Received: 30 November 2021 | Revised: 8 June 2022 | Accepted: 4 July 2022

DOI: 10.1002/bimj.202100380

Biometrical Journal

RESEARCH ARTICLE

Individual risk prediction: Comparing random forests with Cox proportional-hazards model by a simulation study

Valia Baralou | **Natasa Kalpourtzi** | **Giota Touloumi**

Department of Hygiene, Epidemiology & Medical Statistics, Medical School, National & Kapodistrian University of Athens, Athens, Greece

Correspondence
Giota Touloumi, Department of Hygiene, Epidemiology, and Medical Statistics, Medical School, National & Kapodistrian University of Athens, Athens 115 27, Greece.
Email: gtouloum@med.uoa.gr

Valia Baralou and Natasa Kalpourtzi contributed equally to this work.

Funding information
Hellenic Foundation for Research and Innovation, Grant/Award Number: 679

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity and data confidentiality issues.

Abstract

With big data becoming widely available in healthcare, machine learning algorithms such as random forest (RF) that ignores time-to-event information and random survival forest (RSF) that handles right-censored data are used for individual risk prediction alternatively to the Cox proportional hazards (Cox-PH) model. We aimed to systematically compare RF and RSF with Cox-PH. RSF with three split criteria [log-rank (RSF-LR), log-rank score (RSF-LRS), maximally selected rank statistics (RSF-MSR)]; RF, Cox-PH, and Cox-PH with splines (Cox-S) were evaluated through a simulation study based on real data. One hundred eighty scenarios were investigated assuming different associations between the predictors and the outcome (linear/linear and interactions/nonlinear/nonlinear and interactions), training sample sizes (500/1000/5000), censoring rates (50%/75%/93%), hazard functions (increasing/decreasing/constant), and number of predictors (seven, 15 including noise variables). Methods' performance was evaluated with time-dependent area under curve and integrated Brier score. In all scenarios, RF had the worst performance. In scenarios with a low number of events (≤ 70), Cox-PH was at least noninferior to RSF, whereas under linearity assumption it outperformed RSF. Under the presence of interactions, RSF performed better than Cox-PH as the number of events increased whereas Cox-S reached at least similar performance with RSF under nonlinear effects. RSF-LRS performed slightly worse than RSF-LR and RSF-MSR when including noise variables and interaction effects. When applied to real data, models incorporating survival time performed better. Although RSF algorithms are a promising alternative to conventional Cox-PH as data complexity increases, they require a higher number of events for training. In time-to-event analysis, algorithms that consider survival time should be used.

KEYWORDS
Cox model, machine learning, random survival forest, survival analysis

1 | INTRODUCTION

The increasing availability of big data in healthcare [e.g., electronic health records (Goldstein et al., 2017b) and omics data (Kruppa et al., 2012)] has boosted the use of machine learning (ML) algorithms as an alternative approach to classic statistical methods in numerous biomedical research fields, such as risk prediction. A large body of studies have compared the predictive performance of ML and regression methods with conclusions being contradictory about which method behaves best (Choi et al., 2018; Christodoulou et al., 2019; Goldstein et al., 2017a; Gong et al., 2018; Spooner et al., 2020; Steele et al., 2018; van der Ploeg et al., 2014). It still remains an open question as to whether and under what circumstances ML outperforms regression models.

Individual risk prediction holds important benefits for clinical decision-making as it enables the identification of high-risk individuals (Holmberg & Parascandola, 2010), raises their awareness regarding their health status, and provides guidance to health professionals so that appropriate interventions, either for population surveillance or for individual treatment, are made (Moons et al., 2009). In traditional approaches, regression models have been applied to construct a points-based risk score. Such risk scores have been developed for a variety of diseases [i.e., cardiovascular disease (CVD): Conroy et al., 2003; D'Agostino et al., 2008; Hippisley-Cox et al., 2007; type 2 diabetes: Chen et al., 2010; Schulze et al., 2007; breast cancer: Gail et al., 1989] using logistic regression (Chen et al., 2010) as well as Weibull (Conroy et al., 2003) and Cox proportional hazards (Cox-PH) regression model (D'Agostino et al., 2008; Gail et al., 1989; Hippisley-Cox et al., 2007; Schulze et al., 2007) with the Cox-PH (Cox, 1972) being the most popular one.

In recent years, ML algorithms (e.g., artificial neural networks, random forest (RF), support vector machines) have also been employed to predict individual risk of developing a disease (Choi et al., 2018; Dimopoulos et al., 2018; Ward et al., 2020). However, in the majority of the applications, the algorithms used ignore the time until the event occurrence. Indicatively, a recent review of risk prediction modeling studies that used electronic health records found that only about 20% of the studies developed a time-to-event model, whereas almost half of them did not consider loss to follow-up at all (Goldstein et al., 2017b). However, in the presence of censored data, that is, when the event of interest has not occurred during the study period, the standard statistical approach is to develop a regression model that can handle censoring by taking into account the time until the event occurrence.

Random survival forest (RSF) seems to be quite popular in survival modeling of long-term outcomes. Although many studies have compared its performance with that of Cox-PH on real data (Qiu et al., 2020; Steele et al., 2018; Spooner et al., 2020), only a few have used simulations to compare RSF with either the classic Cox-PH (Gong et al., 2018; Omurlu et al., 2009; Wright et al., 2017) or its modern extensions such as penalized Cox-PH (Goldstein et al., 2017b; Wang et al., 2018). Moreover, no study has compared the performance of RF that ignores time-to-event information with RSF that can handle right-censored data.

In this study, we performed extensive simulation experiments to assess (i) whether and under which circumstances RSF outperforms Cox-PH, (ii) whether different split criteria used by RSF significantly impact its performance, and (iii) whether RF's performance is significantly inferior to the survival models in time-to-event analysis. Time-dependent area under curve (AUC) and integrated Brier score (IBS) were used to evaluate and compare the performance of the methods applied.

The remainder of this paper is organized as follows. In Section 2, the motivating example is introduced and the simulation study along with the methods investigated is described. In Section 3, the results from the simulation study and the real data application are presented, whereas conclusion remarks are discussed in Section 4.

2 | METHODS

2.1 | Motivating example

Data from the AMACS study ("Athens multicenter AIDS cohort study") were used as an illustrative example. AMACS is a cohort study initiated in 1996, containing a large database of people with HIV-1 (PWH). A variety of data are recorded at study entry and updated every year including demographic characteristics, clinical events and deaths, antiretroviral therapy, and laboratory test results.

AIDS-related mortality has substantially decreased since the introduction of antiretroviral therapy (Palella et al., 1998). However, the increased life expectancy of PWH has introduced new challenges in their healthcare management, as, in addition to the HIV infection, physicians have also to handle other noninfectious comorbidities related to aging such as diabetes, hypertension, dyslipidemia—known risk factors of CVD, which is the most frequent cause of death worldwide (World Health Organization, 2016). The estimation of the risk of developing a CVD event in PWH is of high importance for the assessment and management of CVD modifiable risk factors.

Several risk scores have been developed for the general population (Conroy et al., 2003; D'Agostino et al., 2008; Hippisley-Cox et al., 2007), and one specifically for PWH (Friis-Møller et al., 2003) that also contains HIV-related characteristics. All these risk prediction algorithms are based on traditional statistical methods, and, thus, their performance might be poor under more complex associations in the data that cannot be easily detected through Cox-PH or Weibull models. In addition, these risk prediction algorithms are based on specific datasets. The specific nature of each country's data may demand recalibration of the existing risk prediction algorithms or even the development of a new one. Thus, identifying through a simulation study the best performing method will help develop a robust recommendation for CVD risk prediction algorithms to be developed in the future.

2.2 | Simulation studies

The majority of simulation scenarios were generated under a realistic framework, roughly based on the AMACS data. In total, 180 simulation scenarios were set up. We assumed different associations between the predictors and the outcome [linear (L), linear and interactions (LI), nonlinear (NL), nonlinear and interactions (NLI)], training sample sizes (500, 1000, 5000), censoring rates (50%, 75%, 93%), and hazard functions (increasing, decreasing, constant). For all of these settings, seven predictors were considered, but for the NL and the L scenarios eight additional uninformative variables were also generated. Finally, to investigate models' performance under stronger nonlinear or interaction effects, we increased the value of some of the relevant coefficients in the NL and the LI scenario (NL-unrealistic, LI-unrealistic). The NL- and LI-unrealistic scenarios were only investigated assuming 50% censoring rate. In total, 500 replications were performed for all the scenarios.

Simulation of the covariates was based on parameters obtained from the AMACS data. Continuous variables were simulated from normal or truncated normal distribution, binary ones from Bernoulli distribution, and categorical variables with more than two categories from multinomial distribution. The means and the probabilities used in the simulation procedure were obtained from models applied to the AMACS data. Specifically, for the scenarios with seven covariates X_1 (age) was simulated from a truncated normal distribution, X_2 (gender) from a Bernoulli distribution with probabilities varying by X_1; X_3 (current smoking) also from a Bernoulli distribution, allowing the probabilities to depend on X_1, X_2; X_4 (total cholesterol) from a truncated normal distribution allowing mean to depend on X_1, X_2, X_3; X_5 (hdl) from a truncated normal with mean varying by $X_1, \ldots, X_4$; X_6 (triglycerides) from a truncated normal with mean varying by $X_1, \ldots, X_5$, X_7 (systolic blood pressure) with mean varying by $X_1, \ldots, X_5$ and X_8 (glucose) from a truncated normal distribution with mean varying by $X_1, \ldots, X_7$. Instead of X_5 and X_6 their ratio X_r was included in the models.

For the scenarios with 15 covariates, eight more variables were generated: X_9 (Greek origin; yes/no) from a Bernoulli distribution allowing the probability to be dependent on all previously mentioned variables. X_{10} [educational level: primary (1), secondary (2), higher (3)] was simulated from a multinomial logistic distribution with probabilities depending on $X_1, \ldots, X_9$. X_{11} (BMI), X_{12} (CD4), X_{13} (albumin), X_{14} (ggt), X_{15} (hemoglobin), and X_{16} (thrombocytes) were generated from a normal or truncated normal distribution, allowing the mean to differ by all previously simulated variables. The parameters used to simulate the covariates are presented in the Supplementary Appendix.

Survival times T were generated from a Weibull distribution with three different shape parameters (γ): 0.8 for decreasing, 1 for constant, and 2 for increasing hazard. The scale parameter is given by $\lambda_i = e^{-(\beta_0/\gamma)-\boldsymbol{\beta}^\top \boldsymbol{X}_i/\gamma}$ (Wan, 2017), where $\boldsymbol{X}_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{ri}, X_{7i}, X_{8i})$, $i = 1, 2, \ldots, N$, N the number of individuals, and $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_7)^\top$ the corresponding coefficients. Thus, the hazard of the event was allowed to depend on $X_1, X_2, X_3, X_4, X_r, X_7$, and X_8. The $\boldsymbol{\beta}$ values and the corresponding functional form of $\boldsymbol{X}_i$ in the scale parameter (through $\boldsymbol{\beta}^\top \boldsymbol{X}_i$) are presented in Supplementary Table S1.

Censoring times (C) were also simulated from a Weibull distribution with the same shape parameters with the distribution of survival times T. The scale parameters corresponding to survival (through β_0) and censoring time were manually tuned to obtain the desired censoring rate (50%, 75%, 93%), and the corresponding values are provided in Supplementary Tables S2 and S3. The final observed time was the minimum value of the survival and censoring time. If the minimum value was equal to the survival time, then the event indicator was set to 1 or 0 otherwise. For a more realistic context, we also considered administrative censoring at 10 years of follow-up.

2.3 | Models

2.3.1 | Random forest

RF (Breiman, 2001) can be used for both classification and regression tasks with focus here on RF as a classifier since we have a binary outcome. Specifically, RF is a collection of multiple decision trees, each generated from a bootstrap sample of the training data but only using a random subset of variables for node splitting. The covariate that best separates the data based on a split criterion, such as Gini impurity, is then selected to split the node into two daughter nodes, and this random node splitting process is recursively repeated until a stopping criterion is met (i.e., all terminal nodes contain no less than a given number of observations). The predictions of all trees are finally combined through averaging for probability estimation.

2.3.2 | Random survival forest

RSF (Ishwaran et al., 2008) is an extension of Breiman's RF (Breiman, 2001) that can handle right-censored survival data. It is an ensemble of a large number of survival trees which unlike RF uses a split rule so as to maximize the survival difference between daughter nodes. Its main output is an estimated cumulative hazard function computed by averaging the Nelson–Aalen cumulative hazard function of each tree.

In this study, we applied the following split criteria: the traditional log-rank test (RSF-LR) (Ishwaran et al., 2008) and the log-rank score (RSF-LRS), which is a modification of the log-rank split rule (Hothorn & Lausen, 2003). Although these approaches hold important advantages since they can handle high-dimensional data and model complex relationships, a disadvantage is that variables with many possible split points (i.e., continuous or categorical variables with multiple levels) have higher chances of being selected as optimal for splitting.

An alternative to overcome this issue is to conduct the optimal split variable selection using maximally selected rank statistics (Wright et al., 2017), which performs a binary split that adjusts for the multiple testing of multiple possible split points. Since we included a considerable number of continuous covariates, we also used the maximally selected rank statistics random survival forest (RSF-MSR) that avoids selection bias.

Based on preliminary results and in order to have reasonable run times, we set the number of trees equal to 250, the number of covariates tried at each split to the classic value $\lfloor \sqrt{p} \rfloor$, and the minimum number of observations in a terminal node to 15. For a fair comparison, we used the same parameter values for all RSF and RF models.

2.3.3 | Cox proportional hazards model

Two different Cox-PH (Cox, 1972) were used as the standard regression methods against which RSF and RF were compared: (i) one in which all continuous variables were included as linear terms (Cox-PH), (ii) one in which natural cubic splines with four knots were included for two of the continuous variables and more specifically for X_r and X_7 (Cox-S).

2.4 | Evaluation metrics

Models' performance was evaluated on test data of size N=10,000 that were generated independently from the training data. We compared models' performance with two popular criteria: time-dependent AUC (Uno et al., 2007) and IBS (Graf et al., 1999).

The time-dependent AUC at a given time point t denotes the probability that a randomly selected patient who experiences the event at time t has a higher predicted risk than a randomly selected patient without the event until that time point. Here, we report the inverse probability weighting (IPCW) estimate of the time-dependent AUC proposed by Uno et al. (2007). The Kaplan–Meier method was used for estimating the censoring weights.

The Brier score (Graf et al., 1999) is another popular measure of a model's accuracy. At a given time point t, the Brier score can be interpreted as the mean square error between the predicted survival function $\hat{S}(t|\boldsymbol{X})$, and the actual event status of the test data weighted by the inverse probability of censoring as follows:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left\{[0-\hat{S}(t|\boldsymbol{X}_i)]^2\frac{I(\tau_i \leqslant t, \delta_i = 1)}{\hat{G}(\tau_i)} + [1-\hat{S}(t|\boldsymbol{X}_i)]^2\frac{I(\tau_i > t)}{\hat{G}(t)}\right\}, \tag{1}$$

where N is the test sample size, $I(\cdot)$ the indicator function, τ_i the observed survival time, δ_i the censoring indicator for the individual i and $\hat{G}(t)$ the Kaplan–Meier estimate of the censoring distribution. The IBS over a certain time period $[0, t^*]$ is given by

$$IBS(t^*) = \frac{1}{t^*}\int_0^{t^*} BS(t)\,dt, \tag{2}$$

Survival models with a small IBS value are preferable.

2.5 | Implementation in R

The study was implemented in R version 4.1.0. The `survival` package was used for Cox-PH implementation, `randomForestSRC` for RSF-LR and RSF-LRS, `ranger` for RSF-MSR, and `randomForest` for RF. Survival models' performance was evaluated with the `riskRegression` and `pec` packages for time-dependent AUC and IBS calculation, respectively. The `ROCR` and `ModelMetrics` packages were used to assess RF's AUC and Brier score.

3 | RESULTS

3.1 | Simulation results

Since differences in performance were negligible among different hazard types, here we present results of scenarios assuming increasing hazard. Results corresponding to constant and decreasing hazard functions are presented in the Supplementary Appendix (Figures S1–S28).

In the vast majority of scenarios, RF that does not incorporate survival time had by far the worst performance among all methods considered in this study. In contrast with all the remaining models which gave similar IBS values, RF resulted in an approximately 56.4% relative increase in IBS and in an approximately 7.3% relative decrease in time-dependent AUC on average. It is also worth mentioning that in comparison with constant and decreasing hazard function, RF's performance slightly deteriorated when we assumed increasing hazard in contrast to the remaining survival models that performed similarly across different hazard types.

Under linearity assumption (scenario L), Cox-PH outperformed all three RSF algorithms in terms of both time-dependent AUC and IBS (average relative difference of 4.3% and 1.9%, respectively). As the number of events increased, the differences in performance between Cox-PH and RSF tended to decline (Supplementary Figures S29, S30). Although in settings with a relatively low number of events (e.g., ≤ 125 corresponding to scenarios with 93% censoring rate and training sample size of 500%, 1000%, and 75% censoring rate and training sample size of 500) Cox-PH outperformed Cox-S (average relative difference in AUC 2.8% and 2.2% in IBS), the performance of both methods became similar when the number of events further increased.

When assuming nonlinear effects (scenario NL), although Cox-PH outperformed RSF in the scenario with a low number of events (≤ 70 corresponding to scenarios with 93% censoring rate and training sample size of 500, 1000), the differences in their performance tended to attenuate as compared to that in the linear settings (average relative difference in AUC 1.9% and 0.5% in IBS). As the number of events increased, RSF algorithms had larger improvement in their predictive ability in comparison with that of Cox-PH. Particularly, in the scenario with 93% censoring rate and sample size 5000 as well as in those with 75% censoring rate and sample sizes of 500 or 1000, all methods had quite similar performance. However, under the scenario with 75% censoring rate and sample size of 5000 as well as all those with 50% censoring rate, RSF outperformed Cox-PH by 1.1–3.2% in terms of AUC. No significant differences in performance were observed between the three RSF variants (Figure 1 and Supplementary Figure S31). When including splines, the performance of Cox-PH

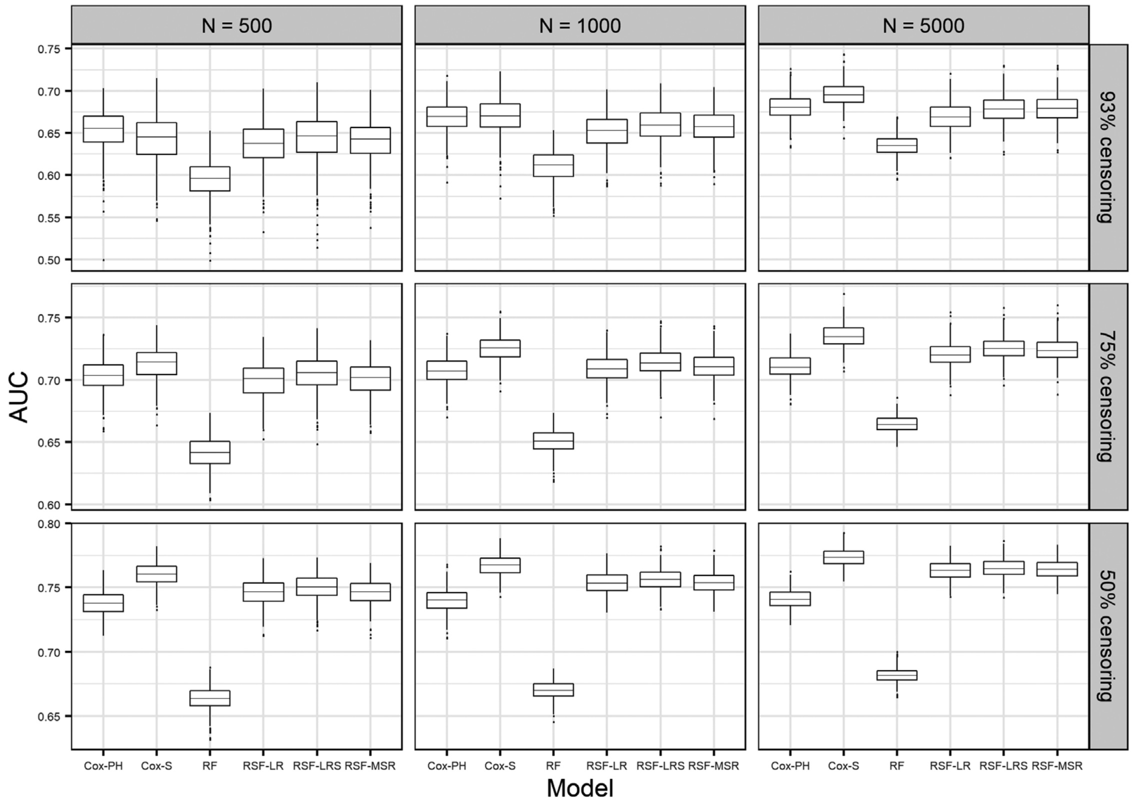

FIGURE 1 Predictive performance in terms of time-dependent AUC under the scenario with nonlinear effects and increasing hazard function when $N = 500, 1000, 5000$ and censoring rates 93%, 75%, and 50%

improved (1.5–4.3% relative increase in AUC) in all settings apart from the scenarios with a number of events ≤ 70. Cox-S performed similarly with RSF in the scenario with 93% censoring rate and training sample size of 500, whereas in the rest of the scenarios it outperformed RSF (1.6% average relative difference in AUC and 1.5% in IBS).

Regarding the LI scenario, Cox-PH performed better than RSF when the number of events was ≤ 70, whereas when the number of events was between 125 and 500 (corresponding to scenarios with 93% censoring rate and training sample size of 5000, 75% censoring rate and training sample size of 500 and 1000, and 50% censoring rate and training sample size of 500 and 1000) both AUC and IBS of RSF-LR, RSF-MSR, and Cox-PH were pretty similar ($<$ 1% relative difference). In the remaining scenarios (censoring rates 75% and 50% with 5000 sample size) RSF-MSR and RSF-LR had the best performance in terms of both AUC and IBS. In all the settings, RSF-LRS had slightly worse performance in comparison with RSF-MSR and RSF-LR (Figure 2 and Supplementary Figure S32). As for Cox-S, it performed worse than Cox-PH when the number of events was ≤ 125 (2.4% and 9% average relative difference in AUC and IBS, respectively) but reached similar performance with Cox-PH and, thus, lower than that of RSF in the remaining settings.

In the NLI scenario, RSF and Cox-PH performed similarly when the censoring rate was 93% (average relative difference in both AUC and IBS $<$ 1%). In the remaining cases (i.e., 75% and 50% censoring rate), RSF outperformed Cox-PH with average relative differences in AUC ranging from 1.2% to 5.1%. Differences in performance between the three RSF algorithms were negligible (Figure 3 and Supplementary Figure S33). Cox-S performed similarly with RSF in all settings (average relative difference in both AUC and IBS $<$ 1% apart from a few scenarios (93% censoring rate and training sample size of 5000, 75% censoring rate and training sample size of 500 or 1000) in which it had slightly better performance (1.6% average relative increase in AUC).

When assuming stronger interaction or nonlinear effects (Figure 4 and Supplementary Figure S34), greater differences between the methods' performance were observed. Particularly, in the NL-unrealistic scenario, in all the settings RSF algorithms outperformed Cox-PH with an average relative difference in AUC and IBS of 6.3% and 4.3%, respectively. The inclusion of splines in Cox-PH led to substantial improvement in its performance (7.8% and 8.4% average relative difference in AUC and IBS, respectively). In addition, when the training sample size was relatively small (i.e., 500 or 1000), Cox-S outperformed RSF by 1.4% and 3.4% on average in terms of AUC and IBS, respectively. In the LI-unrealistic scenario, RSF-

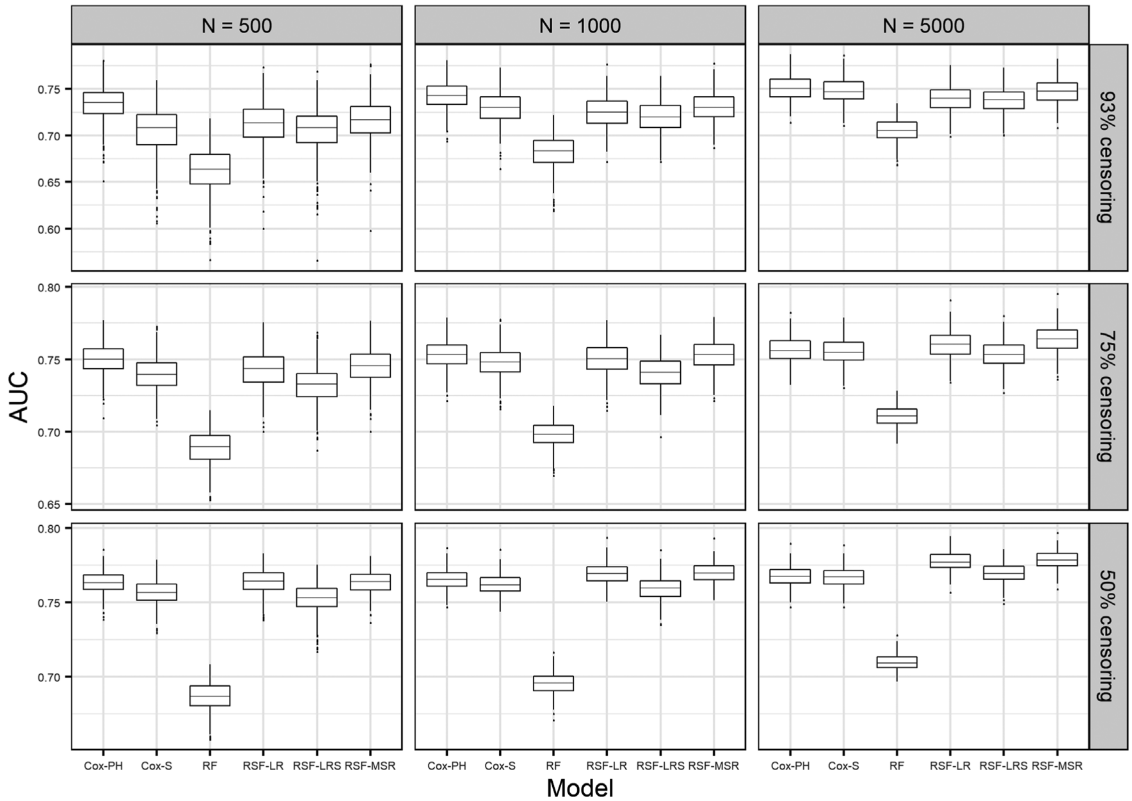

FIGURE 2 Predictive performance in terms of time-dependent AUC under the scenario with interaction effects and increasing hazard function when $N = 500$, 1000, 5000 and censoring rates 93%, 75%, and 50%

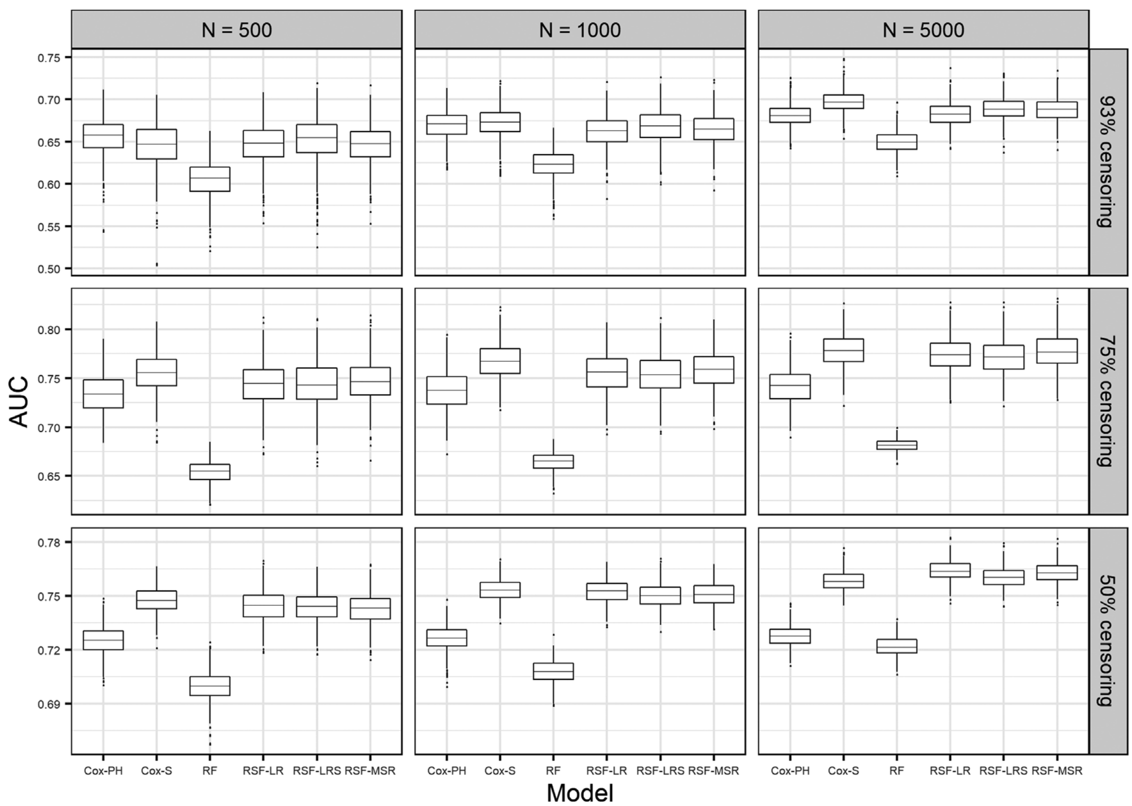

FIGURE 3 Predictive performance in terms of time-dependent AUC under the scenario with both nonlinear and interaction effects and increasing hazard function when $N = 500$, 1000, 5000 and censoring rates 93%, 75%, and 50%

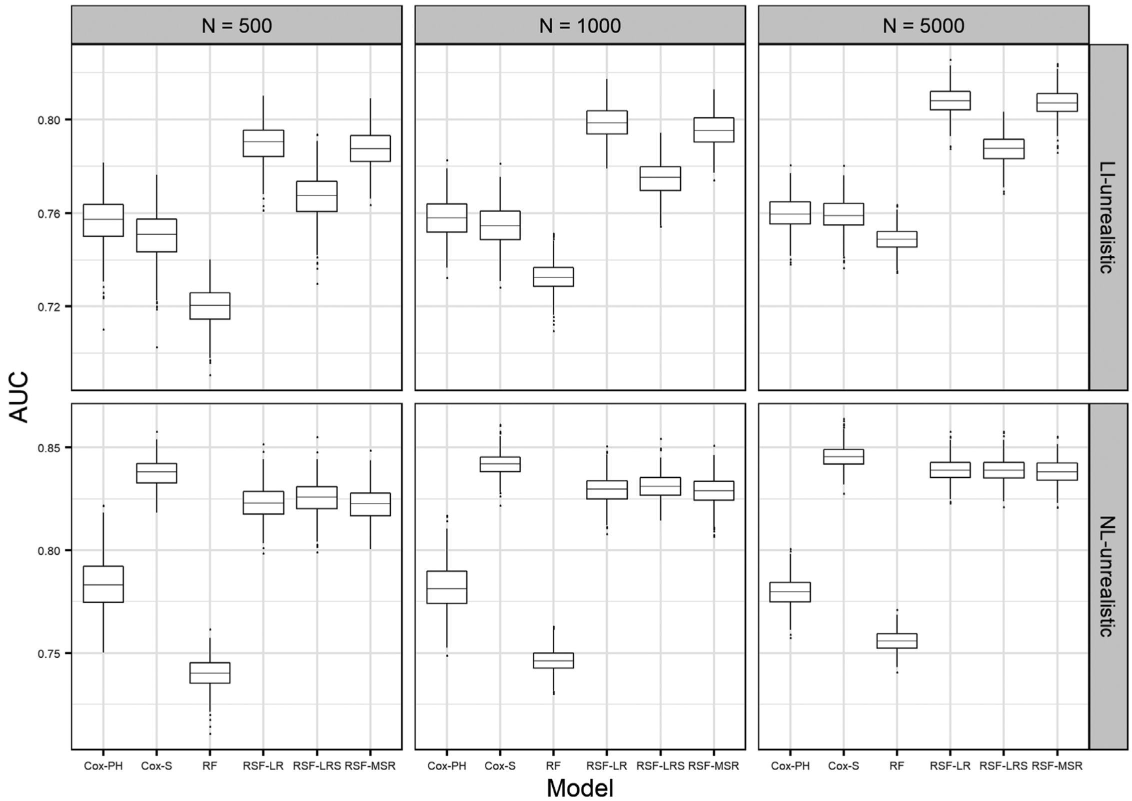

FIGURE 4 Predictive performance in terms of time-dependent AUC under the scenario with stronger nonlinear or interaction effects and increasing hazard function when $N = 500, 1000, 5000$ and 50% censoring rate

LR, and RSF-MSR had the best performance (5.2% average relative difference in AUC and 7.4% in IBS in comparison with Cox-PH) followed by RSF-LRS which also outperformed Cox-PH (2.4% and 7.3% average relative difference in AUC and IBS, respectively). Cox-S had similar performance with Cox-PH (average relative difference in AUC $< 1\%$).

When including uninformative covariates (Supplementary Figures S35–S38), our conclusions regarding the models' comparative performance were in accordance with those corresponding to the scenarios that did not contain noise variables. It is worth mentioning, though, that when the number of events was high (≥ 250), RSF-MSR and RSF-LR performed similarly with the respective scenarios that included only seven predictors whereas their performance improved notably in the remaining cases (i.e., number of events < 250). Furthermore, RSF-LRS performed worse than RSF-MSR and RSF-LR in all settings (2.5% and 2.1% average relative difference in AUC and IBS, respectively).

3.2 | Real data application

As an illustrative example, all models were applied and evaluated on the AMACS data to estimate the 7-year risk of developing a first CVD event in PWH. Our sample consisted of PWH aged between 30 and 60 years old at baseline and diagnosed with HIV after 2000. As baseline, we defined the study enrolment date and as the CVD event we considered myocardial infarction, coronary heart disease, heart failure, invasive coronary procedure, stroke, transient ischemic attack, peripheral artery disease (intermittent claudication), or death from a CVD event. Out of 5054 participants, 7.8% developed a first CVD event during 7 years of follow-up. We used multiple imputations by chained equations to handle the missing values in covariates; Five multiply imputed datasets were generated with a burn-in period of 500 iterations.

Ten covariates measured at baseline were finally included: age, gender, smoking status, diabetes status, systolic blood pressure, triglycerides, CD4 count, albumin, the total and high-density lipoprotein cholesterol. We fitted (i) a Cox-PH model with all continuous variables included as linear terms; (ii) a Cox-S model where total cholesterol and CD4 count were modeled through natural cubic splines with four knots, with these two covariates having the strongest evidence for deviations from linearity based on the Martingale residuals plots; (iii) RSF and RF models. To be consistent with our

TABLE 1 Predictive performance metrics (time-dependent AUC and IBS) of algorithms applied in the real data

	AUC	IBS
Cox-PH	0.758	0.048
Cox-S	0.757	0.048
RF	0.703	0.071
RSF-LR	0.723	0.048
RSF-LRS	0.743	0.048
RSF-MSR	0.735	0.048

Abbreviations: AUC, area under curve; COX-PH, Cox proportional hazards; Cox-S, COX-PH with splines; IBS, integrated Brier score; RF, random forest; RSF-LR, random survival forest log-rank; RSF-LRS, random survival forest log-rank score; RSF-MSR, random survival forest maximally selected rank statistics.

simulation study, we randomly split each imputed dataset into a training (70% of participants) and a test set (remaining 30%) for model assessment, keeping the same RSF and RF parameter values. Performance metrics were averaged across the imputed datasets and are presented in Table 1. In agreement with the simulation results, RF had the worst performance in terms of both AUC and IBS. Similarly with scenarios assuming a linear relationship and 93% censoring rate, Cox-PH and Cox-S had the highest AUC compared to the RSF algorithms among which RSF-LRS performed best. However, IBS values of all but RF models were very close.

4 | DISCUSSION

In this study, we investigated the performance of several random forest algorithms and classic statistical methods under a survival framework. The predictive ability of RSF with different split criteria, RF, Cox-PH, and Cox-S was examined through an extensive simulation study based on real data under scenarios with different hazard functions, training sample sizes, and associations between the covariates and the outcome (i.e., linear, nonlinear, interactions) and evaluated using time-dependent AUC and IBS. Models' performance was also assessed on real data.

In all scenarios, RF had the worst performance among all models considered. When assuming linear effects as well as when the number of events was low (≤ 70), Cox-PH was at least noninferior to RSF. When deviating from linearity (i.e., LI, NL, NLI scenarios), RSF outperformed Cox-PH as the number of events increased. However, including splines in the Cox-PH model improved its performance indicating that Cox-S is a robust alternative to RSF. Particularly, it performed similarly or better than RSF in scenarios with linear nonlinear effects. Although in the majority of the scenarios investigated, all three RSF variants had similar performance, RSF-LRS performed slightly worse than RSF-LR and RSF-MSR when including noise variables and interaction effects. Real data application led to similar results as those from the simulation study.

An enlightening finding of our study is the poor performance of RF in comparison with the remaining models. Although methods not incorporating time-to-event information are expected to perform poorly in survival analysis problems, several previous studies have applied a variety of ML algorithms that ignore survival time, including RF, for individual risk prediction and compared their performance with well-established risk scores developed using Cox-PH (Choi et al., 2018; Dimopoulos et al., 2018; Ward et al., 2020). Their findings show no substantial difference in the performance of ML and Cox-PH. Our results though strongly suggest using algorithms that account for survival time and avoiding those that ignore it (e.g., RF) in time-to-event analysis.

Recent literature on survival forests reveals that there is no algorithm that best fits all data (Nasejje et al., 2017; Wright et al., 2017). In their simulation study, Nasejje et al. (2017) compared the performance of RSF-LR, RSF-LRS, and Conditional Inference Forest (CIF) in time-to-event analysis and concluded that the appropriateness of each survival forest depends on the type of the covariates; suggesting the use of CIF in case of variables with many split points (i.e., nonbinary categorical and continuous variables). On the other hand, Wright et al. (2017) compared the predictive performance of different survival forests (including RSF-LR, CIF, and RSF-MSR) on both simulated and real data and recommended the use of RSF-MSR in the presence of nonlinear effects or covariates with many split points. Although mixed data were included in our study with the majority of the covariates being continuous, all three variants of the RSF algorithm performed similarly in all scenarios investigated apart from those that included uninformative variables or interaction effects where RSF-LRS had lower performance than RSF-MSR and RSF-LR.

Our study sheds light on the ongoing debate of whether and under what circumstances ML outperforms traditional statistical methods. Although a large body of studies have intended to answer this question in time-to-event analysis, they have mainly used real data and, hence, neither the underlying data mechanism was known nor could their conclusions be generalized (Qiu et al., 2020; Steele et al., 2018; Spooner et al., 2020). Only a few studies have used simulations to conduct a more solid comparison (Gong et al., 2018; Omurlu et al., 2009; Wright et al., 2017). Wright et al. (2017) found that survival forests were noninferior to Cox-PH in their simulation scenarios including nonlinear effects and uninformative variables, whereas Gong et al. (2018) undertook more extensive simulation experiments that contained nonlinear effects and interactions among covariates and showed that RSF-LR was by far superior to Cox-PH. On the other hand, Omurlu et al. (2009) showed that Cox-PH performed slightly better than RSF investigating only scenarios with $\leq$ 500 sample size while van der Ploeg et al. (2014) suggested that ML algorithms need more data for training compared to regression in a classification context though. Our findings are consistent with these studies, confirming that RSF is a promising alternative to Cox-PH as data complexity increases (i.e., nonlinear relationships, high order interactions) but requires a higher number of events for training. Particularly, in all scenarios investigated (apart from the unrealistic ones) when the number of events was $\leq$ 70, the performance of Cox-PH was at least noninferior to RSF despite the presence of relationships deviating from linearity or multiplicativity. None of the prementioned studies though included a Cox-PH model that allows for nonlinearity. Our study shows that the inclusion of splines in Cox-PH leads to at least similar performance with RSF in all settings and hence adds to the existing literature that the incorporation of nonlinear effects in Cox-PH can substantially boost its performance, making it a robust alternative to RSF.

In a preliminary analysis, Baralou et al. (2021) found that the variability of the performance metrics was larger when using equal-sized training and testing datasets rather than a testing dataset of 10,000 size. Based on this finding, here we evaluated models' performance in testing datasets of 10,000 size independently of the training dataset size. In addition to time-dependent AUC, in the preliminary analysis (Baralou et al., 2021) we calculated the concordance index, known as C-index, with both the IPCW (Uno et al., 2011) and Harrell's (Harrell Jr et al., 1996) estimates which were found to be quite close. Here, we opted to present only the time-dependent AUC as it is considered a more appropriate measure of the models' discriminative ability when predicted risk at a fixed time horizon is of interest (Blanche et al., 2019). It is worth noting though that conclusions regarding models' comparison did not change when using C-index as evaluation metric.

A limitation of this study is that no tuning of forests' parameters was performed despite the potential improvement of its performance in order to have reasonable run times. The number of trees was chosen on the basis of a trade-off between predictive performance and computational time, while the minimum node size and the number of split candidate variables were intentionally restricted to commonly used values. Moreover, no interaction terms were included in Cox-PH which burdens its performance in the presence of strong interaction effects oand, hence, can possibly explain its low performance, even close to RF, in extreme cases such as the LI-unrealistic scenario. Although ignoring interaction terms in Cox-PH is a model misspecification, their inclusion would unfairly boost Cox-PH's performance given that simulated data were generated assuming proportional hazards. Moreover, since it is impossible to detect all potential interactions included in the data in real situations, the need to define explicitly interaction terms in Cox-PH constitutes a considerable disadvantage against RSF which seems to be more flexible in cases of effects' heterogeneity.

There are still open research questions that need to be answered in future work. In addition to the type of the association between the predictors and the outcome, the required balance between the training sample size and the number of predictors remains an open issue. Since our study investigated different training sample sizes with a fixed number of covariates related to the outcome, future simulation studies are needed to conduct this comparison for different numbers of predictors. Moreover, our comparison experiments focus on forests. We plan to assess the performance of additional tree-based methods such as Gradient Boosting Machine (Ridgeway, 1999) as well as extensions of other ML algorithms in survival analysis such as survival support vector machines (Van Belle et al., 2011) and artificial neural networks (Biganzoli et al., 1998). Since simulated data were generated under the proportional hazards assumption, future studies are needed to investigate the relative performance of RSF and traditional statistical methods for survival data in scenarios that violate this assumption. We also relied on the assumption that the censoring mechanism does not depend on covariates. It would be very interesting to investigate how random survival forests behave under different censoring mechanisms; currently, the literature on this topic is very limited (Gerber et al., 2021).

In conclusion, our results showed that in survival studies, RF that does not incorporate survival time should be avoided, as in comparison with all the survival models applied, it had by far the worst performance. Under linearity assumption as well as in scenarios with a low number of events ($\leq$ 70), Cox-PH was at least noninferior to RSF. On the other hand, when deviating from linearity, RSF outperformed Cox-PH as the number of events increased and the nonlinear or interaction effects became stronger whereas the inclusion of splines in Cox-PH in the presence of nonlinear effects can improve its performance making it a fair alternative to RSF. However, given large sample sizes, as the complexity of the relationship among many covariates increases, RSF would be preferable for individual risk prediction.

ACKNOWLEDGMENTS

NK was supported for this research by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI Ph.D. Fellowship grant (Fellowship Number: 679). VB was co-funded by the European Union and Greek national funds through the Operational Program "Competitiveness, Entrepreneurship and Innovation" under the call "RESEARCH-CREATE-INNOVATE" (project code: T1EΔK-03812).

CONFLICT OF INTEREST

GT has received EU and national funds through the Operational Program "Competitiveness, Entrepreneurship and Innovation" (project code: T1EΔK-03812) to support this study and grants unrelated to this study from Gilead Sciences Europe, UCL, ECDC, University of Bristol, EU, and National funds, all paid to her institution. The rest of the authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity and data confidentiality issues.

ORCID

Valia Baralou https://orcid.org/0000-0001-7089-6787
Natasa Kalpourtzi https://orcid.org/0000-0002-2828-3140

REFERENCES

Baralou, V., Kalpourtzi, N., & Touloumi, G. (July 18–22, 2021). Individual risk prediction: Comparing random forests with cox proportional-hazards model by a simulation study [conference presentation abstract]. In 42nd Annual Conference of the International Society for Clinical Biostatistics, Lyon, France.

Biganzoli, E., Boracchi, P., Mariani, L., & Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine*, *17*(10), 1169–1186.

Blanche, P., Kattan, M. W., & Gerds, T. A. (2019). The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics*, *20*(2), 347–357.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, L., Magliano, D. J., Balkau, B., Colagiuri, S., Zimmet, P. Z., Tonkin, A. M., Mitchell, P., Phillips, P. J., & Shaw, J. E. (2010). Ausdrisk: An Australian type 2 diabetes risk assessment tool based on demographic, lifestyle and simple anthropometric measures. *Medical Journal of Australia*, *192*(4), 197–202.

Choi, S. B., Lee, W., Yoon, J.-H., Won, J.-U., & Kim, D. W. (2018). Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. *Journal of Affective Disorders*, *231*, 8–14.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22.

Conroy, R. M., Pyörälä, K., Fitzgerald, A. E., Sans, S., Menotti, A., De Backer, G., De Bacquer, D., Ducimetiere, P., Jousilahti, P., Keil, U., Njølstad, I., Oganov, R. G., Thomsen, T., Tunstall-Pedoe, H., Tverdal, A., Wedel, H., Whincup, P., Wilhelmsen, L., Graham, I. M., & SCORE project group. (2003). Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *European Heart Journal*, *24*(11), 987–1003.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care. *Circulation*, *117*(6), 743–753.

Dimopoulos, A. C., Nikolaidou, M., Caballero, F. F., Engchuan, W., Sanchez-Niubo, A., Arndt, H., Ayuso-Mateos, J. L., Haro, J. M., Chatterji, S., Georgousopoulou, E. N., Christos, P., & Demosthenes, B. P. (2018). Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC Medical Research Methodology*, *18*(1), 1–11.

Friis-Møller, N., Weber, R., Reiss, P., Thiébaut, R., Kirk, O., d'Arminio Monforte, A., Pradier, C., Morfeldt, L., Mateu, S., Law, M., El-Sadr, W., De Wit, S., Sabin, C. A., Phillips, A. N., Lundgren, J. D., & DAD study group. (2003). Cardiovascular disease risk factors in HIV patients–association with antiretroviral therapy. results from the DAD study. *AIDS*, *17*(8), 1179–1193.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, *81*(24), 1879–1886.

Gerber, G., Faou, Y. L., Lopez, O., & Trupin, M. (2021). The impact of churn on client value in health insurance, evaluation using a random forest under various censoring mechanisms. *Journal of the American Statistical Association*, *116*(536), 2053–2064.

Goldstein, B. A., Navar, A. M., & Carter, R. E. (2017a). Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *European Heart Journal*, *38*(23), 1805–1814.

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. (2017b). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, *24*(1), 198–208.

Gong, X., Hu, M., & Zhao, L. (2018). Big data toolsets to pharmacometrics: Application of machine learning for time-to-event analysis. *Clinical and Translational Science*, *11*(3), 305–311.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*(17-18), 2529–2545.

Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361–387.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., & Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *BMJ*, *335*(7611), 136.

Holmberg, C., & Parascandola, M. (2010). Individualised risk estimation and the nature of prevention. *Health, Risk & Society*, *12*(5), 441–452.

Hothorn, T., & Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, *43*(2), 121–137.

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

Kruppa, J., Ziegler, A., & König, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. *Human Genetics*, *131*(10), 1639–1654.

Moons, K. G., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: What, why, and how? *BMJ*, *338*, b375.

Nasejje, J. B., Mwambi, H., Dheda, K., & Lesosky, M. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Medical Research Methodology*, *17*(1), 1–17.

Omurlu, I. K., Ture, M., & Tokatli, F. (2009). The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer. *Expert Systems with Applications*, *36*(4), 8582–8588.

Palella, F. J., Jr, Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Satten, G. A., Aschman, D. J., Holmberg, S. D., & HIV Outpatient Study Investigators. (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine*, *338*(13), 853–860.

Qiu, X., Gao, J., Yang, J., Hu, J., Hu, W., Kong, L., & Lu, J. J. (2020). A comparison study of machine learning (random survival forest) and classic statistic (Cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. *Frontiers in Oncology*, *10*, 551420.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, *31*, 172–181.

Schulze, M. B., Hoffmann, K., Boeing, H., Linseisen, J., Rohrmann, S., Möhlig, M., Pfeiffer, A. F., Spranger, J., Thamer, C., Häring, H.-U., Fritsche, A., & Joost, H.-G. (2007). An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care*, *30*(3), 510–515.

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, *10*(1), 1–10.

Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS ONE*, *13*(8), e0202344.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L.-J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, *30*(10), 1105–1117.

Uno, H., Cai, T., Tian, L., & Wei, L.-J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, *102*(478), 527–537.

Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. (2011). Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, *53*(2), 107–118.

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*(1), 1–13.

Wan, F. (2017). Simulating survival data with predefined censoring rates for proportional hazards models. *Statistics in Medicine*, *36*(5), 838–854.

Wang, H., Shen, L., Geng, J., Wu, Y., Xiao, H., Zhang, F., & Si, H. (2018). Prognostic value of cancer antigen-125 for lung adenocarcinoma patients with brain metastasis: A random survival forest prognostic model. *Scientific Reports*, *8*(1), 1–9.

Ward, A., Sarraju, A., Chung, S., Li, J., Harrington, R., Heidenreich, P., Palaniappan, L., Scheinker, D., & Rodriguez, F. (2020). Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ Digital Medicine*, *3*(1), 1–7.
World Health Organization. (2016). *World health statistics 2016: Monitoring health for the SDGs sustainable development goals*. World Health Organization.
Wright, M. N., Dankowski, T., & Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine*, *36*(8), 1272–1284.

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: Baralou, V., Kalpourtzi, N., & Touloumi, G. (2023). Individual risk prediction: Comparing random forests with Cox proportional-hazards model by a simulation study. *Biometrical Journal*, *65,* 2100380. https://doi.org/10.1002/bimj.202100380