004

Prediction models are increasingly developed and used in diagnostic and prognostic studies, where the use of machine learning (ML) methods is becoming more and more popular over traditional regression techniques. For survival outcomes the Cox proportional hazards model is generally used and it has been proven to achieve good prediction performances with few strong covariates. The possibility to improve the model performance by including nonlinearities, covariate interactions and time-varying effects while controlling for overfitting must be carefully considered during the model building phase. On the other hand, ML techniques are able to learn complexities from data at the cost of hyper-parameter tuning and interpretability. One aspect of special interest is the sample size needed for developing a survival prediction model. While there is guidance when using traditional statistical models, the same does not apply when using ML techniques. This work develops a time-to-event simulation framework to evaluate performances of Cox regression compared, among others, to tuned random survival forest, gradient boosting, and neural networks at varying sample sizes. Simulations were based on replications of subjects from publicly available databases, where event times were simulated according to a Cox model with nonlinearities on continuous variables and time-varying effects and on the SEER registry data.

KEYWORDS
machine learning, prediction, sample size, simulation, time-to-event

1 | INTRODUCTION

Prediction models are increasingly developed and used in diagnostic and prognostic studies. Within the latter, many applications involve survival outcomes, which implies accommodation of censored data by applying appropriate techniques. The Cox proportional hazards model is generally used in this context.

In recent years, interest in the use of machine learning (ML) techniques has increased (see for example the book of Gerds and Kattan[1]). A PubMed search by Year[2] using meSH terms such as "neural network*," "machine learning," "pattern recognition," "deep learning," and "deep neural network*," restricted to titles, shows a first increase in the 1990s with 35 publications per 100 000 in 1990 to reach 94 in 1996 and then a very rapid surge starting in 2016 with 127 publications per 100 000 and reaching 687 per 100 000 in 2020.

The distinction between statistical methods (SM) and ML methods is matter of debate. While probably there is no doubt in classifying neural networks or random forest as ML algorithms, at a closer look even classical risk scores, such as

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2023 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

the Framingham cardiovascular risk score,[3] are the result of an algorithm learned from the data. According to Beam and Kohane,[4] it is most useful to think to a continuum of methods (the ML spectrum) where the distinction is based on the amount of human guidance on the algorithm. In this sense, the Framingham cardiovascular risk score was the result of a strong interaction between statisticians and clinicians in order to select variables (together with their transformations and interactions) to be specified in the Cox model fitted to the data (low on the ML spectrum). A fully ML approach would have resulted if a large set of variables would have been provided to the algorithm without any pre-specification of the importance of the variables, and on the modeling of their nonlinearities and interactions. In some applications, involving for example high-dimensional data possibly coming from different sources, simply there is no alternative to ML methods and their usefulness is out of doubt (high on the ML spectrum). The framework investigated here is not on the high ML spectrum, as only a limited number of variables for the prediction model are considered. On the other hand, this is the setting in which it is less clear the potential advantage of a ML algorithm, as outlined by Smith et al.[5]

When dealing with prediction models with survival outcomes, at present, there are plenty of published ML tools specifically designed to take into account censored observations. For example, applications of neural networks to this framework dates back at least to 1992, Ravdin and Clark,[6] with developments in the 1990s and 2000.[7-11]

The recent scoping review by Smith et al[5] of studies comparing statistical and ML approaches for time-to-event data with simulation studies found 10 studies on the subject.

One aspect of special interest is the sample size needed for developing a survival prediction model. In clinical literature the rule of thumb of event per variable (EPV) based on the work of Peduzzi et al[12] is traditionally used. Recent work by Riley et al[13] has proposed new guidance for sample size calculation based on event rates, number of covariates to consider and the expected model performance. Such guidance refers to settings where covariate selection is not applied and the number of covariates does not exceed 30. Moreover, it works with traditional regression models and it is not an option for ML techniques.

For logistic regression van der Ploeg et al,[14] found that ML methods need far more EPV than SM to achieve stable results. In the simulation studies retrieved by Smith et al,[5] specifically targeted to censored data, seven articles found that the sample size did not impact the relative performances of the methods.

Taking the Riley's work[13] context as a start, the present study refers specifically to regression models for censored data and develops a time-to-event simulation framework to comparatively evaluate the performance of ML methods such as tuned random survival forest (RSF),[15] gradient boosting (GB)[16] and deep neural networks[17] at varying sample size.

To set the stage for more realistic model comparison, real-world datasets were used to implement the different simulation settings.

In the first setting a plasmode simulation was used. The German Breast Cancer Study[18,19] and the colon randomized trial for adjuvant treatment of colon cancer[20] were used to establish the survival time generating mechanisms. Specifically, event times were simulated according to a Cox model with nonlinearities on continuous covariates, possible time-varying effects and interactions between covariates.

A second set of simulations was based on the SEER registry data on breast and lung cancer, which was used to generate samples of increasing size by repeated resampling. Therefore, in this simulation setting, there is no a benchmark model and the survival data generating mechanism is unknown. For the SEER data on breast cancer, subsamples of the patients with diagnosis of breast cancer in 2010 and 2011 were used for model training while year 2012 was used for validation. For the SEER data on lung cancer patients with diagnosis of lung cancer between 2010 and 2015 were used. One fourth of the data was taken at random for validation purposes, while the remaining sample was used as a training set.

Simulations were run by varying the training set size from 600 to 9000 (first setting) or to 15 000 (second setting).

For each simulated dataset, hyper-parameters tuning was performed by choosing among the procedures available in the software of ML model packages.

The models performance were tested on the validation data using IPA and time-dependent Brier score.[21] Moreover, together with the absolute model performances, the learning curve of each technique, that is how fast each model reaches its optimal performance, was evaluated. The sample sizes calculated for building a prognostic model with traditional statistical regression models were then compared to the sample sizes needed by ML to achieve the same level of performances of traditional models. As in the book of Gerds and Kattan,[1] the data used to build the models are here referred to as "training set," and the portion of the data used to evaluate models' performance is referred to as "validation set." The latter wording is in contrast with ML literature, according to which the validation set is that used to select hyper-parameters (with or without resampling), while the set used to evaluate model performance is named "test set."

The article is structured as follows. In the methods Section 2, the simulation procedures, sample size calculations according to Riley et al,[13] model evaluation and the traditional and ML regression methods considered are described. In

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Section 3, the results of the simulations based on GBCS (Section 3.1), colon cancer (Section 3.2), and SEER data on breast and lung cancer (Section 3.3) are presented. In Section 4, results are discussed.

2 | METHODS

2.1 | Data generating mechanism

In the first setting, using the German Breast Cancer Study[18,19] and colon randomized trial,[20] the data were generated from a known model estimated on each of the two datasets. The simulation strategy was similar to that proposed by van der Ploeg et al[14] for logistic regression:

- We randomly divided the original dataset into a training set ($\frac{3}{4}$ of the patients) and a validation set for performance assessment.
- Two "artificial cohorts," training and validation sets were then obtained by separately replicating 20 times the training and validation sets.
- On the "artificial cohorts" we generated random survival times according to a flexible survival regression model estimated on the training set with nonlinear effects on the numerical covariates and possibly interactions and time dependent effects. In details, the generation of random survival times was done according to the methodology proposed by Crowther and Lambert[22] using the same baseline hazard estimated in the original data. The flexible parametric model proposed by Royston and Parmar[23] was used to easily incorporate time-varying covariate effects. Natural cubic splines with 3 degrees of freedom (df) were used for baseline estimation obtaining the model:

$$ln(-ln(S(t;\mathbf{z}))) = \gamma_0 + \gamma_1\, B_1(t) + \gamma_2\, B_2(t) + \gamma_3\, B_3(t) + \mathbf{z}\,\boldsymbol{\beta}. \tag{1}$$

  In this model time-varying covariate effects were obtained by adding *time* (or $ln(time)$) by covariate interactions. To simulate survival times, a numerical root finder to solve $S(t;\mathbf{z})) - u = 0$, where $u \sim U[0,1]$, was used through the `simsurv` function.[24]

  As regards censoring times generation, for GBCS data an exponential distribution was used; the exponential parameter was set to have approximately 75% censoring as in the observed data. For colon data the root finder returned errors in the generation of survival times and it was necessary to set the maximum event time. The truncation time was set to 10 years to have approximately 45% censoring similarly to the original data.
- Training sets were generated as samples of increasing sizes (varying from 600 to 9000), drawn from the artificial cohort training data.
- The predictions of the model, estimated using the training set, were evaluated on the validation set, as described in Section 2.3 below.

The simulation code is reported in the supporting material with the details of the generating models. The true model coefficients are copied for convenience in the `R` code file.

Another approach was used for the experiment with large SEER registry data on breast and lung cancer. Training and validation data were obtained by dividing the cohort in two groups based on the year of diagnosis for beast cancer data while a random split was used for lung cancer data. Subsamples from the training data of varying size (600-15 000) were used for model development while a subsample of validation data was used for validation. In this two experiments the true data generating mechanism is unknown.

2.2 | Sample size determination

For the sample size calculation the approach of Riley et al[13] was used, according to which sample size for survival endpoints is determined as

$$n = \frac{P}{(C-1)ln\left(1-\frac{R^2_{CS}}{D}\right)}, \tag{2}$$

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where P is the number of model coefficients, C is the targeted shrinkage of the parameter estimates, expressing the predictor effects, and R^2_{CS} is the Cox-Snell R^2 statistic. A first approach to determine the sample size, is choosing a desired shrinkage amount (ie, less than 10%) in order to minimize overfitting. Then R^2_{CS} can be established based on published regression models. When there is no pertinent a priori information, one proposal in Riley et al[13] is to use $R^2_{CS} = \epsilon * \max(R^2_{CS})$, with $\epsilon \in [0.1 - 0.2]$. Differently from continuous outcomes, with time-to-event outcomes, the $\max(R^2_{CS})$ is generally less than 1 and depends on the overall event rate (see the supplementary material S5 of Riley et al[13] for a calculation of $\max(R^2_{CS})$).

A second approach consists in targeting a small optimism in the apparent model fit. Such an optimism can be quantified by the absolute difference, δ, in the apparent model fit (observed on the training data) and the optimism adjusted model fit. According to Riley et al[25] the corresponding shrinkage factor to be used in formula (2) can be obtained as $C = \frac{R^2_{CS}}{R^2_{CS}+\delta \max(R^2_{CS})}$. The difference δ is chosen equal to 0.05.

A third approach is based on the length of 95% confidence interval of the cumulative incidence estimate at a specific time point.

In the first simulation setting, an estimate of the event rate was obtained using the original dataset. The number of covariates was calculated taking into account the need for estimating the baseline risk. Considering model (1), the four coefficients γ were added to the coefficients $\boldsymbol{\beta}$ used for covariate effect estimation.

For the SEER data, the event rate was determined using all the data reserved for training.

The sample size calculations were performed using `pmsampsize`[26] R package. The largest sample size among those calculated using the three methods was chosen.

For the SEER data, for computational reasons, a sample of the validation cohort was used. The sample size for validation data was calculated according to the procedure described in Reference 27, that is, in order to achieve a mean calibration slope standard error less than 0.05.

2.3 | Model evaluation

Evaluation of predictive performance were based on two different criteria, namely the Brier score[28] and the index of prediction accuracy (IPA)[21] at a specified time. The criteria were calculated using the validation data only.

Learning curves based on inverse power law functions were then used to compare the performances for increasing sample sizes. This was already done for evaluation of classifiers[29] or for machine learning in medical imaging research.[30] Letting $Y_r(N)$ the predictive performance of model r with sample size N, the learning curve is:

$$Y_r(N) = a + b * N^c, \tag{3}$$

where a is the maximum achievable performance for $N \to \infty$, b is the learning rate and c the decay rate. If $Y_r(N)$ is the Brier score at a specified time of interest obtained with sample size equal to N, then a is the minimum Brier score obtainable for very large N, b is connected with the Brier score at the minimum value of N, while the parameter c depends on how fast the Brier score decreases when N increases. Considering the transformation of $c = \frac{\ln(p)}{\ln(2)}$, the coefficient p can be interpreted as the learning percentage when the sample size doubles, that is, the ratio $\frac{Y_r(2N)}{Y_r(N)}$. The derivative of the learning curve is a monotonically increasing curve showing the added value, in terms of performance, of each new sample. At the beginning the derivative is increasing very steeply then it slows down for very large N.

Contrary to the Brier score, IPA is increasing with increasing N and derivatives are decreasing curves. Results in terms of IPA are reported in the supporting material.

2.4 | Statistical and machine learning methods for predicting time to event probabilities

We considered two traditional regression methods, namely the Cox model and an accelerated failure time (AFT) model with log-logistic distribution. Both Cox and AFT regression were specified with and without cubic natural splines (with 3 df) for continuous covariates. We considered four ML methods: boosting, both for the Cox and for the AFT with log-logistic distribution, random forest and a deep neural network.

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.4.1 | Cox and AFT boosting

The boosting algorithm originated in the ML community for classification and was subsequently applied to more general regression problems. Hastie, Tibshirani, and Friedman, in their book on statistical learning,[31] present boosting as "stagewise, additive modeling." Additive modeling means that the algorithm combines additively simple models (called base learners), while stagewise means that the combination acts sequentially, trying to improve the performance at each step. The number of steps is a fundamental hyper-parameter crucial for obtaining a balance between goodness of fit and overfitting. Boosting has relationships with penalized estimation and can be viewed as a functional gradient descent algorithm able to optimize a complex loss function of the covariates. The boosting algorithm is therefore specified through a loss function to be minimized (such as the squared-error or a likelihood-based loss function) and by different base learners, such as P-splines, B-spline bases or stumps.

The approach adopted here is the one based on gradient boosting for additive models presented by Bühlmann and Hothorn,[16] using an additive combination of P-splines for continuous covariates as base learners, and the optimization of the partial Cox likelihood for Cox boosting, and of the log-logistic likelihood for AFT modeling.[32] Specifically the function `gamboost`, using regularization and P-splines for continuous covariates, in the R package `mboost` was used.[33]

Hyper-parameter selection, namely the number of boosting steps, was performed by the function `cvrisk`, which uses cross-validation to estimate the empirical risk. Specifically, a 10-fold cross-validation was used.

2.4.2 | Random survival forest

RSFs[15] is a ML method that averages the terminal node statistics of survival trees to obtain an ensemble learner. In survival trees, predictions are based on binary recursive splits of the variable space. Starting from the root node, two daughter nodes are created dividing the sample in two groups (branches). The division is done by selecting over all possible covariates and their split values the one that maximize the survival difference between the two resulting groups using, for example, the log-rank test statistics. The process is continued for each of the branches until the terminal nodes of the tree contain a minimum number of events. Then, for each terminal node, corresponding to specific covariate patterns, the survival probability is calculated. Predictions for new subjects are obtained by dropping them down the tree, that is, by finding their corresponding terminal nodes. As predictions from a single tree are often poorly generalizable, the general idea of combining base learners is effective in improving prediction performances. Random forest is an ensemble learner that combines the predictions of survival trees, the base learners, obtained through randomization. The randomization occurs into the learning process in two forms: to grow a survival tree the data are randomly drawn into a bootstrap sample; a randomly selected subset of variables is chosen as candidates for splitting during the grow stage at each node of the tree. The framework of fast unified random forests for survival, regression, and classification (RF-SRC)[34] was used in this work. In particular, RSFs[35] were implemented through the function `rfsrc.fast` in the R package `randomForestSRC`,[36] which provides a good approximation and fosters computational speed. The two key parameters selected for tuning, through the function `tune`, were (*i*) the minimum number of events in each terminal node, that is, *nodesize*, and (*ii*) the number of candidate variables selected for splitting a node, that is, *mtry*. The tuned parameters were used in a 5000-trees forest on train data to predict the ensemble survival outcome at 250 unique time points (adjusted by setting *ntime*). Any other parameter was set automatically as provided for in `rfsrc.fast` default setup. Hyper-parameters were tuned using out-of-bag observations.

2.4.3 | *PC-hazard* deep neural network

In neural networks the basic computation is performed by the neuron. The neuron has inputs and outputs: the inputs are combined, usually a weighted sum, and transformed using a nonlinear function (ie logistic or RELU functions). Transformations are called activation function as the first use, by McCulloch and Pitts in the 40,[37] provided an activation signal only if the linear combinations was greater than a threshold. In feed-forward neural networks (the ones mostly used for prediction models), neurons are organized in layers. Layers are then connected taking as input the output of the preceding layer. The first layer, or input layer, takes linear combination of the original variables, while the output layer has the task to output the predicted values. The input layer communicate with the first hidden layer, and then the first hidden layer to the second hidden layer, until the output propagating the signal in forward direction. A neural network is a highly

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

nonlinear model able to account for multiple interactions among the inputs. The task of finding the neurons weights for prediction is done by minimizing a loss function, that is, the squared-error or a likelihood-based loss function. As the neural network is a very complex model, the minimization is a complex task and is done through the backpropagation algorithm, which involves a clever and efficient way to update the weights according to the derivatives of such a complex function. As neural networks are very flexible it is very important to control for overfitting. Many tools are available, for example regularization applied to neuron weights, dropout (during training, some number of layer outputs are randomly ignored or "dropped out"), limiting the number of neurons in layers and the number of layers, the learning rate (which control the amount of update of the weights at each iteration).

Applications of neural networks to survival data are not new in biomedical literature. The pioneer work dates back at least to 1992.[6] Many successive implementations were based on discrete times or on some sort of discretization, making the model easily adapted to standard regression techniques.[7-10] In fact, as there is a correspondence between the discrete-time survival likelihood and the Bernoulli likelihood, and between a piecewise exponential model and the Poisson likelihood, it is possible to use standard regression algorithms for generalized linear models to fit a neural network for censored data. It is worth mentioning also a proposal based on pseudo-values,[38] which is another form of time discretization that makes possible to use standard software.

Recent developments in ML with software such as TensorFlow[39] or Keras[40] made possible to extend such approaches in the field of deep neural networks.[17,41,42] Deep neural networks have many hidden layers, then increasing the ability to learn nonlinearities and interactions.

We used the comprehensive `mlr3 R` package, which offers a modern implementation in R through a connection to Pyhton.[43] Trials with the data used for the simulations showed good performances of the *PC-hazard* method[17] in comparison with the other implemented approaches. A piecewise constant hazard is assumed in predefined time intervals. An equidistant grid over 10 cut points was used as for the default method. Five hyper-parameters were tuned: the dropout rate; the weight decay; the learning rate; the number of nodes per layer; the number of hidden layers. A random search was used within the hyper-parameters space with a total of 100 evaluations using a three-fold cross validation. For SEER lung cancer data 250 random searches were performed. The C-index was used to select the optimal hyper-parameters combination, as usual in many applications of neural networks to survival data. It is worth saying that `mlr3` offers the possibility to perform hyper-parameters selection using other metrics, such as the Brier score, and this option is preferred if not only discrimination but also calibration is of interest. As in this work, Brier score is used to compare the different models, in the supporting information, the results of *PC-hazard* with hyper-parameters selection using Brier score for the SEER lung cancer data are reported.

The implementation and the tuning of the deep neural network was performed through the `mlr3` library using a random search on a predefined hyper-parameters domain.

3 | RESULTS

In this section, the results of the simulations based on GBCS data, colon cancer data, and SEER data on breast and lung cancer are presented in terms of Brier score. Results in terms of IPA are in accordance to those of the Brier score and are reported in the supporting material.

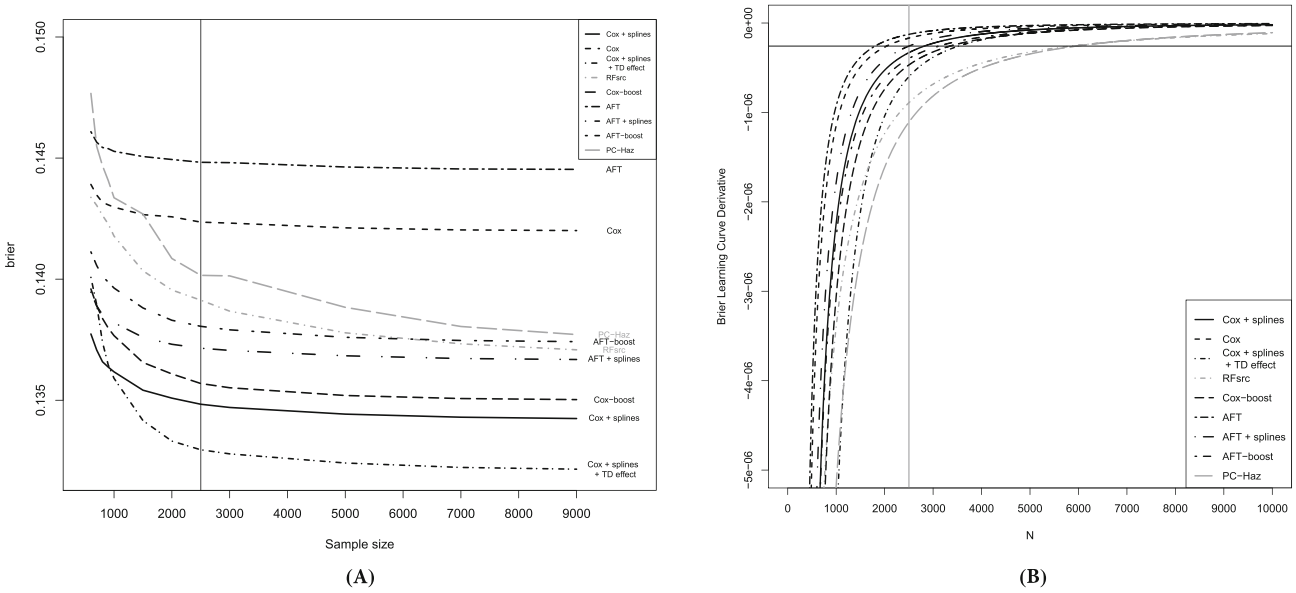
3.1 | German Breast Cancer Study data

In the German Breast Cancer Study (GBCS) a total of 720 patients with primary node positive breast cancer were recruited between July 1984, and December 1989.[18] These data were used to illustrate methods for building prognostic models[44] and also for illustrating external validation of a prognostic model.[45]

The estimated monthly event rate is 0.006. Considering a mean follow-up (restricted at 72 months) of 50 months, a maximum R^2_{CS} equal to 0.72 was calculated.

The generating model was based on the available covariate information, namely age at diagnosis (years), menopausal status, hormone therapy, tumor size (mm), tumor grade (1-3), number of nodes, number of progesterone and estrogen receptors. Numerical variables (nodes excluded) were modeled with natural cubic splines with 3 df, and tumor grade with two dichotomous variables. This amounts to a total of 17 coefficients. In addition, to model the baseline hazard, 3 coefficients were needed plus the intercept, for a total of 21 coefficients.

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



(A) (B)

FIGURE 1 GBCS cancer data. (A) Brier score as a function of sample size for the different regression models. (B) Derivatives of the estimates of the learning curves for the different methods. The horizontal line is at the intersection of the derivative of AFT+splines with Riley's sample size (vertical gray line).

The function `pmsampsize` was used with the following specifications: expected R^2_{CS} equal to 0.1 of max(R^2_{CS}), $P = 21$, a prediction horizon of 72 months. Based on these settings, and fixing a 10% shrinkage, generated a target sample size of 2501 subjects, while the target of an absolute difference $\delta = 0.05$ in the model's apparent and validated R-squared generated 548 subjects. The target sample size was therefore 2501.

In order to highlight the abilities of ML methods to discover interactions and possibly time dependent effects, we included in the generating model these two interactions: interaction between estrogen receptors levels (modeled with a natural cubic spline with 3 df) and the logarithm of time (linear), resulting in a time-dependent estrogen effect; interaction between tumor size (modeled with a natural cubic spline with 3 df) and grade. The generating model has thus a total of 30 coefficients and the calculation with `pmsampsize` generated a target sample size of 3573 patients.

The generated survival data had a monthly event rate of 0.001 with about 75% censoring, as in GBCS data, obtained using an exponential distribution for the censoring times.

The performance of the different models according to the the Brier score, calculated at 6 years, are reported in Figure 1A for different sample sizes.

The absolute benchmark model here is represented by the true model (Cox + Splines + TD), which has the better performance for $N > 1000$. Cox regression with splines (Cox + splines) and the boosted Cox model are performing very well. These models are all proportional hazards and model nonlinear effects using natural spline functions. A fairer benchmark for ML methods is the AFT model with log-logistic distribution using restricted cubic splines to account for nonlinear effects. The same benchmark was used for all the simulations. The simple Cox or AFT regression without modeling of nonlinear effects showed the worst performances.

At the sample size of 2384 patients random forest regression and *PC-hazard* show the worst performances (except for the simple Cox and AFT regression). The Brier score of AFT with splines at sample size 2384, is reached by random forest at a sample size of about 8000, while *PC-hazard* requires even more. Random forest regression shows better performances than boosted AFT at a sample size of about 6000 patients

The result of the fitting of the learning curve for each regression model considered is reported in Supplementary Table 1. Random forest and *PC-hazard* have the smallest absolute value of the c coefficient that summarizes the improvement for growing sample size. The simple Cox and AFT regressions have the highest improvement for small samples. In particular it can be seen how the percentage reduction of the Brier score for doubling of the sample size, that is, p, is inversely correlated with the complexity of the model. Random forest have the slowest learning rate.

Considering the derivative of the learning curves in Figure 1B it is possible to compare the different approaches as far as the learning rate. In Table 1 are reported the required sample size per method obtained by setting the derivative of the

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 GBCS: Sample size per method obtained by setting the derivative of the learning curve reached by AFT regression with cubic natural splines at determined sample size as reference.

	Cox	Cox + splines	AFT	AFT + splines	Cox-boost	AFT-boost	PC-Haz	RFsrc
1	2040	2818	1801	2501	3349	3116	5954	5847
2	82%	113%	72%	100%	134%	125%	238%	234%

learning curve reached by AFT regression with cubic natural splines at $N = 2501$ as reference. In order to reach the same slope of the learning curve reached by the reference model, random forest and *PC-hazard* requires more than double the sample size, while boosting methods need approximately an additional 40% of the sample size.

In Supplementary Figure 1 the estimates of the learning curves for the different methods are reported together with their standard errors. The variability of the estimates is comparable among the different models. In Supplementary Table 1, parameter estimates are also reported.

3.2 | Colon cancer data

We used colon cancer data from a trial of adjuvant chemotherapy for colon cancer comparing Levamisole and Levamisole plus 5-FU (a chemotherapy agent),[20,46] available in the *survival R* package.[47] The re-analysis presented by Eng and Seagle[48] explored the complex pattern of interaction between age and treatment using RMST. In fact it appeared that age was significantly associated with relapse in the Levamisole plus 5-FU arm but not the Levamisole alone arm.

In this study the monthly event rate is 0.01.

Considering a mean follow-up (restricted at 72 months) of 68 months a maximum R^2_{CS} equal to 0.85 was calculated.

There are two records per person, one for recurrence and one for death. Only the data for recurrences were considered.

The generating model was based on the available covariate information, namely age at diagnosis (years), sex, treatment, obstruction of colon by tumor, perforation of colon, adherence to nearby organs, number of lymph nodes with detectable cancer, differentiation of tumor (well, moderate, poor), Extent of local spread (submucosa, muscle, serosa, contiguous structures), time from surgery to registration (short, long) and more than four positive lymph nodes. Numerical variables were modeled using truncated power cubic splines with 1 knot (at 50 for age and at 2 for the number of nodes). Extent of local spread was modeled with a dichotomous variable (3 and 4 vs 1 and 2). As suggested by Eng and Seagle[48] an interaction between treatment and age was also inserted. This amount to a total of 19 coefficients. In addition, to model the baseline hazard, 3 coefficients were needed plus the intercept, for a total of 23 coefficients.

The function `pmsampsize` was used with these settings: R^2 equal to 0.1 of the maximum R^2_{CS}, 23 parameters to be fitted and a prediction horizon of 72 months. The calculation based on an expected shrinkage of predictor effects equal to 10% results in 2306 subjects, while the target of an absolute difference of 0.05 in the model's apparent and validated R-squared results in 507 subjects. The target sample size was therefore 2306.

The generated survival data had a monthly event rate is 0.008 with about 45% censoring, as in colon cancer data, obtained using an administrative censoring at 10 years of follow-up. This choice was due to the fact that the `simsurv` function returned errors in the generation of survival times unless setting the `maxt` option to censor large event times.

In Figure 2A the average Brier score at different sample sizes for the considered regression models is reported. The spread of the Brier scores among the different models is more limited than with the GBCS data. In this setting the simple Cox and AFT regression, without nonlinear effects, perform very well and near to the true model (Cox + splines + interaction). It was not possible, to the best of our abilities, to find a tuning for *PC-hazard* with reasonable performances. The Brier scores were much higher than that of the other regression approaches and the IPA were negatives. It is therefore not reported in Figure 2A.

At the sample size of 2306 patients, random forest shows the worst performance. The Brier score of AFT with splines at sample size 2306, is reached by random forest at a sample size of about 4500 while that of AFT at about 6000. Random forest regression becomes competitive with respect to boosted AFT at a sample size of about 5000 patients.

The result of the fitting of the learning curve for each method considered is reported in Supplementary Table 2. The learning curve is reported also for *PC-hazard* to see how the algorithm improves its performance. Also in this case random forest and *PC-hazard* have the smallest absolute value of the c coefficient that summarizes the improvement for growing sample size. The percentage reduction of the Brier score for doubling of the sample size, that is, p, is lower for random

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

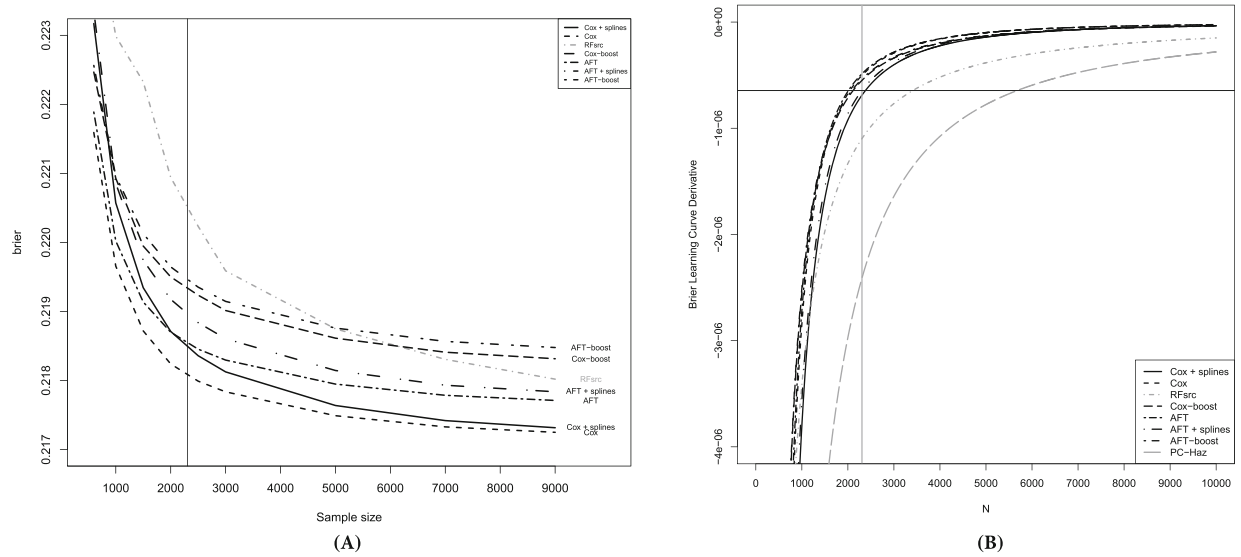


FIGURE 2 Colon cancer data. (A) Brier score as a function of sample size for the different regression models. (B) Derivatives of the estimates of the learning curves for the different methods. The horizontal line is at the intersection of the derivative of AFT+splines with Riley's sample size (vertical gray line).

TABLE 2 Colon: Sample size per method obtained by setting the derivative of the learning curve reached by AFT regression with cubic natural splines at determined sample size as reference.

Cox	Cox + splines	AFT	AFT + splines	Cox-boost	AFT-boost	PC-Haz	RFsrc
2045	2382	2012	2306	2114	2087	5679	3412
89%	103%	87%	100%	92%	91%	246%	148%

forest and *PC-hazard* have the slowest learning rate. In Supplementary Figure 2 the estimates of the learning curves for the different methods are reported together with their standard errors. Also in this application, the variability of the estimates is comparable among the different models.

The derivative of the learning curves are reported in Figure 2B. In Table 2 are reported the required sample size per method obtained by setting the derivative of the learning curve reached by AFT regression with cubic natural splines at $N = 2306$ as reference. In order to reach the same slope of the learning curve reached by the reference model, *PC-hazard* requires more than double the sample size, random forest needs about 50% more observations, while boosting methods need approximately a 10% less records of the calculated sample size.

3.3 | SEER registry data

The data extraction was performed on SEER Research Data 2000-2018[49] using the SEER*Stat software version 8.3.9.2.[50]

3.3.1 | Breast cancer

The selection criteria included female patients who were at least 18 years of age and who received a primary malignant breast cancer diagnosis, staged as localized or regional, between 2010 and 2012. The exclusion criteria aimed to select only complete records or factors having levels with frequency >1%. Retrieved variables were patient' characteristics, that is, age in years and race group (three-levels factor: black, other, white), and tumor characteristics, that is, year of diagnosis (2010 and 2011 in the train set, 2012 as validation set), size in millimeters (for those cases reporting "less than × cm," it was chosen a value of × cm minus 5 mm as reasonable), tumor grade (three-levels factor: 1, 2, 3), AJCC 6th ed. tumor stage (three-levels factor: I, II, III), total number of regional lymph nodes that were removed and examined by the pathologist (at least one) and total number of those that were found to contain metastases, breast tumor subtype (four-levels factor:

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Triple negative, HER2 enriched, Luminal A, Luminal B). In addition, each record included the survival status and time in months from diagnosis. A total of 106 823 cases were selected, 70 175 for training set and 36 648 for validation set.

In this study the monthly event rate in the training set was 0.002 with a 83% censoring. Considering a mean follow-up (truncated at 60 months) of 60 months a maximum R^2_{CS} equal to 0.52 was calculated.

The total number of positive nodes is a variable with a spike at zero as some patients have no positive nodes. According to one of the proposal in Reference 51, total number of positive nodes was modeled with two variables: one binary variable, modeling the presence (1) or absence of positive nodes, and a continuous variable with the number of positive nodes. Modeling numerical variables with natural cubic splines with 3 df, and the baseline hazard with 3 coefficients plus the intercept, the total number of coefficients was 21.

The function `pmsampsize` was used with these settings: the expected R^2 equal to 0.1 of the maximum R^2_{CS}, 21 parameters to be fitted and a prediction horizon of 60 months. The calculation based on an expected shrinkage of predictor effects equal to 10% results in 3501 subjects, while the target of an absolute difference of 0.05 in the model's apparent and validated R-squared results in 775 subjects. The target sample size was therefore 3501.

Following the approach described in Riley et al,[27] a validation sample size of 6000 patients was identified to have a mean calibration slope standard error less than 0.05. A validation sample size 6000 patients was then drawn at random from the test set, for computational reasons, and used for validation in all simulations runs. The validation set had a monthly event rate is 0.002 with about 87% censoring.

In Figure 3, the distributions of the hyper-parameters selected through the simulations are reported. Regarding *PC-hazard*, for increasing sample sizes there was a decrease in the number of nodes per layer and an increase in weight decay while the amount of dropout and the learning rate were approximately stable. Rarely more than one hidden layer was selected (not shown). Boosting algorithms show a clear increase in the number of step values. Random forest shows a decrease in the frequency of high mtry values and a slight increase of node size for increasing sample sizes.

In Figure 4A, the average Brier score at different sample sizes for the considered regression models is reported. In this setting AFT regression with splines, or boosted, performs very well. At the sample size of 3501 patients, *PC-hazard* shows the worst performance while random forest performs better than the simple Cox and AFT regressions without nonlinear effects. The Brier score of AFT regression with splines at sample size of 3501, was reached by random forest at a sample size of about 6000 and *PC-hazard* at about 7000. Random forest regression and *PC-hazard* reaches the same performances of boosted AFT at the last sample size of 15 000 and outperform AFT regression with splines. A larger sample size was also evaluated, namely 20 000 to see for a further possible improvement of *PC-hazard*. In fact the performance at 20 000 was practically the same of that at 15 000 samples.

The results in terms of learning curve for each regression model investigated are reported in Supplementary Table 3. Random forest has the smallest absolute value of the c coefficient that summarizes the improvement for growing sample size, while *PC-hazard* has the largest. The percentage reduction of the Brier score for doubling of the sample size, that is, p, was the lowest for *PC-hazard* and the highest for random forest. However *PC-hazard* has the worst performance for small sample sizes (top-left of Figure 4A) as demonstrated by the very high b coefficient. In fact, *PC-hazard* has very poor performances with low sample sizes of 600 (that was excluded from the evaluations) and 1000. Considering the sample size of 1000, in 30 out of 300 simulations the selected learning rate was about a half the one selected in the other simulations with a Brier score greater than 0.25 and a negative value for IPA. These 30 simulations, at sample size 1000, were therefore excluded from the mean calculation. In Figure 4B the derivative plot of the estimates of the learning curves for the different methods are reported.

In Table 3 are reported the required sample size per method obtained by setting the derivative of the learning curve reached by AFT regression with cubic natural splines at $N = 3501$ as reference. In order to reach the same slope of the learning curve reached by the reference model, *PC-hazard* requires more than twice the sample size, random forest needs about 50% more observations, while boosting methods need quite the same sample size.

In Supplementary Figure 3, the estimates of the learning curves for the different methods are reported together with their variability. In this application also, the variability of the performances of ML methods is similar to that of the other methods.

3.3.2 | SEER lung cancer

The selection criteria included patients who were at least 18 years of age and who received a primary malignant lung cancer diagnosis between 2010 and 2015. The exclusion criteria aimed to select only complete records. Retrieved variables

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

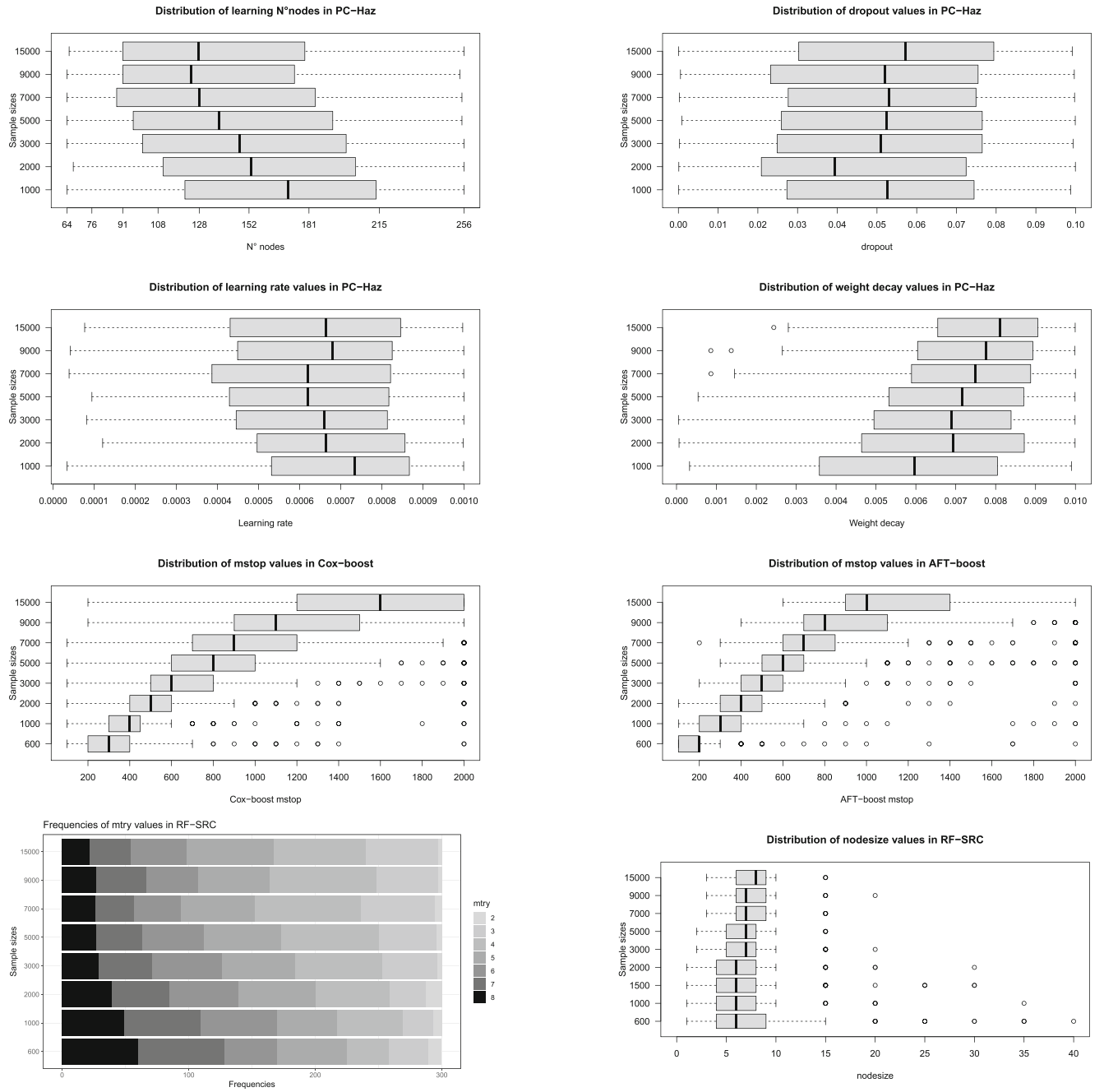


FIGURE 3 SEER breast: Hyperparameters selection.

were patient' characteristics, that is, age in years, sex, and race group (three-levels factor: black, other, white), and tumor characteristics. The tumor size, in millimeters, was analysed as numeric variable (for those cases reporting "less than × cm," it was chosen a value of × cm minus 5 mm as reasonable), the same way as the total number of regional lymph nodes that were removed and examined by the pathologist (at least one) and total number of those that were found to contain metastases. Other tumor-specific features were employed as categorical, that is, the laterality of primary site (left lung or right lung only), tumor grade (four-levels factor: 1, 2, 3, 4), AJCC 7th ed. tumor stage (four-levels factor: I, II, III, IV), three-levels factor to distinguish regional, localized or distant disease, and histology (seven-levels factor: adenomas and adenocarcinomas, squamous cell neoplasms, epithelial neoplasms NOS, acinar cell neoplasms, cystic or mucinous or serous neoplasms, complex epithelial neoplasms, and other types). The last level of histology gathered ICD-O-3 codes 8000-8009, 8120-8139, 8430-8439, 8500-8549, and 8930-8999. In addition, each record included the survival status and time in months from diagnosis.

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

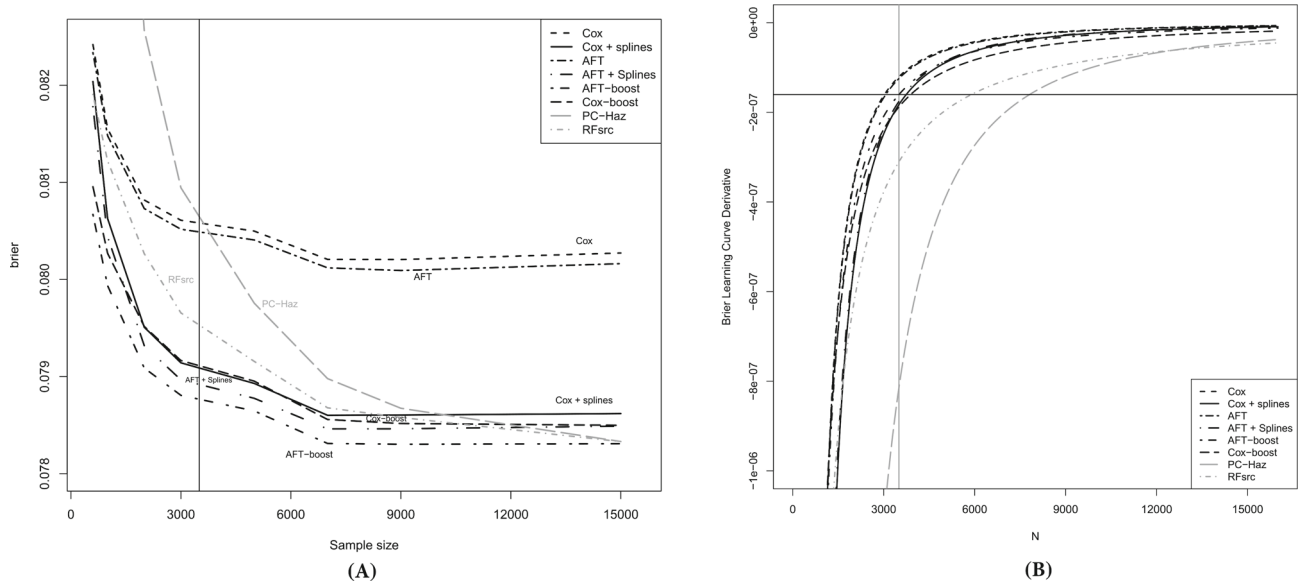


FIGURE 4 SEER breast cancer data. (A) Brier score as a function of sample size for the different regression models. (B) Derivatives of the estimates of the learning curves for the different methods. The horizontal line is at the intersection of the derivative of AFT+splines with Riley's sample size (vertical gray line).

TABLE 3 SEER breast: Sample size per method obtained by setting the derivative of the learning curve reached by AFT regression with cubic natural splines at determined sample size as reference.

Cox	Cox + splines	AFT	AFT + Splines	Cox-boost	AFT-boost	PC-Haz	RFsrc
2882	3573	2925	3501	3663	3321	7467	5484
82%	102%	84%	100%	105%	95%	213%	157%

A total of 33 693 cases were selected, randomly divided in 75% (n = 25 270) for the training set and 25% (n = 8423) for the validation set.

In the training data the monthly event rate was 0.009 with a 55% censoring. Considering a mean follow-up (truncated at 36 months) of 36 months a maximum R^2_{CS} equal to 0.75 was calculated.

The same strategy used for the data on breast cancer was used to model the total number of positive nodes with the spike at zero. Modeling numerical variables with natural cubic splines with 3 df, plus the 3 coefficients needed for the baseline hazard and the intercept, this amounts to a total of 33 coefficients.

The function `pmsampsize` was then used specifying the expected R^2 equal to 0.1 of the maximum R^2_{CS}, 33 coefficients to be fitted and a prediction horizon of 36 months. Considering these settings and the 10% target of an expected shrinkage of predictor effects, 3801 subjects were calculated. The target of an absolute difference of 0.05 in the model's apparent and validated R-squared resulted in 832 subjects. The target sample size was therefore 3801.

We decided to use all the 8423 patients for validation as, according to the approach in Riley et al,[27] a mean calibration slope standard error less than 0.05 is achieved with a sample size of approximately 5000 patients.

In the validation survival data, the monthly event rate was 0.009 with about 56% censoring.

In Figure 5, the distributions of the hyper-parameters selected through the simulations are reported. For increasing sample sizes, *PC-hazard* choose more than one layer more often, the number of nodes per layer and the dropout values do not follow a specific trend, while a relevant decrease in weight decay and learning rate are appreciable. Boosting algorithms show a clear increase in the number of step values. Random forest shows an increase in the frequency of high mtry values and a slight increase for increasing sample sizes of nodesize.

In Figure 6A, the average Brier score at different sample sizes for the considered regression models is reported. Boosting, both with Cox and AFT regressions shows the lowest Brier scores. The difference between Cox and AFT regressions with and without splines is lower than that observed with breast cancer data. At the sample size of 3801 patients, random forest and *PC-hazard* show the worst performances. For larger samples, *PC-hazard* performs slightly better than

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

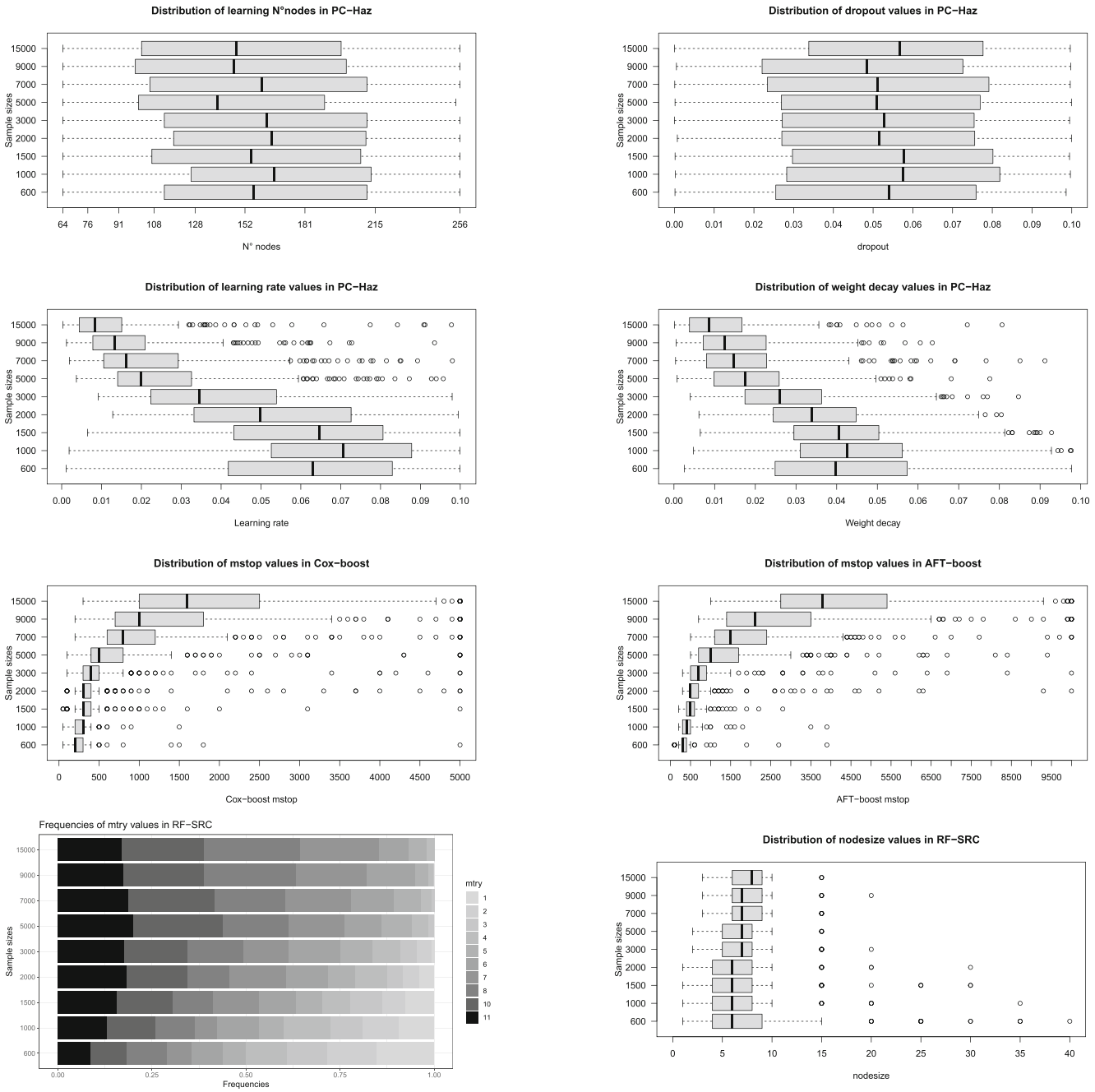


FIGURE 5 SEER lung: Hyperparameters selection.

random forest. For increasing sample sizes, random forest and *PC-hazard* reach the Brier score of of AFT regression without splines, while never reach the performance of AFT regression with splines. The Brier score of AFT without splines at sample size of 3801, is reached by *PC-hazard* (although the learning curve is fluctuating) and by random forest at about 8000 and 9000 respectively. Random forest regression and *PC-hazard* reach approximately the same performances of AFT without splines at the last sample size of 15 000 and 9000, respectively.

The results in terms of learning curve for each regression model considered are reported in Supplementary Table 4. Random forest has the smallest absolute value of the c coefficient that summarizes the improvement for growing sample size, while *PC-hazard* has the largest. The percentage reduction of the Brier score for doubling of the sample size, that is, p, is the highest for random forest and the lowest for *PC-hazard*. However *PC-hazard* has the worst performance for small sample sizes as demonstrated for the very high b coefficient.

In Figure 6B the derivative plot of the estimates of the learning curves for the different methods are reported.

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

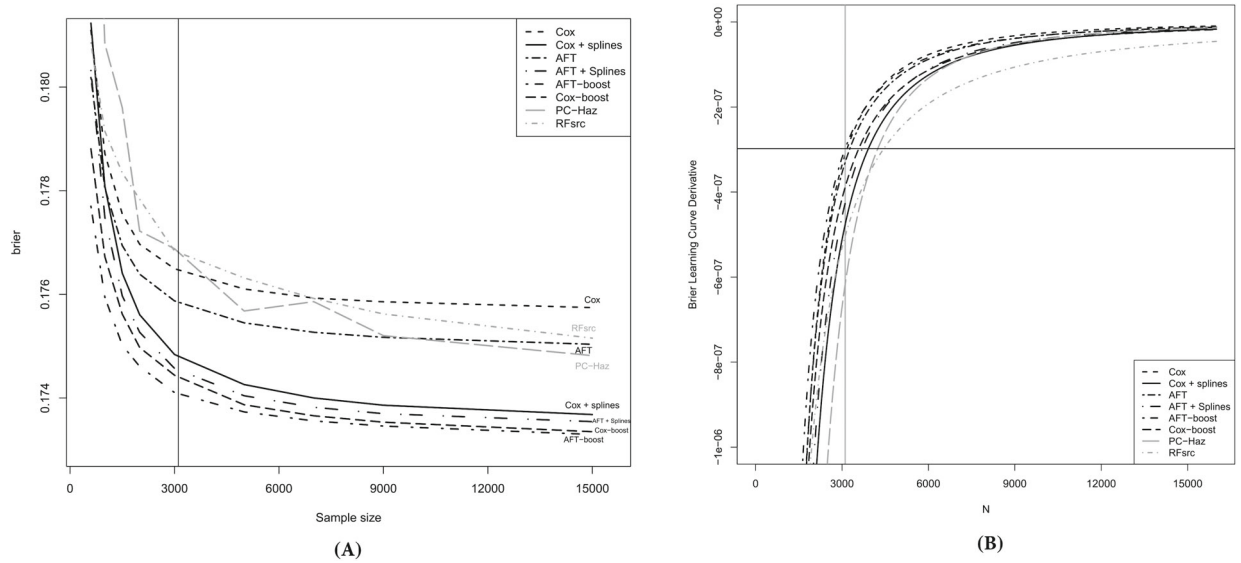


FIGURE 6 SEER lung cancer data. (A) Brier score as a function of sample size for the different regression models. (B) Derivatives of the estimates of the learning curves for the different methods. The horizontal line is at the intersection of the derivative of AFT+splines with Riley's sample size (vertical gray line).

TABLE 4 SEER lung: Sample size per method obtained by setting the derivative of the learning curve reached by AFT regression with cubic natural splines at determined sample size as reference.

Cox	Cox + splines	AFT	AFT + Splines	Cox-boost	AFT-boost	PC-Haz	RFsrc
2687	3287	2741	3110	2962	2586	3644	3489
86%	106%	88%	100%	95%	83%	117%	112%

In Supplementary Figure 4, the estimates of the learning curves for the different methods are reported together with their variability. The variability of the performances of ML methods is larger than that of the other methods. *PC-hazard* has the largest variability.

In Table 4 are reported the required sample size per method obtained by setting the derivative of the learning curve reached by AFT regression with cubic natural splines at $N = 3801$ as reference. In order to reach the same slope of the learning curve reached by the reference model, *PC-hazard* and random forest require an additional 17% and 12%, respectively, of the calculated sample size, while boosting methods need approximately a 10% less records of the calculated sample size.

3.4 | Sample size

The different simulations showed some specificities but also common traits. Trying to summarize a common message, in Table 5 it is reported the sample size needed by RF and NN to reach the performance of the reference model at Riley's sample size. Moreover the sample size needed by RF and NN to reach the same learning curve slope of the reference model at Riley's sample size is reported. As reference it was taken the level of performances of AFT regression with splines for nonlinear effects. *PC-hazard* and random forests need at least from 2 to 3 times the sample size calculated according to Riley's method to achieve the performance of the reference. To achieve the learning rate of the reference, *PC-hazard* and RF need a double sample size.

4 | DISCUSSION

The use of risk prediction models in medicine can be of help to clinicians in many decision making tasks, for example, patient counseling, treatment choice or selection of patients eligible for a clinical trial. There is a large body of literature

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 Summary of the sample sizes needed by *PC-hazard* and random forest to reach the brier score and the slope of the learning curve of the reference model at Riley's sample size estimate.

			Brier score		Learning curve slope	
Dataset	**Reference model (RM)**	**Riley's sample size**	***PC-hazard***	**RF**	***PC-hazard***	**RF**
GBCS	AFT + splines	2501	>9000 (3.8)	≈8000 (3.4)	≈6000 (2.4)	≈5900 (2.3)
Colon	AFT + splines	2306	NA	≈4500 (2.0)	≈5700 (2.5)	≈3400 (1.5)
SEER breast	AFT + splines	3501	≈7000 (2.0)	≈6000 (1.7)	≈ 7500 (2.1)	≈5500 (1.6)
SEER lung	AFT + splines	3801	>15 000 (4.0)	>15 000 (4.0)	≈3600 (1.2)	≈3500 (1.1)

on model development using multivariable regression, and their validation and application (see for example Moons et al[52] and references in the series about prognosis and prognostic research).

In a classic paper contrasting two approaches devoted to the solution of prediction problems, that is, statistical and algorithmic modeling, the latter most widely known as ML, Breiman describes how these two different paradigms were generated by two distinct research communities or "modeling cultures."[53] One of the effects of the coexistence of these two communities are conflicting opinions, such as the claims that ML methods outperform traditional models and, conversely, that there is no advantage of using ML over traditional statistical models.

The distinction between traditional statistical models and ML models is itself matter of debate. In the article by Beam and Kohane[4] the distinction is not clear-cut, and a ML spectrum is defined by the trade-off between the predictive algorithm a priori specification against the ability of the algorithm to learn from the (possibly big) data. At the very bottom of the spectrum is the human decision making and at the top the complex deep learning methods fitted on high-dimensional data. The possibility to learn from the data is obviously linked to the amount of data available.

In a work comparing ML techniques and traditional statistical methods in cardiovascular diseases by Wallisch et al,[54] the authors stated that, as long as sample size is sufficient, predictive accuracy is not largely affected by the choice of algorithm. Again according to Beam and Kohane,[4] while risk calculators developed using traditional techniques require sample size of the order of 10^4, successful applications of ML were obtained with sample size of order 10^6 or more. Such a huge amount of data is required to carefully tune the hyperparameters used by ML methods in order to balance the trade-off between the reduction of the training error and the ability to generalize predictions to new data (the bias-variance trade-off).

One of the issues regarding the use of ML is the guidance on the minimum sample size needed to develop a predictive model. For traditional statistical models, the rule of thumb based on the event per variable is generally used but more recent proposals are available combining information on the event rate, the number of variables considered and the expected algorithm performance.[13] However, such considerations cannot generalize to ML. Existing work on the required sample size for ML relies mainly on logistic regression[14] or image analysis.[30]

To have insights on the sample size needed by ML methods, one can resort to data simulation. In a recent publication by Smith et al,[5] a scoping methodological review on risk prediction for survival outcomes which compares statistical and ML approaches through simulations, it has been underlined the limited number of articles on the topic and the focus on ML, with the frequent comparison with the Cox model only, in its simplest formulation. Frequently the simulation setting was favorable to ML methods, for example focusing on high-dimensional data. Many articles reported that sample size did not impact the relative performance of the methods. This finding may be due to the limited number of model investigated. The results reported in this work shows that the relative performance of the models is highly dependent on the sample size.

The main difficulty in building a simulation framework to have insights on the sample size needed by ML methods is the heavy computational burden when learning from the data, especially that related to cross-validation (or resampling method) for hyper-parameters tuning. This fact, coupled with the need of huge datasets, makes simulations very slow and difficult to set up.

To circumvent this problem, van der Ploeg et al[14] used default algorithms settings in their simulation study, thus avoiding the burden of hyper-parameters tuning. Such a procedure has the disadvantage of using ML techniques in a suboptimal way as the hyper-parameters tuning is a key feature for a proper learning from the data. As an alternative to simulation, in the search for a neutral comparison among different methods,[55] a competition of experts framework, as done by Wallish et al,[54] can be engaged.

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We adopted two different simulation strategies according to Morris et al.[56] In the first two simulations we defined a data-generating mechanism with a true model benchmark, while in the last two simulations we draw repeatedly random samples from specific datasets and we compared the different approaches in presence of an unknown data-generating mechanism. Simulations were performed at different sample sizes. The data generating mechanism is very important as suggested by Austin et al.[57] In their study, comparing traditional and ML methods using different data generating mechanisms, no method was uniformly superior to the other. Therefore, in the first two simulations, we focused more on learning rate (how fast each model reaches its optimal performance) than on the absolute performance of the different methods.

In our work, we made heavy use of parallel computation with a pre simulation setup for defining a plausible domain for hyper-parameters optimization search. For example, when using boosting, we used some example simulated datasets to perform a cross validation procedure to determine the optimal number of boosting steps and decide a maximum number of boosting steps to be used in the simulations. For neural networks the setup was particularly difficult as we had to decide the range of four hyper-parameters and, most importantly, the number of random searches to be performed. In principle, it would be wise to specify a very large number of random searches but this will dramatically increases computational times of the simulation procedure. We found good neural network configurations for GBCS and SEER breast cancer data, while we did not succeed in finding a neural network configuration with good performance for colon cancer data, even increasing the number of random searches. Considering SEER lung cancer data, increasing the number of random searches from 100, as in breast cancer applications, to 250 was instead effective for achieving good performances. The main distinction between breast cancer data and the other data considered is the number of continuous variables: in colon cancer data, only age and the number of lymph nodes with detectable cancer were present with limited possibilities for nonlinear effects. It is possible that ML regression is more effective in settings where continuous variables are involved.

Another difficulty was to deal with censored data. While for binary response variables, there is huge availability of ML methods to be compared with logistic regression, it is more difficult to find ML regression able to properly treating censored data. Many ML methods exploit the Cox likelihood and this may limit the potentiality of the methods, especially in settings where the standard Cox regression can be used. For this reason we tested other models not relying on Cox proportional hazard likelihood, such as boosting based on a non-proportional hazard regression model, the random forest and the *PC-hazard* method.

The approach we used to first calculate the minimum sample size for building a prognostic model with traditional statistical regression techniques, along the lines suggested by Riley et al.[13] This calculated sample sizes was then compared to that needed by ML techniques to achieve the same learning rate or the same performance level of traditional methods.

The setting investigated was therefore the same of Riley et al,[13] considering a number of predictors, or better coefficients in terms of standard regression models, not exceeding 30. Variable selection was therefore not considered. This setting is the one where potential benefits of ML are more uncertain while, probably, ML methods become much more appealing in more complex scenarios with many putative prognostic factors.

Interestingly, boosting algorithms are working excellently in these settings both in terms of absolute performances and learning rates. Considering more extreme ML techniques, namely random forests and deep neural networks, absolute performances are good especially in complex applications (presence of nonlinear effects) while the learning rate is somewhat slower than that of traditional techniques. It seems that random forests requires at least 150% the minimum sample size suggested for traditional regression models while neural networks requires about 200% the minimum sample size. This is true at least in situations where neural networks successfully outperforms the simpler modeling strategies (Cox and AFT regression without nonlinear effects). Concerning the performances of the methods in terms of time dependent Brier score, the simulations are quite concordant in suggesting the need for a double or triple sample size to obtain the same performance level of standard regression.

As pointed out by a reviewer, a clarification should be made about considering random forests more extreme ML methods than boosting. In fact, they look conceptually very similar, two ensemble methods in which the base learners are fitted on parallel (random forests) or sequential (boosting) transformations of the data. Boosting however is implemented in many flexible ways and, as outlined in Reference 58, can be seen both as a statistical model, when a statistical model is fitted, or as an algorithmic approach, when it is implemented using for example stumps. In the application presented here, the base learner used for boosting is in fact a traditional regression model, making the actual implementation more similar to traditional regression than random forests.

Another aspect that was investigated is that of the variability of the model performances. Variability in predictive performances is very important and contributes to the reliability of the model results. Looking at the results obtained it

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

seems that, when ML methods are effective, the variability of the performance scores is in line with that of traditional models. In difficult applications for ML, such as the one presented here on lung cancer, also the variability of ML is much larger than that of standard models.

At the end, looking at the results presented here, it seems that there would be no room for using extreme ML methods, such as random forests or neural networks, for developing a predictive model in a standard setting. In fact, these methods were never able to outperform boosting regression of simple Cox or AFT regression with splines to model nonlinear effects. Even performances of the simple regressions without nonlinearities were in some cases difficult to reach. Apart from considerations inherent to prediction performance and learning rates, the actual use of one particular strategy is however matter of many different considerations and "cultures." We share one sentence from the book of Kattan and Gerds[1] in which they say: " … we start by noting that for practical purposes it is often reasonable to expect that all sound strategies when applied to the same dataset should end-up with comparable results … " We therefore did not expect to see big differences among the methods and our main motivation was to determine whether an additional sample size was needed by RF and NN to achieve the same performances of traditional models. To this respect, it seems that, although the sample sizes needed are in fact larger, it is not an order of magnitude difference, and doubling or tripling the sample size is already effective. To this respect, RF, NN and even more boosting, could be a good way for benchmarking a standard, explainable, regression method.

In conclusion, the results most relevant both for practical and research purposes regard the very initial goal of the study, that is, clarification on the minimum sample size required for developing a reliable survival prediction model using ML models compared to traditional regression methods. While ML models require larger sample sizes, it is sufficient doubling or tripling the minimum sample size required by traditional regression models to obtain similar performances and not to increase by one or more order of magnitude. It is interesting that boosting algorithms, in the setting investigated here, need a lower or equal sample size than traditional regression models. Another key factor when building a prediction model is if and how possible nonlinear effects of covariates should be taken into account. ML techniques are of value for modeling nonlinearities and interactions, but must be accurately trained to find optimal hyper-parameters to balance the bias-variance trade-off. On the other hand, traditional regression models can easily account for nonlinear effects that must be explicitly modeled and more emphasis must be put on this topic in applied literature. Future applied and methodological research comparing ML and traditional regression cannot ignore inclusion of nonlinear effects in traditional regression models. These results come from a framework considering prediction models with small number of covariates but, rather than being of a limited use, they can be potentially helpful to biomedical researchers interested in exploiting clinical and biological variables for predicting patients' outcome.

AUTHOR CONTRIBUTIONS

Federico Ambrogi, Rosalba Miceli, and Gabriele Infante contributed to the conception, design, analysis, interpretation, and drafted the work.

ACKNOWLEDGEMENTS

The work was partially supported by Italian Ministry of Education, University and Research project PRIN 2017, prot. 20178S4EK9_004, Innovative Statistical methods in biomedical research on biomarkers: from their identification to their use in clinical practice.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

R code for simulation with GBCS and colon data are also available in github: https://github.com/federicoambrogi/Sample-size-and-ML-with-survival-data. The code uses the publicly available datasets colon,[20,46] available in the *survival R* package[47] (https://CRAN.R-project.org/package=survival) and the German Breast Cancer Study (GBCS) available in *R* package *condSURV*.[59] The SEER cancer data are accessible upon request (https://seer.cancer.gov/data/access.html).

ORCID

Federico Ambrogi https://orcid.org/0000-0001-9358-011X

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

REFERENCES

1. Gerds TA, Kattan MW. *Medical Risk Prediction Models: with Ties to Machine Learning*. 1st ed. Boca Raton, FL: Chapman and Hall/CRC; 2021.
2. Sperr E. PubMed by Year [Internet]; 2016. http://esperr.github.io/pubmed-by-year/
3. Brand RJ, Rosenman RH, Sholtz RI, Friedman M. Multivariate prediction of coronary heart disease in the Western collaborative group study compared to the findings of the Framingham study. *Circulation*. 1976;53(2):348-355.
4. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318.
5. Smith H, Sweeting M, Morris T, Crowther MJ. A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagn Progn Res*. 2022;6(1):1-15.
6. Ravdin PM, Clark GM. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Res Treat*. 2005;22:285-293.
7. Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat Med*. 1998;17(10):1169-1186.
8. Biganzoli E, Boracchi P, Marubini E. A general framework for neural network models on censored survival data. *Neural Netw*. 2002;15(2):209-218.
9. Fornili M, Boracchi P, Ambrogi F, Biganzoli E. Modeling the covariates effects on the hazard function by piecewise exponential artificial neural networks: an application to a controlled clinical trial on renal carcinoma. *BMC Bioinformatics*. 2018;19(Suppl 7):186.
10. Liestøl K, Andersen PK, Andersen U. Survival analysis and neural nets. *Stat Med*. 1994;13(12):1189-1200.
11. Ambrogi F, Lama N, Boracchi P, Biganzoli E. Selection of artificial neural network models for survival analysis with genetic algorithms. *Comput Stat Data Anal*. 2007;52(1):30-42. doi:10.1016/j.csda.2007.05.001
12. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995;48(12):1503-1510.
13. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441
14. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14(1):137. doi:10.1186/1471-2288-14-137
15. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860. doi:10.1214/08-AOAS169
16. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci*. 2007;22(4):477-505. doi:10.1214/07-STS242
17. Kvamme H, Borgan O. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal*. 2021;27(4):710-736.
18. Schmoor C, Sauerbrei W, Bastert G, Schumacher M. Role of isolated locoregional recurrence of breast cancer: results of four prospective studies. *J Clin Oncol*. 2000;18(8):1696-1708.
19. Schumacher M, Bastert G, Bojar H, et al. Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German breast cancer study group. *J Clin Oncol*. 1994;12(10):2086-2093.
20. Moertel CG, Fleming TR, Macdonald JS, et al. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *N Engl J Med*. 1990;322(6):352-358.
21. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagn Progn Res*. 2018;2:7.
22. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med*. 2013;32(23):4118-4134. doi:10.1002/sim.5823
23. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002;21(15):2175-2197.
24. Brilleman SL, Wolfe R, Moreno-Betancur M, Crowther MJ. Simulating survival data using the simsurv R package. *J Stat Softw*. 2020;97(3):1-27. doi:10.18637/jss.v097.i03
25. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II—binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296.
26. Ensor J, Martin EC, Riley RD. pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model. R package version 1.1.2; 2022.
27. Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med*. 2022;41(7):1280-1295.
28. Gerds TA, Schumacher M. Efron-type measures of prediction error for survival analysis. *Biometrics*. 2007;63(4):1283-1287.
29. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012;12:8.
30. Balki I, Amirabadi A, Levman J, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J*. 2019;70(4):344-353. doi:10.1016/j.carj.2019.06.002
31. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer Series in Statistics. New York: Springer; 2001.
32. Schmid M, Hothorn T. Flexible boosting of accelerated failure time models. *BMC Bioinformatics*. 2008;9:269.
33. Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B. Mboost: Model-Based Boosting. R package version 2.9-5; 2021.
34. Ishwaran H, Lu M, Kogalur UB. randomForestSRC: getting started with randomForestSRC vignette; 2021.
35. Ishwaran H, Lauer MS, Blackstone EH, Lu M, Kogalur UB. randomForestSRC: random survival forests vignette; 2021.

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

36. Ishwaran H, Kogalur U. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.12.1; 2021.
37. Mcculloch W, Pitts W. A logical calculus of ideas immanent in nervous activity. *Bull Math Biophys*. 1943;5:127-147.
38. Zhao L, Feng D. Deep neural networks for survival analysis using pseudo values. *IEEE J Biomed Health Inform*. 2020;24(11):3308-3314.
39. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems; 2015. https://www.tensorflow.org/
40. Chollet F. Keras; 2015.
41. Lee C, Yoon J, Schaar MV. Dynamic-DeepHit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans Biomed Eng*. 2020;67(1):122-133.
42. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. *PeerJ*. 2019;7:e6257.
43. Sonabend R. Neural networks for survival analysis in R: a demonstration of training, tuning, and comparing survival networks. https://towardsdatascience.com/neural-networks-for-survival-analysis-in-r-1e0421584ab. Accessed September 3, 2022.
44. Royston P, Sauerbrei W. *Multivariable Model-Building: A Pragmatic Approach to Regression Anaylsis Based on Fractional Polynomials for Modelling Continuous Variables*. Chichester, UK: John Wiley & Sons Ltd.; 2008.
45. Royston P, Altman DG. External validation of a cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13:33.
46. Moertel CG, Fleming TR, MacDonald JS, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report. *Ann Intern Med*. 1991;122:321-326.
47. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.
48. Eng KH, Seagle BL. Covariate-adjusted restricted mean survival times and curves. *J Clin Oncol*. 2017;35(4):465-466.
49. Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Database: Incidence - SEER Research Data, 18 Registries, Nov 2020 Sub (2000-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission.
50. Surveillance Research Program. National Cancer Institute SEER*Stat software version 8.3.9.2.
51. Lorenz E, Jenkner C, Sauerbrei W, Becher H. Modeling variables with a spike at zero: examples and practical recommendations. *Am J Epidemiol*. 2017;185(8):650-660.
52. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338:b375. doi:10.1136/bmj.b375
53. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16(3):199-231. doi:10.1214/ss/1009213726
54. Wallisch C, Agibetov A, Dunkler D, et al. The roles of predictors in cardiovascular risk models-a question of modeling culture? *BMC Med Res Methodol*. 2021;21(1):1-12.
55. Boulesteix AL, Lauer S, Eugster MJA. A plea for neutral comparison studies in computational sciences. *PLoS ONE*. 2013;8(4):1-11. doi:10.1371/journal.pone.0061562
56. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102.
57. Austin PC, Harrell FE, Steyerberg EW. Predictive performance of machine and statistical learning methods: impact of data-generating processes on external validity in the "large N, small p" setting. *Stat Methods Med Res*. 2021;30(6):1465-1483. doi:10.1177/09622802211002867
58. Rahnenführer J, De Bin R, Benner A, et al. Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. *BMC Med*. 2023;21(1):182. doi:10.1186/s12916-023-02858-y
59. Meira-Machado L, Sestelo M. condSURV: Estimation of the Conditional Survival Function for Ordered Multivariate Failure Time Data. R package version 2.0.1; 2016. https://CRAN.R-project.org/package=condSURV

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Infante G, Miceli R, Ambrogi F. Sample size and predictive performance of machine learning methods with survival data: A simulation study. *Statistics in Medicine*. 2023;42(30):5657-5675. doi: 10.1002/sim.9931

10970258, 2023, 30, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9931 by Universita Di Milano, Wiley Online Library on [11/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License