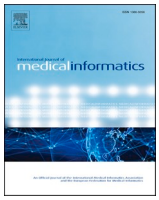
International Journal of Medical Informatics 194 (2025) 105700

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Balancing accuracy and Interpretability: An R package assessing complex relationships beyond the Cox model and applications to clinical prediction

Diana Shamsutdinova [a,*], Daniel Stamate [b,c], Daniel Stahl [a]

[a] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
[b] Data Science and Soft Computing Lab, Computing Department, Goldsmiths University of London, United Kingdom
[c] School of Health Sciences, University of Manchester, Manchester, United Kingdom

ARTICLE INFO

Keywords:
Clinical prediction model
Interpretability
Survival analysis
Ensemble methods
Internal validation
R

ABSTRACT

Background: Accurate and interpretable models are essential for clinical decision-making, where predictions can directly impact patient care. Machine learning (ML) survival methods can handle complex multidimensional data and achieve high accuracy but require post-hoc explanations. Traditional models such as the Cox Proportional Hazards Model (Cox-PH) are less flexible, but fast, stable, and intrinsically transparent. Moreover, ML does not always outperform Cox-PH in clinical settings, warranting a diligent model validation. We aimed to develop a set of R functions to help explore the limits of Cox-PH compared to the tree-based and deep learning survival models for clinical prediction modelling, employing ensemble learning and nested cross-validation.

Methods: We developed a set of R functions, publicly available as the package "survcompare". It supports Cox-PH and Cox-Lasso, and Survival Random Forest (SRF) and DeepHit are the ML alternatives, along with the ensemble methods integrating Cox-PH with SRF or DeepHit designed to isolate the marginal value of ML. The package performs a repeated nested cross-validation and tests for statistical significance of the ML's superiority using the survival-specific performance metrics, the concordance index, time-dependent AUC-ROC and calibration slope. To get practical insights, we applied this methodology to clinical and simulated datasets with varying complexities and sizes.

Results: In simulated data with non-linearities or interactions, ML models outperformed Cox-PH at sample sizes $\geq$ 500. ML superiority was also observed in imaging and high-dimensional clinical data. However, for tabular clinical data, the performance gains of ML were minimal; in some cases, regularised Cox-Lasso recovered much of the ML's performance advantage with significantly faster computations. Ensemble methods combining Cox-PH and ML predictions were instrumental in quantifying Cox-PH's limits and improving ML calibration. Traditional models like Cox-PH or Cox-Lasso should not be overlooked while developing clinical predictive models from tabular data or data of limited size.

Conclusion: Our package offers researchers a framework and practical tool for evaluating the accuracy-interpretability trade-off, helping make informed decisions about model selection.

1. Introduction

Digitization of healthcare records has given rise to clinical prediction modelling, while the advancements in machine learning (ML) have introduced more complex models to clinicians and researchers. Clinical prediction models deal with a wide range of outcomes, including disease risk, treatment efficacy, or mortality. Clinical models often estimate risks of future events, requiring longitudinal data where predictors are collected prior to the outcome. In such cases, survival analysis methods are typically preferred to make the full use of longitudinal data and accommodate censored observations [1–5]. Among these, the Cox Proportionate Hazard model [6] is one of the most widely used. ML models, including XGBoost, Survival Random Forests, FastCPH, DeepHit, and DeepSurv, have been adapted for time-to-event data and offer more

Abbreviations: ML, machine learning; AI, artificial intelligence; SRF, survival random forest; Cox-PH, Cox Proportional Hazards model; LASSO, least absolute shrinkage and selection operator; Cox-Lasso, Cox Proportional Hazards model with LASSO regularisation; ROC-AUC, area under the receiver-operating characteristic curve; BMI, body mass index; T2DM, type 2 diabetes mellitus.

* Corresponding author at: IoPPN, King's College London, De Crespigny Park, Box PO20, London SE5 8AF, United Kingdom.
E-mail address: diana.2.shamsutdinova@kcl.ac.uk (D. Shamsutdinova).

https://doi.org/10.1016/j.ijmedinf.2024.105700
Received 18 October 2024; Accepted 8 November 2024
Available online 10 November 2024

flexible approaches [7–12].

While these models have been tested for their ability to predict clinical outcomes, there is often a focus on how newer, more complex models outperform Cox-PH [3,11,13]. However, interpretability has become a critical factor in model selection, for the lack of which ML models are often criticised. Interpretable models enable practitioners to assess whether the predictions align with clinical knowledge, mitigate concerns about algorithmic bias, and aid integration into clinical practice. Although post-hoc explanation methods such as LIME and SHAP [14,15] have been developed, many argue that complex methods should only be employed when classical models meet their limits [16,17]. Others directly call for interpretable models over black-box algorithms [18,19]. However, interpreting survival models poses additional challenges due to the time dimension involved, compared to non-longitudinal data. New methods like SurvLIME, SurvShap, and SurvNAM [20–22] have been adapted for survival data but are still at early stages of validation.

In light of these challenges, we aimed to develop a set of user-friendly functions that would help researchers determine if added ML's complexity translates into a better prediction accuracy for a given dataset. The package evaluates model performances using repeated nested cross-validation, and the potential ML's superiority is tested for statistical significance. Focusing on ML abilities to capture data complexities beyond the Cox model, we selected a tree-based ensemble method, Survival Random Forest (SRF), and a deep-learning model, DeepHit, as ML alternatives to Cox-PH. SRF is particularly suited for handling interaction terms and dealing with outliers due to the tree structure. It also addresses non-proportional hazards by making predictions using Kaplan-Meier non-parametric survival curves in the final leaves. DeepHit, on the other hand, is a well-established deep learning model, which is not based on the Cox loss and does not rely on the proportional hazard assumption, unlike many other deep learning survival models [12].

To enhance the performance of the underlying models and provide an additional insight into the marginal contribution of ML over Cox-PH, we introduced two ensemble approaches. The first ensemble incorporates Cox-PH predictions as an additional feature in SRF or DeepHit, ensuring that ML models learn at least as much as Cox-PH does. Any improvement in performance in this ensemble can then be attributed to the ML's ability to capture relationships that Cox-PH could not. The second ensemble stacks the Cox-PH and ML models, optimizing their linear combination, expressed as $(1-\lambda)$ Cox + λ ML, where $\lambda \in [0, 1]$. This setup allows us to quantify the 'black box' contribution needed to enhance the Cox-PH's predictions.

To provide practical guidance on when Cox-PH reaches its limits, we applied this methodology to real-life and simulated medical datasets. These datasets represented a variety of health outcomes (mortality, new disease cases), settings (clinical trials, observational data), and data domains (socioeconomic and behavioural information, genetic biomarkers, physical and mental health indicators, clinical imaging). Simulated data allowed us to explore ML's superiority across a range of training set sizes (200–5000 cases), and complexities, including data with linear, non-linear, and interaction terms.

2. Methods

We first describe the underlying models and define the performance metrics. Then, tuning and validation methods are explained followed by a brief dataset description, and a code example. More details are given in the Supplementary Materials.

2.1. Models

Survival data is typically presented by a p-dimensional vector of predictors, x, and a pair of the outcome variables, (t, σ), where t is the observation time and σ is the binary outcome. If $\sigma = 1$, the event had occurred at t; if $\sigma = 0$, then no event was observed from 0 to t. Further, let T be a random variable representing time-to-event, then the survival function is the probability of being event-free past t, $S(t|x) = P(T > t|x)$, and the hazard rate is an instantaneous rate of event:

$$h(t|x) = \lim_{dt \to 0} \frac{1}{dt} P(t \leqslant T < t + dt | T \geqslant t, x) = -\frac{S'(t|x)}{S(t|x)}. \quad (1)$$

2.1.1. Cox Proportionate hazards model (Cox-PH)

Cox-PH [6] is one of the mostly used survival models, celebrated its 50-th anniversary in 2022. This semiparametric model does not make assumptions about the baseline hazard function, $h_0(t)$, but assumes that the risk factors act on it multiplicatively:

$$h(t|x, \beta) = h_0(t) \cdot \exp(\beta^T x). \quad (2)$$

From that it follows that Cox-PH is a linear regression in the log-hazard space. Another distinct feature of Cox-PH is the proportionality of hazards (PH) assumption. Indeed, the ratio of two individual hazards is time-invariant,

$$\frac{h(t|x_i, \beta)}{h(t|x_j, \beta)} = \exp(\beta^T (x_i - x_j)). \quad (3)$$

The regression parameters are estimated by maximisation of the partial likelihood [6]:

$$L_{partial}(\beta) = \prod_{i:\sigma_i=1} \frac{h(T_i \mid x_i, \beta)}{\sum_{j:t_j \geqslant t_i} (T_j \mid x_j, \beta)}. \quad (4)$$

2.1.2. Cox-Lasso

The regularized version of the Cox model, known as Cox-Lasso [23], incorporates an *L1*-norm penalty on the model's parameters into the maximization process:

$$LL_{Lasso}(\beta) = \log(L_{partial}(\beta)) - \lambda \| \beta \|_{L1}, \quad \| \beta \|_{L1} = \sum_{i=1..p} |\beta_i|. \quad (5)$$

Regularization is often used to reduce overfitting to training data by shrinking or eliminating the influence of weak predictors. It is especially useful for addressing ill-posed problems, where classical regressions fail, such as when the number of predictors exceeds the sample size, or there is multicollinearity among the predictors.

2.1.3. Survival random Forest (SRF)

SRF is an adaptation of Random Forests to survival outcomes. SRF aggregates predictions of many decision trees, grown in parallel on a bootstrapped version of the data. Each tree recursively partitions the data in the predictor space, splitting the data into homogeneous subsamples. Among the splitting rules proposed [24,25], the log-rank-based proved particularly effective [26]. Individual prognoses are made from the observations in the terminal leaves using a nonparametric survival estimate, such as Nelson-Aalen estimator, and averaging predictions across the trees. The advantages of SRF over Cox-PH is its ability to account for the predictor's interaction and operate in nonproportionate hazards [27].

2.1.4. DeepHit

DeepHit [10] is a deep learning discrete-time survival model, based on a fully connected neural network, or several networks when handling competing events. The final layer contains several *softmax* outputs, modelling survival probabilities for different discrete time points, therefore estimating time-to-event's probability distribution functions, $\widehat{P}(T = t|x)$, and $\widehat{F}(t|x) = P(T \leq t|x)$. A distinctive feature of DeepHit is the loss function which combines the log-likelihood of the censored data ($\mathscr{L}_1$) with the ranking loss ($\mathscr{L}_2$). Simplifying [10] for a single outcome without competing risks, the loss is as follows:

$$\mathscr{L} = \mathscr{L}_1 + \alpha \cdot \mathscr{L}_2,$$
$$\mathscr{L}_1 = -\sum_{i:\sigma_i=1} \log(\widehat{P}(T = T_i|x_i)) - \sum_{i:\sigma_i=0} \log(1 - \widehat{F}(T_i|x_i)), \quad (6)$$
$$\mathscr{L}_2 = \sum_{i \neq j:\sigma_i=1, T_i<T_j} \exp\left(-\frac{1}{\sigma_{DH}}\left(\widehat{F}(T_i|x_i) - \widehat{F}(T_i|x_j)\right)\right).$$

Here, $\mathscr{L}_1$ ties the estimated probability mass function to the observed event rates, while $\mathscr{L}_2$ encourages higher risk estimates for observations experiencing the event earlier. This optimizes both discrimination and calibration during model fitting [28,29].

2.1.5. Ensemble 1: Sequential approach

In this ensemble, the out-of-sample Cox predictions are added to the ML model as an additional predictor (Fig. 1). This approach has two key benefits. First, this may help improve the overall performance, especially in smaller datasets where ML models are particularly prone to overfitting and may not capture linearity as effectively as linear models do. Second, by ensuring that the ML model learns at least as much as Cox, the performance difference can be viewed as the predictive value of the relationships that the Cox model could not capture. This ensemble was validated in our previous work [30] and for cross-sectional data in [31].

2.1.6. Ensemble 2: Stacking approach

The second ensemble employs a stacking technique [32] common in ML. The underlying models (Cox-PH and either SRF or DeepHit) are trained independently, after which a *meta*-learner is tuned that combines their predictions. The *meta*-learner is a linear combination of the models' predictions, regulated by a tuning parameter λ (Fig. 1),

$$\widehat{S}_{stack}(t) = (1-\lambda)\cdot\widehat{S}_{CoxPH}(t) + \lambda\cdot\widehat{S}_{ML}(t), \quad \lambda \in [0,1]. \quad (7)$$

Equation (7) can be rewritten as $\widehat{S}_{CoxPH}(t) + \lambda\cdot(\widehat{S}_{ML}(t) - \widehat{S}_{CoxPH}(t))$, therefore λ represents the share of ML's outperformance over CoxPH's contribution to the final predictions.

The condition $\lambda \in [0,1]$ ensures that predictions remain in the [0,1] range. The lambda is tuned using out-of-sample predictions and is similar to the meta-learner's optimisation in Scikit-learn [33]. Fitting the ensemble involves: (1) using cross-validation to generate out-of-sample predictions from the underlying models, (2) selecting the lambda maximizing C-index from a set of 100 evenly spaced λ in [0,1]; 3) re-fitting the underlying models to the data supplied and returning those models with the optimised lambda as the components of the final model.

2.2. Performance metrics

The models were assessed in discrimination and calibration [34]. *Discrimination* is how well a model separates high and low risk observations and is measured by Harrell's concordance index (C-index) [35], and the time-dependent area under the receiver operating curve (AUC-ROC) [36]. The concordance index for predictions of a binary (1/0) outcome measures the probability of assigning a higher risk to an outcome of 1 compared to those with a 0 outcome, for a random pair of observations. It has been shown that the c-statistic is equivalent to the AUC-ROC for binary outcomes [37]. For survival data, Harrell's C-index measures the model's probability of assigning a greater risk of failure for an observation with the shorter survival time, while time-dependent AUC-ROC reflects the concordance of the predicted survival probabilities by a certain time with the observed survival. The equivalence does not hold for these measures, and we compute both.

$$c\text{-}index = \frac{\#\ concordant\ pairs}{\#\ permissible\ pairs} = \frac{\sum_{i,j:T_i>T_j} I\left(\widehat{y}_i < \widehat{y}_j\ \&\ \sigma_j = 1\right)}{\sum_{i \neq j} I(T_i > T_j\ \&\ \sigma_j = 1)} \quad (8)$$

Here, $\widehat{y}_i$ is an estimated risk score, such that a higher score indicates a shorter time-to-event.

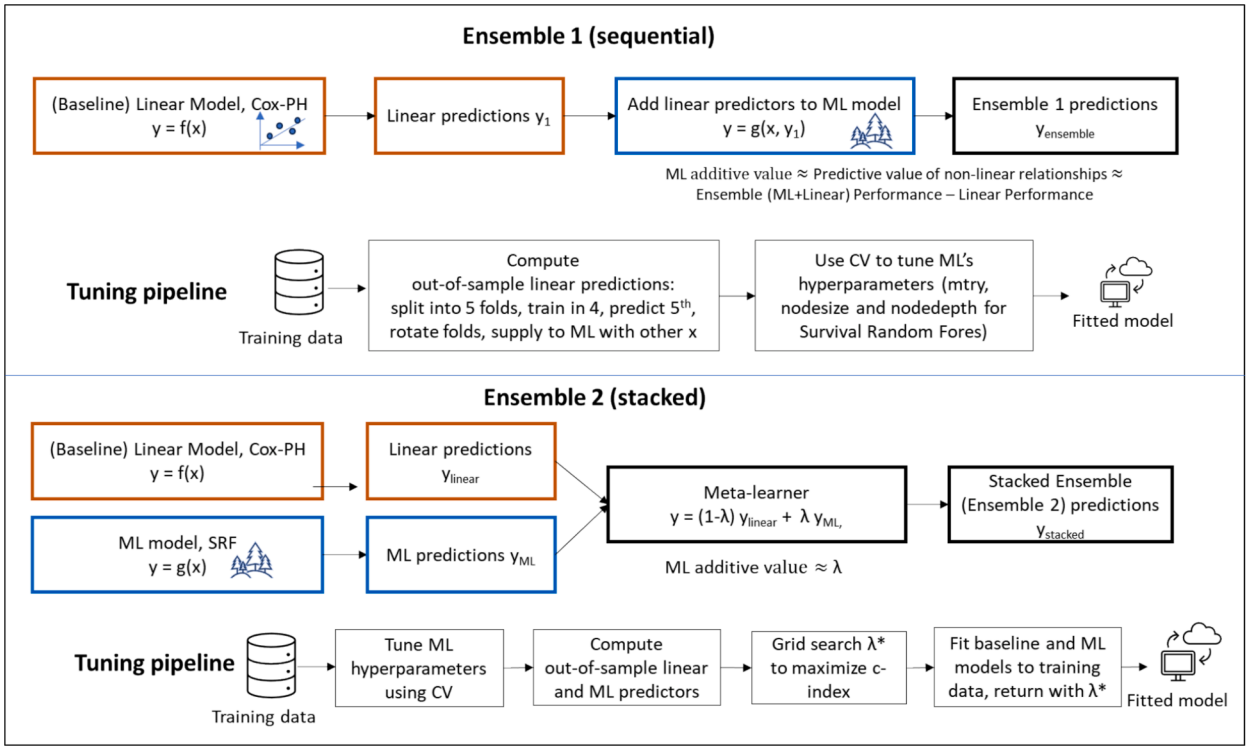


Fig. 1. Ensemble models description. The figure explains the underlying idea of the sequential ensemble and uses generic names such as "Linear" and "ML" models. In this paper and in the 'survcompare' package, we use Cox-PH or Cox-Lasso as the linear models, and Survival Random Forests, and DeepHit as nonlinear ML alternatives.

$$AUCROC(t) = \int_{c=-\infty}^{\infty} Sensitivity(c,t)\, d(1 - Specificity(c,t)),$$
$$Sensitivity(c,t) = P(\widehat{y}_i > c | D_i(t) = 1),$$
$$Specificity(c,t) = P(\widehat{y}_i \leq c | D_i(t) = 0). \quad (9)$$

Here, $D_i(t)$ is the event status by time t. Cases are the observations with observed events by t, that is, $D_i(t) = 1$ if $T_i \leq t$ and $\sigma_i = 1$. Controls are those observed past t with no events, $D_i(t) = 0$ if $T_i > t$. It was shown that an inverse probability of censoring weights (IPCW) adjustment can increase stability of the AUC-ROC estimates across various censoring scenarios [38], which is implemented by Blanche et al, and utilised in our code [39].

Calibration is how well the estimated chances correspond to the observed event rates. For example, if estimates survival chances are 0.85, it is expected that among the people with similar risk profiles the observed survival would also be 0.85. Calibration can be measured by the calibration slope [34]. For binary outcomes, it is computed by regressing observed outcomes on the model's risk score and taking the slope coefficient. An ideal slope is 1; a lower value indicates overfitting, and vice versa. For survival data, we fix the time of interest and evaluate calibration slope with respect to the observed binary survival by this time. If $\widehat{y}_i(t)$ are predicted event probabilities by t, and we defined time-dependent cases and controls as $D_i(t) = 1$ or 0 as in Equation (9), calibration slope is the slope coefficient in the logistic regression of $D_i(t)$ on $\widehat{y}_i(t)$ [40],

$$\log \frac{P(D(t) = 1)}{1 - P(D(t) = 1)} = b_0 + slope(t) \cdot \widehat{y}_t. \quad (10)$$

2.3. Model tuning and validation

Our code performs k_1-fold cross-validation repeated r times, and model parameters were tuned using a k_2-fold CV on the training data. CV involves splitting the data into k folds, training on all but one, and evaluating performance on the residual fold. The process rotates trough the folds, averaging performances. CV can also be employed for model tuning to identify the hyperparameters with the best averaged performance. In nested CV both techniques are used simultaneously: the external loop defines testing data while internal CV tunes hyperparameters. Repeated CV splits the data differently each time to reduce the validation bias, and stratified CV ensures similar event rates across the folds. SRF and DeepHit are tuned by C-index across a wide grid of hyperparameters (see Supplementary Table 1); Cox-Lasso uses "glm" package. Our code allows random hyperparameter search with user-defined search sizes [41].

To compare the models, we validate them with the same number of repetitions, CV folds, and random seeds that define the data splits, ensuring the performance estimates corresponds to the same train/test sets. The superiority of one model is assessed with a one-sided paired Student's *t*-test applied to the performance metrics, with p-value < 0.05 considered significant, using the C-index to define the difference.

In the experiments for this paper, we used $k_1 = 5$, $k_2 = 3$, and $r = 10$, and the prediction time was set to the 0.90-th quantile of the observed event times. For clinical data, we set *early_stopping* $= 0$, *weight_decay* $= 0$, *activation_function* $=$ '*relu*', and performed a random grid search over 100 combinations for DeepHit tuning; *n_tree* $= 300$ and random search over 25 combinations was used for SRF. For simulations, DeepHit hyperparameters were set to *early_stopping* $=$ *FALSE*, *weight_decay* $= 0$, *epochs* $= 100$, optimized *learning_rate* from (0.001, 0.01, 0.1), *batch_size* was 100 for datasets sized up to 500, and 250 for larger samples, *mod_alpha* $= 0.2$, *sigma* $= 0.1$, *num_nodes* was chosen from the combinations of 1, 2, or 3 layers, containing 4, 8, or 16 nodes each. For the SRF we used *n_trees* $= 300$, *mtry* $= 3$, *n_depth* $= 5$.

2.4. Software employed

Our code relies on a number of R packages for fitting individual models and computing performance metrics. Namely, the package 'survival' for Cox-PH fitting; 'glm' package for Cox-Lasso [42]; 'randomForestSRC' for SRF with log-rank splitting [9]; 'survivalmodels' package for DeepHit [43]. C-index was computed using the 'concordance' command from 'survival' package [44]. Time-dependent AUC-ROC computations utilised 'timeROC' [45]. All calculations were run using R version 4.3.1, on the Intel® Core™ i7-9700 K CPU, 3.60 GHz processor, 16 GB RAM.

2.5. Code example

The full version of our package 'survcompare' can be downloaded or installed from GitHub [46]. A shorter version, supporting SRF but not DeepHit, is available from the Comprehensive R Archive Network (CRAN) [47] and can be installed as 'install.packages("survcompare")'. In this example, we simulated a dataset using the function 'simulate_nonlinear(N = 200)', with the columns 'time' and 'event' defining the survival outcome, as required by the package. The function survcompare() performs a repeated CV for Cox-Lasso and SRF, and returns an output, similar to that in Fig. 2. The map of the package functions and other code examples are provided in the Supplementary materials.

2.6. Datasets

We employed six clinical datasets (Table 1).

- English Longitudinal Study of Ageing (ELSA) is an ongoing multidisciplinary study of the older UK residents (aged > 50) with publicly available data [48,49]. We utilised processed data from the [50] for the outcome of new diagnoses of type 2 diabetes.
- Alzheimer's Disease Neuroimaging Initiative data (ADNI) is an ongoing longitudinal study (https://adni.loni.usc.edu/about/) promoting Alzheimer's disease research [51]. We analysed the processed data used in [3] available from the authors by request and after getting the ADNI approvals.
- Foot ulcer study (FUS). The data were collected by Ismail and colleagues [52] for the original study investigating the impact of depression on mortality among the people diagnosed with diabetic foot ulcer.
- Worcester Heart Attack Study (W500) included 500 patients admitted to Worcester's hospitals, USA, between 1975 and 2001 with an acute myocardial infarction to investigate mortality risk. We analysed data stored in scikit-learn package.
- Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) [53]. The data is available in the 'pycox' Python package [54] and includes information of 8873 severely ill hospitalised adults and 14 predictors for the risk of death.
- Head and neck squamous cell carcinoma data (HNSCC). HNSCC data were collected by the Anderson Cancer Center Head and Neck Quantitative Imaging Working Group [55]. We used a processed HNSCC dataset analysed by Yang et al. [13].
- Simulated datasets. The generating function was the same as described in our previous work [30], and largely similar to simulations in the 'survivalmodels' package [43]. We simulated datasets of the sizes 200 to 5000, and validated model performances on an independently simulated testing dataset (n = 5000). The experiment was repeated 50 times to compute the means and standard deviations of the performance metrics.

More detailed information can be found in the Supplementary Materials.

```
> mydata <- simulate_nonlinear(N=200)
> mypredictors <- names(mydata)[1:4]
> survcompare(
+   mydata,
+   mypredictors,
+   fixed_time = 9,
+   randomseed = 100,
+   repeat_cv = 3,
+   outer_cv = 5,
+   inner_cv = 3,
+   useCoxLasso = TRUE
+ )
[1] "Cross-validating CoxLasso using 3 repeat(s), 5 outer, 3 inner loops)."
[1] "Repeated CV 1 / 3"
  |=============================================================================================| 100%
[1] "Repeated CV 2 / 3"
  |=============================================================================================| 100%
[1] "Repeated CV 3 / 3"
  |=============================================================================================| 100%
Time difference of 1.533418 secs
[1] "Cross-validating Survival Random Forest using 3 repeat(s), 5 outer, 3 inner loops)."
[1] "Repeated CV 1 / 3"
  |=============================================================================================| 100%
[1] "Repeated CV 2 / 3"
  |=============================================================================================| 100%
[1] "Repeated CV 3 / 3"
  |=============================================================================================| 100%
Time difference of 22.36462 secs

Internally validated test performance of CoxLasso and Survival Random Forest over 3 repeated 5 fold cross-vali
dations (inner k = 3 ). Mean performance:
                         T C_score AUCROC Calib_slope   sec
CoxLasso                 9  0.5425 0.5425      0.4286  1.53
Survival Random Forest   9  0.6988 0.6960      0.8692 22.36
Diff                     0  0.1563 0.1535      0.4406 20.83
pvalue                 NaN  0.0000 0.0000      0.0528   NaN

Median performance:
                         T C_score AUCROC Calib_slope   sec
CoxLasso                 9  0.5300 0.5172      0.4880  1.53
Survival Random Forest   9  0.7143 0.7301      0.6867 22.36
Diff                     0  0.1843 0.2129      0.1987 20.83
pvalue                 NaN  0.0000 0.0000      0.0528   NaN

Survival Random Forest has outperformed CoxLassoby 0.1563 in C-index.
The difference is statistically significant with the p-value 1.72e-06***.
The supplied data may contain non-linear or cross-term dependencies,
better captured by Survival Random Forest.
Mean C-score:
  CoxLasso  0.5425(95CI=0.5301-0.5638;SD=0.0199)
  Survival Random Forest 0.6988(95CI=0.6896-0.7049;SD=0.0086)
Mean AUCROC:
  CoxLasso  0.5425(95CI=0.5194-0.5698;SD=0.0269)
Survival Random Forest 0.696(95CI=0.6914-0.7026;SD=0.0063)
```

Fig. 2. Code example.

Table 1
Dataset descriptions.

Dataset	Event description	N	Events observed	Predictors	Mean observation time	Data domains
ELSA	Type 2 diabetes	5957	456 (8 %)	13	8.7 years	Epidemiological
SUPPORT	Hospital mortality	8873	6036(68 %)	14	479 days	Clinical
ADNI	Alzheimer's disease	285	34 (12 %)	55	33 months	Imaging
FUS	Death	253	92 (32 %)	10	35 months	Clinical
W500	Death	500	215 (43 %)	16	2.4 years	Clinical
HNSCC	Death	451	56(12 %)	107	6.6 years	Imaging
Linear	simulated	up to 5000	20 %	4	10	–
Non-linear	simulated	up to 5000	20 %	4	10	–
Cross-terms	simulated	up to 5000	20 %	4	10	–

3. Results

3.1. Clinical data analysis

Table 2 and Fig. 3 present the results of repeated nested cross-validation for six prognostic clinical models using Cox-PH, Cox-Lasso, SRF, DeepHit, and their ensembles, showing the C-index. Table 3 displays the calibration slopes. Machine learning (ML) models outperformed the Cox models in 3 out of 6 datasets, with significant variation. No difference was observed for ELSA, W500, and FUS studies, while differences in the C-index were 0.0384 for SUPPORT, 0.0606 for ADNI, and 0.1057 for HNSCC (detailed in Supplementary Table 2).

Calibration was generally better in the models that also showed better discrimination (Table 3). In datasets where Cox-PH was not outperformed (ELSA, FUS, W500), Cox models' calibration slopes were nearly perfect. SRF and DeepHit tended to underfit (slope > 1), but this improved when ensembled with Cox-PH. The ensembled approach worked better with DeepHit, while SRF sequential ensemble (Ensemble 1) underfit considerably with calibration slope close to 2. Where ML models outperformed (SUPPORT, HNSCC), SRF ensembles demonstrated better calibration than other models. DeepHit was overfit, especially in smaller datasets (ADNI, HNSCC). Worth highlighting Cox-Lasso's close-to-optimal calibration in the SUPPORT and ADNI datasets. However, Cox-Lasso did not show good discrimination in ADNI, and it is important to consider both metrics.

3.2. Analysis of the stacked ensembles of clinical and simulated data

As described in the methods, alternative analysis of the ML outperformance is possible by computing the lambdas in the stacked ensembles of Cox-PH and ML models (Equation (7). Fig. 4 and Supplementary Table 2 show lambdas for all datasets across the splits in repeated nested CV. In simulated data, both DeepHit and SRF stacked ensembles correctly learned to rely on Cox-PH predictions in the linear (lambda < 0.1), and on ML predictions in the non-linear and cross-terms datasets (lambda > 0.90). In clinical data, low lambdas were seen in the clinical datasets with no ML outperformance: mean lambda for ELSA was 0.14. 0.32 for FUS and 0.20 for W500 for SRF stacked ensemble. In accordance with ML's outperformance in C-index, the contribution of the ML to the stacked ensemble predictions was high (> 0.70) in the SUPPORT and HNSCC datasets. In ADNI, DeepHit's lambda was 0.40, while SRF's was 0.97, indicating a better fit of SRF compared to DeepHit, also seen in the C-index results.

3.3. Learning curves of the ML models in the simulated data

Fig. 5 presents the simulated learning curves for DeepHit and SRF. Each subplot shows the mean performance of Cox-PH, ML models, and Cox-PH ensembles on linear, nonlinear, and cross-terms data. As expected, ML models consistently outperformed Cox-PH on nonlinear and cross-terms data, even with small sample sizes ($\geq$500). For linear data, Cox-PH was always superior, though the difference diminished with the sample size increase.

3.4. Computational time

We estimated computational times required for the 10 repeats of 5-fold nested CV with 3 internal folds and 100 random grid searches for DeepHit and 25 for SRF (Supplementary Table 4). Averaging across the datasets and normalizing to 1000 observations, Cox-PH and Cox-Lasso could be validated in 3 seconds, SRF and its ensembles in under 10 minutes, DeepHit and its sequential ensemble in 1–1.3 hours, and its stacked ensemble in 6 about hours due to additional out-of-sample computations to tune lambda.

4. Discussion

Accurate and interpretable models are essential for clinical decision-making, which is increasingly acknowledged by the healthcare professionals, regulators, and patients [18,19]. Generalized Linear models (GLMs) are considered transparent and interpretable, while "black-box" state-of-the-art ML models handle complex data and can achieve higher accuracy [11,30,56]. We focused on time-to-event medical data and evaluated the performance of Survival Random Forest (SRF) and DeepHit against the traditional Cox-PH and Cox-Lasso models, examining the trade-offs between model complexity and transparency. To facilitate this, we developed a set of R functions to perform nested cross-validation and compare the predictive performance of these models.

Our analysis of the clinical data revealed that the superiority of ML models was largely domain and size dependent. Imaging datasets benefited significantly from the ML, as well as a larger non-imaging data, that included prediction of cancer mortality for HNSCC study, Alzheimer's disease progression in the ADNI data, and hospital mortality in the SUPPORT study. However, in other epidemiological studies (prediction of diabetes in ELSA, cancer mortality in W500, and mortality in FUS), Cox-PH performed comparably to the ML. The ensemble models, combining Cox-PH with ML predictions, provided another quantification of the potential ML's ability to capture data complexities beyond Cox-PH. For example, in the SUPPORT and HNSCC studies, the stacked ensembles mostly relied on the ML component, with $\lambda > 0.70$.

Table 2
Validated performances of the SRF, DeepHit, their ensembles with Cox-PH, and their outperformance of the baseline Cox models.

C-index, mean (SD)	Cox-PH	Cox-Lasso	SRF	SRF Ens1	SRF Stack Ens2	DeepHit	DeepHit Ens1	DeepHit Stack Ens2	ML outperformed
ELSA	**0.7522** (0.0029)[b]	0.7523 (0.0028)	0.7344 (0.0042)	0.7292 (0.0042)	**0.7529** (0.0029)	0.7469 (0.003)	0.7406 (0.0068)	0.7524 (0.0032)	NO
FUS	0.7240 (0.0100)[b]	0.7232 (0.0088)	0.7138 (0.0147)	0.7089 (0.0109)	0.7163 (0.0092)	0.7184 (0.0121)	0.7195 (0.0112)	**0.7254 (0.0126)**	NO
W500	**0.7715 (0.0046)**[b]	0.7699 (0.0064)	0.7692 (0.0053)	0.7694 (0.0054)	0.7762 (0.0043)	0.7587 (0.0084)	0.7616 (0.0088)	**0.7710 (0.0050)**	NO
SUPPORT	0.5706 (0.0007)[b]	0.5715 (0.0006)	0.609 (0.0012) ***	0.6093 (0.0012) ***	**0.6088** (0.0011) ***	0.5914 (0.0049) ***	0.5825 (0.0048) ***	0.5916 (0.0063) ***	YES
ADNI	0.5576 (0.0489)	0.7120 (0.0501)[b]	**0.7726 (0.0178)** ***	0.7708 (0.0182) ***	0.7684 (0.0184) **	0.7263 (0.0529)	0.7164 (0.0587)	0.7301 (0.0371)	YES
HNSCC	0.547 (0.0462)	0.6109 (0.0496)[b]	**0.7189 (0.0104)** ***	0.7165 (0.0114) ***	0.7130 (0.0140) ***	0.6833 (0.0240) ***	0.6983 (0.0238) ***	0.6925 (0.0245) ***	YES

*** – *p-value* < *0.001*, ** – *p-value* < *0.01*, * – *p-value* < *0.05; SD- standard deviation*
[b]- baseline model, Cox-PH for ELSA, SUPPORT, FUS, W500, and Cox-Lasso for ADNI and HNSCC.
The table shows the mean and standard deviation of the validated C-scores over the 10 repeated 5-fold CVs. The stars indicate statistical significance of the outperformance of the ML model over the baseline Cox model.

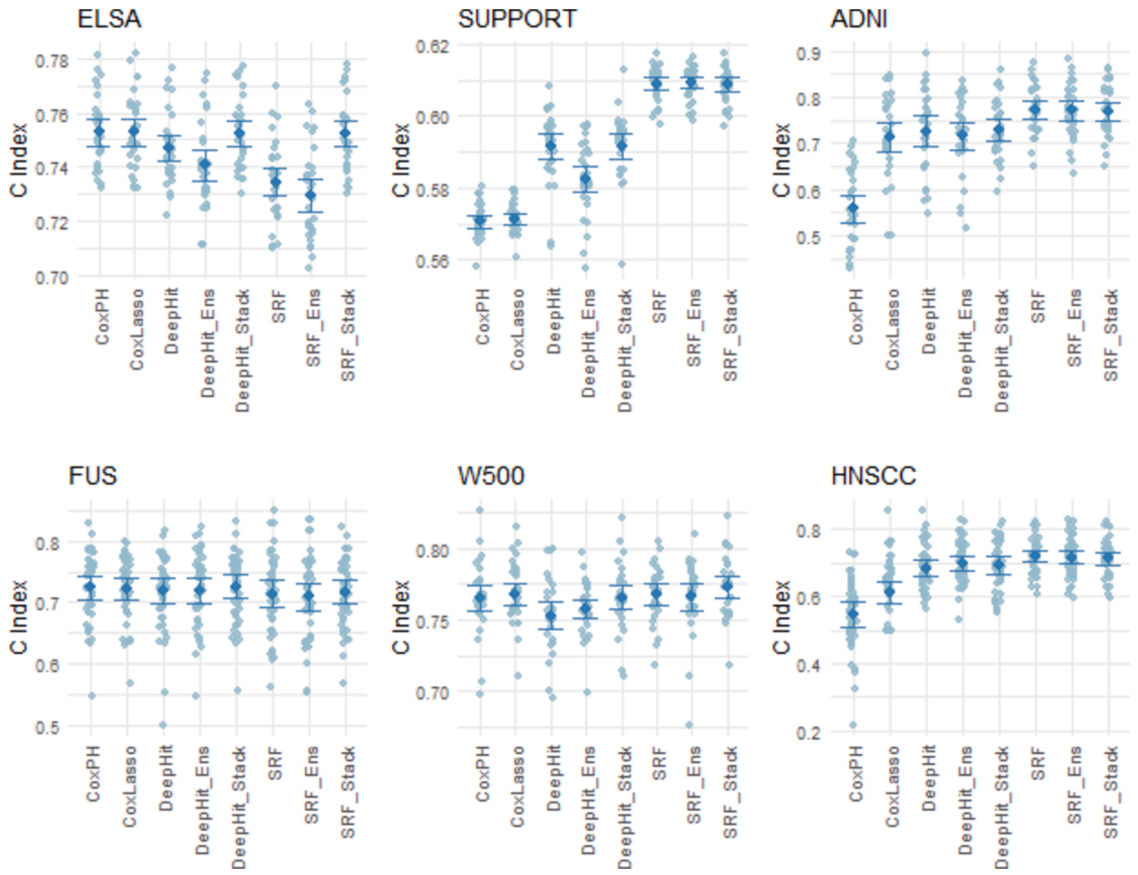


Fig. 3. Validated C-scores for the clinical datasets. Estimated mean C-index with the 95% confidence intervals for the mean estimated from the 10 repeated 5-fold cross-validations.

Table 3
Validated calibration slopes of the Cox-PH, SRF, DeepHit and their sequential and stacked ensembles.

Calib slope, median (SD)	Cox-PH	Cox-Lasso	SRF	SRF Ens1	SRF Stack Ens2	DeepHit	DeepHit Ens1	DeepHit Stack Esn2
ELSA	**0.92 (0.02)**	1.16 (0.02)	1.26 (0.37)	2.28 (0.28)	**0.95 (0.05)**	0.86 (0.46)	0.69 (0.05)	0.91 (0.02)
FUS	0.74 (0.07)	0.95 (0.08)	1.45 (0.2)	1.78 (0.27)	1.02 (0.22)	1.38 (1.67)	**1.01 (0.17)**	0.94 (0.19)
W500	**0.81 (0.08)**	1.30 (0.04)	1.50 (0.28)	2.26 (0.51)	1.23 (0.17)	0.86 (0.48)	**0.89 (0.16)**	0.89 (0.1)
SUPPORT	0.98 (0.01)	1.04 (0.01)	0.96 (0.06)	1.12 (0.03)	**0.96 (0.03)**	0.66 (0.08)	0.71 (0.06)	0.84 (0.04)
ADNI	0.03 (0.03)	1.13 (1.26)	**0.75 (0.11)**	0.77 (0.15)	0.74 (0.13)	0.39 (0.3)	0.41 (0.17)	0.77 (0.38)
HNSCC	0.05 (0.05)	0.37 (0.38)	**0.71 (0.04)**	0.68 (0.08)	0.68 (0.08)	0.87 (0.16)	0.93 (0.18)	0.96 (0.23)

Median and standard deviation of the validated calibration slopes over the 10 repeated 5-fold CV (apart from ELSA and SUPPORT with 3-fold CVs). In bold are the calibration scores which correspond to the models with the highest C-index, as indicated in Table 2.

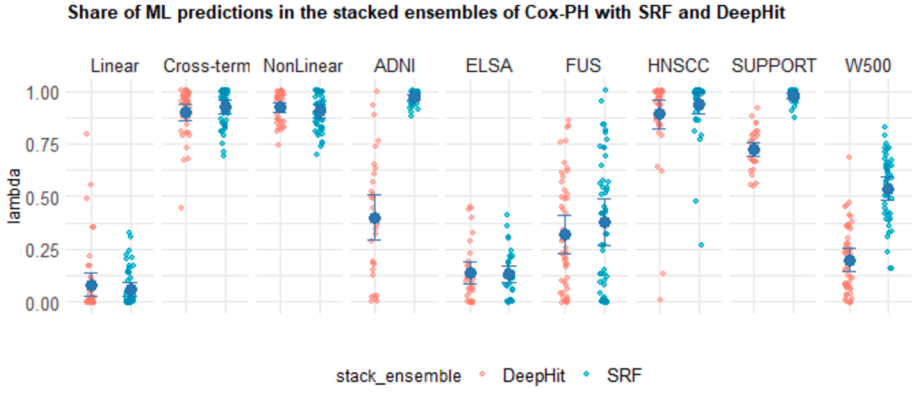


Fig. 4. Share of the ML predictions in the stacked ensembles of Cox-PH (lambda parameter) with SRF and DeepHit.

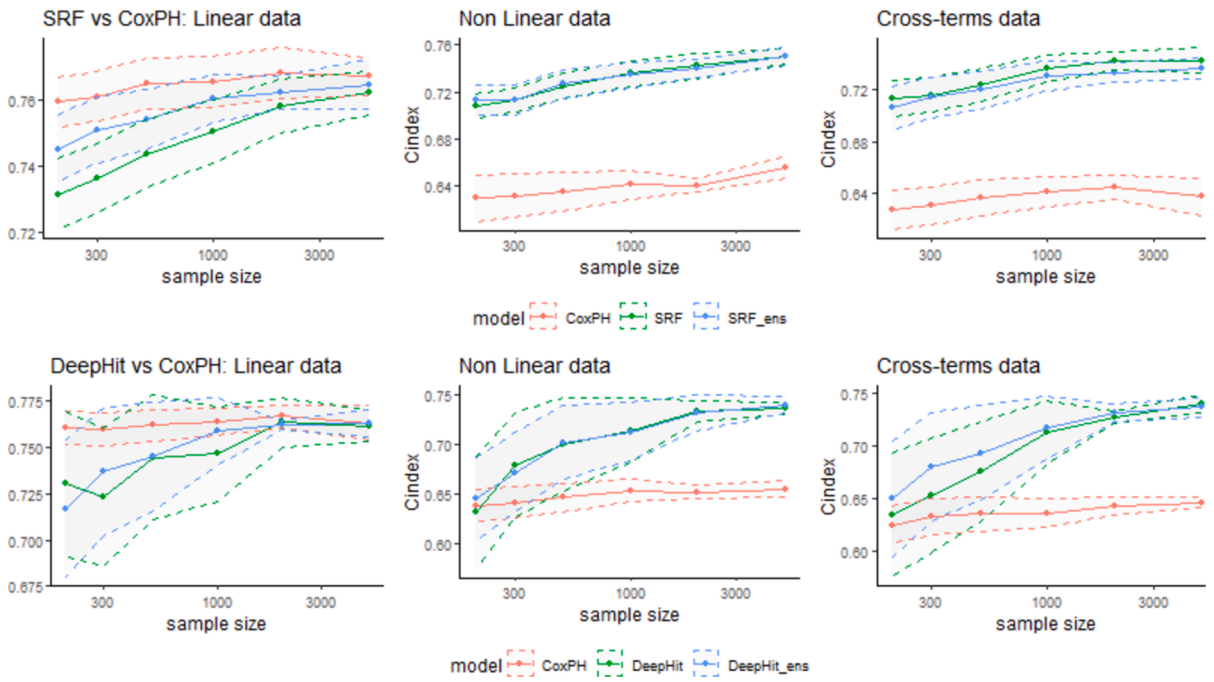


Fig. 5. Learning curves of Cox-PH, SRF, and their ensembles for the simulated datasets.

Conversely, in ELSA and W500, the share of ML's contribution to the stacked ensemble was minimal (0.1–0.3), reinforcing that Cox-PH captured the majority of the predictive information.

The investigation of the learning curves revealed that the Cox models can achieve near-optimal performance even with relatively small sample size (200–500 observations). In contrast, ML models required larger datasets (1000–5000 observations) to fully leverage their feature transformation capabilities. Notably, the benefit of incorporating Cox-PH predictions into more complex models, such as SRF, was evident in smaller datasets, where ML models struggled to capture the generalisable data's complexity.

4.1. *Our findings in the context of previous research*

Other studies also reported that classical methods can perform as good as ML models in predicting health outcomes, or only marginally inferior. A systematic review by De Silva et al. (2020) [57] found that among 27 ML models of type 2 diabetes, the pooled C-index for neural networks was 0.825, closely followed by logistic regression at 0.815. If interpretability is favoured, such minor differences may not justify choosing ML models over logistic regression, particularly given longer computation times. Nusinovici et al. (2020) [58] and Wu et al. (2023) [59] found that in low-dimensional settings logistic regression performed on par with ML models. The preference of Cox-PH in low sample size was also reported in the simulation study by Baralou et al. (2023) [56]. We further found that simple solutions such as opting for Cox-Lasso could recover a part of ML's outperformance in high-dimensional data such as ADNI, and HNSCC. Similarly, Baralou et al. advocates for the inclusion of splines in Cox-PH to handle potential non-linearity [56].

Further, our results align with previous analyses of the SUPPORT, ADNI, and W500 data [11,13,60]. Some variations could be due to the differences in model tuning, data heterogeneity and randomness of train-test splits. Unlike some prior studies, our code averages performance estimates over multiple splits (50 test performances are generated by 10 repeated 5-fold CVs), which provides a more robust assessment of the model performance and its stability. The variability of results is further illustrated in the histograms of C-index differences (Supplementary Fig. 1).

4.2. *Calibration for clinical prediction modelling*

While good discrimination enables the identification of high-risk individuals, calibration, or accurate probability estimates, are crucial for medical decision-making, yet calibration measures are rarely reported [28,29,61]. Traditional models like Cox-PH, or Logistic Regression, typically exhibit good calibration, as they are optimized to match observed event rates. ML models such as SRF may luck inherent calibration mechanisms. Other studies highlight the importance of higher sample size for calibration of the ML models [62]. In our analysis, the Cox models resulted in better calibration in the data with limited ML outperformance, while the ensembles incorporating Cox-PH predictions into SRF have shown improved calibration where SRF itself had failed (Table 3).

4.3. *Future directions and methodological considerations*

This study focused on SRF and DeepHit as ML alternatives to the Cox regression. Other survival models can be tested, and our code can offer a basis for such extensions. Second, external validity, such as performance assessment in data collected from different geographic locations, or time periods, remained untested and can be further explored [5,63], including validation in synthetic data [64]. Third, we acknowledge the emergence of the interpretable ML models such as rule-based, prototype-based, and ensemble methods [30,65]. Such models may offer both interpretability and predictive superiority, though GLMs could still challenge them in computational efficiency and prediction stability. Future research could include the comparisons of the Cox-PH with such models. Finally, future package versions will include missing data handling.

4.4. *Conclusion*

We presented a systematic approach for assessing the predictive value of complex data relationships in survival analysis. The findings indicate that Cox-PH or Cox-Lasso may be sufficient in many clinical applications, while embedding Cox-PH predictions into the ML models such as SRF or DeepHit may help assessing the predictive advantage of the ML and improve calibration. Our methodology and R package

"survcompare" offer researchers a practical framework for diligent model validation and assessing the necessity of complex models in clinical prediction settings.

5. Code availability

Code: The full version of our R package 'survcompare' can be downloaded or installed from GitHub page, https://github.com/dianashams/survcompare/tree/DeepHit [46]. A shorter version, supporting SRF but not DeepHit, is available from CRAN https://github.com/cran/survcompare [47] and can be installed in a standard way.

The data used in this project were publicly available apart from the FUS as described below. The secondary analyses of the anonymised data performed in this study did not need ethical approval. Here, we provide the approval information for the original studies and data collections.

- English Longitudinal Study of Ageing. The ELSA data are managed by the UK Data Services [66] and is publicly available from https://www.elsa-project.ac.uk/. All participants gave informed consent, ethical approvals were received for each data collection wave as detailed in https://www.elsa-project.ac.uk/ethical-approval.
- Alzheimer's Disease Neuroimaging Initiative data. The ADNI data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defence award number W81XWH-12–2-0012). The study declared compliance with the ethics principles derived from international ethics guidelines, including the Declaration of Helsinki and Council for International Organizations of Medical Sciences (CIOMS) International Ethical Guidelines. The study further complied with Title 21 of the United States Code of Federal Regulations (US 21 CFR) regarding clinical studies, including Part 50 and Part 56 concerning informed consent and IRB regulations and applicable sections of US 21 CFR Part 312. All participants gave informed consent. The data is accessible as detailed in https://adni.loni.usc.edu/data-samples/access-data/.
- Foot ulcer study data. The study protocol was approved by the ethics committees of King's College London, and the local participating National Health Service Trusts; all participants provided written informed consent. The presented secondary analysis of the anonymized data does not require separate approval. The anonymized data can be shared with approved researchers upon request, ethical approval, and permission.
- Worcester Heart Attack Study. The data were used in the textbook by Hosmer and Lemeshow [67] and publicly available from https://web.archive.org/web/20170517071528/http://www.umass.edu/statdata/statdata/data/whas500.txt, or using Python's scikit-survival package and the function 'sksurv.datasets.load_whas500()' https://scikit-survival.readthedocs.io/en/stable/api/generated/sksurv.datasets.load_whas500.html.
- Head and neck squamous cell carcinoma data. The data were collected by the Anderson Cancer Center Head and Neck Quantitative Imaging Working Group [55] and are publicly available at The Cancer Imaging Archive [68] as detailed in https://www.cancerimagingarchive.net/collection/hnscc/. In this project, we used a pre-processed HNSCC dataset analysed by Yang et al. (2022) [13] and can be downloaded from https://github.com/lasso-net/lassonet/tree/master/examples/data [69].
- Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) [53]. The data is available in the 'pycox' Python package and includes information of 8873 severely ill hospitalised adults and 14 predictors for the risk of death [54].

CRediT authorship contribution statement

Diana Shamsutdinova: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Daniel Stamate:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Daniel Stahl:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Funding

D Shamsutdinova and D Stahl are funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, NIHR or the Department of Health and Social Care. This paper represents independent research part-funded by the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the author(s) used Grammarly (https://www.grammarly.com) in order to improve the readability of the manuscript, including spelling and grammar check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Diana Shamsutdinova reports financial support was provided by National Institute for Health Research Maudsley Biomedical Research Centre. Daniel Stahl reports financial support was provided by National Institute for Health Research Maudsley Biomedical Research Centre. Daniel Stamate reports financial support was provided by Alzheimer's Research UK. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Dr Khalida Ismail and Dr Kirsty Winkler, both King's College London, for the collection and sharing the Foot ulcer study (FUS) data, and The Wellcome Trust and NIHR UK for supporting the FUS study.

Consent for publication

Not applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2024.105700.

References

[1] D. Stamate, H. Musto, O. Ajnakina, D. Stahl, Predicting Risk of Dementia with Survival Machine Learning and Statistical Methods: Results on the English

Longitudinal Study of Ageing Cohort, in: 2022: pp. 436–447. DOI: 10.1007/978-3-031-08341-9_35.
[2] J.H.F. Oosterhoff, A.A.H. de Hond, R.M. Peters, L.N. van Steenbergen, J.C. Sorel, W.P. Zijlstra, R.W. Poolman, D. Ring, P.C. Jutte, G.M.M.J. Kerkhoffs, H. Putter, E. W. Steyerberg, J.N. Doornberg, M.L. Consortium, Machine learning did not outperform conventional competing risk modeling to predict revision arthroplasty., Clin Orthop Relat Res (2024).
[3] H. Musto, D. Stamate, I. Pu, D. Stahl, Predicting Alzheimer's Disease Diagnosis Risk Over Time with Survival Machine Learning on the ADNI Cohort, in: International Conference on Computational Collective Intelligence, 2023: pp. 700–712.
[4] O. Ajnakina, D. Agbedjro, R. McCammon, J. Faul, R.M. Murray, D. Stahl, A. Steptoe, Development and validation of prediction model to estimate 10-year risk of all-cause mortality using modern statistical learning methods: a large population-based cohort study and external validation, BMC Med. Res. Method. 21 (2021) 1–11.
[5] B. Perry, F. Vandenberghe, E.F. Osimo, C. Grosu, M. Piras, P. Jones, P. Mallikarjun, J. Stochl, R. Upthegrove, G. Khandaker, others, An International External Validation and Revision of the PsyMetRiC Cardiometabolic Risk Prediction Algorithm for Young People with Psychosis, European Psychiatry 65 (2022) S676–S677.
[6] D.R. Cox, Regression models and life-tables, J. Roy. Stat. Soc.: Ser. B (Methodol.) 34 (1972) 187–202.
[7] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[8] A. Barnwal, H. Cho, T. Hocking, Survival regression with accelerated failure time model in XGBoost, J. Comput. Graph. Stat. 31 (2022) 1292–1302.
[9] H. Ishwaran, M.S. Lauer, E.H. Blackstone, M. Lu, U.B. Kogalur, Randomforestsrc: Random survival forests vignette, (2021).
[10] C. Lee, W. Zame, J. Yoon, M. Van der Schaar, DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks, in: Proceedings of the AAAI Conference on Artificial Intelligence 32, 2018, https://doi.org/10.1609/aaai.v32i1.11842.
[11] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Med. Res. Method. 18 (2018) 1–12.
[12] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, A. Bender, Deep learning for survival analysis: a review, Artif. Intell. Rev. 57 (2024) 65, https://doi.org/10.1007/s10462-023-10681-3.
[13] X. Yang, L. Abraham, S. Kim, P. Smirnov, F. Ruan, B. Haibe-Kains, R. Tibshirani, FastCPH: Efficient Survival Analysis for Neural Networks, ArXiv Preprint ArXiv: 2208.09793 (2022).
[14] M.T. Ribeiro, S. Singh, C. Guestrin, " Why should i trust you?" Explaining the predictions of any classifier, in, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[15] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv Neural Inf Process Syst 30 (2017).
[16] T.P. Quinn, S. Jacobs, M. Senadeera, V. Le, S. Coghlan, The three ghosts of medical AI: Can the black-box present deliver? Artif. Intell. Med. 124 (2022) 102158.
[17] G. Erion, J.D. Janizek, C. Hudelson, R.B. Utarnachitt, A.M. McCoy, M.R. Sayre, N. J. White, S.-I. Lee, A cost-aware framework for the development of AI models for healthcare applications, Nat. Biomed. Eng. 6 (2022) 1384–1398, https://doi.org/10.1038/s41551-022-00872-8.
[18] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215.
[19] D.W. Joyce, A. Kormilitzin, K.A. Smith, A. Cipriani, Explainable artificial intelligence for mental health through transparency and interpretability for understandability, NPJ Digit Med 6 (2023) 6.
[20] M. Krzyziński, M. Spytek, H. Baniecki, P. Biecek, SurvSHAP(t): Time-dependent explanations of machine learning survival models, Knowl Based Syst 262 (2023) 110234, https://doi.org/10.1016/j.knosys.2022.110234.
[21] M.S. Kovalev, L.V. Utkin, E.M. Kasimov, SurvLIME: A method for explaining machine learning survival models, Knowl Based Syst 203 (2020) 106164, https://doi.org/10.1016/j.knosys.2020.106164.
[22] L.V. Utkin, E.D. Satyukov, A.V. Konstantinov, SurvNAM: The machine learning survival model explanation, Neural Netw. 147 (2022) 81–102, https://doi.org/10.1016/j.neunet.2021.12.015.
[23] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, J R Stat Soc Series B Stat Methodol 58 (1996) 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[24] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, M.J. Van Der Laan, Survival ensembles, Biostatistics 7 (2006) 355–373.
[25] M. Leblanc, J. Crowley, Survival Trees by Goodness of Split, J. Am. Stat. Assoc. 88 (1993) 457–467.
[26] A. Shimokawa, Y. Kawasaki, E. Miyaoka, Comparison of splitting methods on survival tree, Int. J. Biostat. 11 (2015) 175–188.
[27] N. Korepanova, H. Seibold, V. Steffen, T. Hothorn, Survival forests under test: Impact of the proportional hazards assumption on prognostic and predictive forests for amyotrophic lateral sclerosis survival, Stat. Methods Med. Res. 29 (2020) 1403–1419.
[28] L. Famiglini, A. Campagner, F. Cabitza, others, Towards a Rigorous Calibration Assessment Framework: Advancements in Metrics, Methods, and Use, Frontiers in Artificial Intelligence and Applications 372 (2023) 645–652.
[29] B. Van Calster, D.J. McLernon, M. Van Smeden, L. Wynants, E.W. Steyerberg, topic group evaluating diagnostic tests, prediction models of the STRATOS initiative, Calibration: the Achilles heel of predictive analytics, BMC Med. 17 (2019) 230.
[30] D. Shamsutdinova, D. Stamate, A. Roberts, D. Stahl, Combining Cox Model and Tree-Based Algorithms to Boost Performance and Preserve Interpretability for Health Outcomes, in: Proceedings of the Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Part II, Springer International Publishing, 2022: pp. 170–181.
[31] D. Shamsutdinova, J. Das-Munshi, M. Ashworth, A. Roberts, D. Stahl, Predicting type 2 diabetes prevalence for people with severe mental illness in a multi-ethnic East London population, Int. J. Med. Inf. 172 (2023) 105019, https://doi.org/10.1016/j.ijmedinf.2023.105019.
[32] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (1992) 241–259, https://doi.org/10.1016/S0893-6080(05)80023-1.
[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in python journal of machine learning research, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[34] E.W. Steyerberg, Clinical prediction models: A practical approach to development, validation, and updating, Springer International Publishing, 2019.
[35] F.E. Harrell Jr, K.L. Lee, R.M. Califf, D.B. Pryor, R.A. Rosati, Regression modelling strategies for improved prognostic prediction, Stat. Med. 3 (1984) 143–152.
[36] P. Blanche, J. Dartigues, H. Jacqmin-Gadda, Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks, Stat. Med. 32 (2013) 5381–5397, https://doi.org/10.1002/sim.5958.
[37] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36, https://doi.org/10.1148/radiology.143.1.7063747.
[38] H. Uno, T. Cai, M.J. Pencina, R.B. D'Agostino, L.-J. Wei, On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data, Stat. Med. 30 (2011) 1105–1117.
[39] P. Blanche, A. Latouche, V. Viallon, Time-dependent AUC with right-censored data: a survey study, (2012).
[40] C.S. Crowson, E.J. Atkinson, T.M. Therneau, Assessing calibration of prognostic risk scores, Stat. Methods Med. Res. 25 (2016) 1692–1706.
[41] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: Theory and practice, Neurocomputing 415 (2020) 295–316, https://doi.org/10.1016/j.neucom.2020.07.061.
[42] J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, J. Stat. Softw. 33 (2010) 1–22.
[43] R. Sonabend, survivalmodels: Models for Survival Analysis, (2024). https://CRAN.R-project.org/package=survivalmodels.
[44] T. Therneau, E. Atkinson, The concordance statistic, A Package for Survival Analysis in R, Vignettes, 2022.
[45] P. Blanche, M.P. Blanche, Package 'timeROC, (2019).
[46] D. Shamsutdinova, D. Stahl, survcompare/DeepHit, (2024). https://github.com/dianashams/survcompare/tree/DeepHit (accessed October 14, 2024).
[47] D. Shamsutdinova, D. Stahl, survcompare, (2024). DOI: DOI: 10.32614/CRAN.package.survcompare.
[48] A. Steptoe, E. Breeze, J. Banks, J. Nazroo, Cohort profile: the English longitudinal study of ageing, Int. J. Epidemiol. 42 (2013) 1640–1648.
[49] A. et al. Steptoe, English Longitudinal Study of Ageing, (2024).
[50] D. Shamsutdinova, O. Ajnakina, A. Roberts, D. Stahl, Schizophrenia polygenic risk score and type 2 diabetes onset in older adults with no schizophrenia diagnosis, Psychiatr. Genet. 33 (2023) 191–201.
[51] M.W. Weiner, Alzheimer's Disease Neuroimaging Initiative, (2024). https://adni.loni.usc.edu/about/ (accessed June 27, 2024).
[52] K. Ismail, K. Winkley, D. Stahl, T. Chalder, M. Edmonds, A cohort study of people with diabetes and their first foot ulcer: the role of depression on mortality, Diabetes Care 30 (2007) 1473–1479.
[53] W.A. Knaus, The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults, Ann. Intern. Med. 122 (1995) 191, https://doi.org/10.7326/0003-4819-122-3-199502010-00007.
[54] H. Kvamme, Ø. Borgan, I. Scheel, Time-to-event prediction with neural networks and Cox regression, J. Mach. Learn. Res. 20 (2019) 1–30.
[55] A.J. Grossberg, A.S.R. Mohamed, H. El Halawani, others, Data descriptor: Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy, Sci. Data 2018 (5) (2018) 1–10.
[56] V. Baralou, N. Kalpourtzi, G. Touloumi, Individual risk prediction: Comparing random forests with Cox proportional-hazards model by a simulation study, Biom. J. 65 (2023) 2100380.
[57] K. De Silva, W.K. Lee, A. Forbes, R.T. Demmer, C. Barton, J. Enticott, Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis, Int. J. Med. Inf. 143 (2020) 104268.
[58] S. Nusinovici, Y.C. Tham, M.Y.C. Yan, D.S.W. Ting, J. Li, C. Sabanayagam, T. Y. Wong, C.-Y. Cheng, Logistic regression was as good as machine learning for predicting major chronic diseases, J. Clin. Epidemiol. 122 (2020) 56–69.
[59] T. Wu, Y. Wei, J. Wu, B. Yi, H. Li, Logistic regression technique is comparable to complex machine learning algorithms in predicting cognitive impairment related to post intensive care syndrome, Sci. Rep. 13 (2023) 2485.
[60] H. Musto, D. Stamate, I. Pu, D. Stahl, Predicting Alzheimer's Disease Diagnosis Risk Over Time with Survival Machine Learning on the ADNI Cohort, in: Proceedings of the Computational Collective Intelligence: 15th International Conference, ICCCI 2023, Springer Nature Switzerland, 2023: pp. 700–712.

[61] V.E. Staartjes, J.M. Kernbach, Letter to the Editor. Importance of calibration assessment in machine learning–based predictive analytics, J. Neurosurg. Spine 32 (2020) 985–987, https://doi.org/10.3171/2019.12.SPINE191503.
[62] F.M. Ojeda, M.L. Jansen, A. Thiéry, S. Blankenberg, C. Weimar, M. Schmid, A. Ziegler, Calibrating machine learning approaches for probability estimation: A comprehensive comparison, Stat. Med. 42 (2023) 5451–5478, https://doi.org/10.1002/sim.9921.
[63] B. Van Calster, E.W. Steyerberg, L. Wynants, M. van Smeden, There is no such thing as a validated prediction model, BMC Med. 21 (2023) 70.
[64] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D.A. Calian, T.A. Mann, Improving Robustness using Generated Data, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, J.W. Vaughan (Eds.), Adv Neural Inf Process Syst, Curran Associates, Inc., 2021: pp. 4218–4233. https://proceedings.neurips.cc/paper_files/paper/2021/file/21ca6d0cf2f25c4dbb35d8dc0b679c3f-Paper.pdf.
[65] L. Xu, C. Guo, CoxNAM: An interpretable deep survival analysis model, Expert Syst. Appl. (2023) 120218.
[66] J. Banks, G.D. Batty, J. Breedvelt, K. Coughlin, I.F.F.S. (IFS) Crawford R., M. Marmot, J. Nazroo, I.F.F.S. (IFS) Oldfield Z., N. Steel, A. Steptoe, M. Wood, P. Zaninotto, English Longitudinal Study of Ageing: Waves 0-9, 1998-2019, UK Data Service (2021). DOI: 10.5255/ukda-sn-5050-24.
[67] D.W. Hosmer Jr, S. Lemeshow, S. May, Applied survival analysis: regression modeling of time-to-event data, John Wiley & Sons, 2008.
[68] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, others, The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository, J. Digit. Imaging 26 (2013) 1045–1057.
[69] I. Lemhadri, F. Ruan, L. Abraham, R. Tibshirani, LassoNet: a new family of models to incorporate feature selection and neural networks, (2022). https://github.com/lasso-net/lassonet.