New perspectives on semi-parametric and pseudo-value approaches for modeling clinical time-to-event data

Doctoral thesis
to obtain a doctorate (PhD)
from the Faculty of Medicine
of the University of Bonn

Alina Schenk
from Wittlich
2025

Written with authorization of
the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. Matthias Schmid
Second reviewer: Prof. Dr. Markus Neuhäuser

Day of oral examination: May 13, 2025

From the Institute of Medical Biometry, Informatics and Epidemiology

Table of Contents

List of abbreviations

AFT	Accelerated failure time
APS	Acute physiology score
GEE	Generalized estimation equation
HR	Hazard ratio
IML	Interpretable machine learning
PH	Proportional hazards
PIRATE	Pre-Interventional Risk Assessment in The Elderly
POSE	Peri-Interventional Outcome Study in the Elderly
PRT	Pseudo-value regression trees
PVRF	Pseudo-value random forest
RCT	Randomized controlled trial
RMST	Restricted mean survival time
SAPS	Simplified acute physiology score

1 Abstract

Clinical decision-making often relies on quantitative measures derived from statistical time-to-event models, enabling risk assessment through the quantification of survival probabilities. A key goal of these modeling approaches in guiding clinical decisions is to provide accurate risk estimates while using as little patient information as possible. Standard time-to-event modeling techniques often rely on restrictive assumptions, and their violation bear the risk of biased estimates. Furthermore, these models may need to be tailored to specific population groups, such as children or elderly patients.

The aim of this cumulative dissertation was to develop and evaluate new modeling approaches for clinical time-to-event outcomes, focusing on interpretability and applicability in clinical settings, minimal model assumptions, and the ability to filter out the most relevant patient information required for accurate risk estimation. To this end, this dissertation presents three modeling strategies that address the aforementioned goals.

In the first work, a tool for the pre-interventional risk assessment of 30-day mortality in the population of elderly patients was developed. Translating the underlying semi-parametric Cox model to a simple scoring system, this tool is user-friendly and only involves three risk factors. The development process focused on interpretability and applicability in clinical settings, while relying on a selection of the most relevant patient information within the Cox model framework.

To further reduce assumptions, the second and third works developed modeling approaches for risk assessment in terms of survival probabilities and the restricted mean survival time. These methods, which are based on pseudo-value regression and machine learning methods, demonstrate the reduction of assumptions compared to standard modeling techniques while being able to automatically select most relevant risk factors and interactions among them. The presented modeling approaches maintain interpretability and are able to quantify causal treatment effects, as illustrated in simulation studies and on clinical datasets.

All research articles included in this dissertation have been published in international peer-reviewed journals.

2 Introduction and aims with references

2.1 Time-to-event outcomes

In clinical practice, questions like *What is my expected survival time?* and *How likely am I to die within the next year?* are often raised by patients diagnosed with life-threatening diseases. This highlights the need for accurate risk communication, relying on the quantification of outcome measures investigated in time-to-event analysis. The aim of time-to-event analysis is the description of the survival time T and its association to covariates, such as age, sex, or treatment (Klein and Moeschberger, 2003). The survival time is defined as the duration from a specified starting point until the occurrence of a target event of interest (Klein and Moeschberger, 2003). In clinical research, these events include, for example, death, disease progression, or therapy success, with corresponding survival times measuring time from birth to death, time from study entry to disease progression, or time from intervention to therapy success. During their longitudinal follow-up, patients may drop-out of the study due to, e.g., relocation or withdrawal of consent. These drop-outs lead to partly incompletely observed survival times, a phenomenon known as *censoring* (Kalbfleisch and Prentice, 2002). Patients are *right-censored* at the last time when they were known to be event-free. If patients cannot be included in the study because their event had occurred before study begin, their survival time is considered *left-truncated*, another type of incompletely observed survival times (Kalbfleisch and Prentice, 2002). The methods discussed and applied in this dissertation account for both right-censoring and left-truncation.

The survival time can formally be described as a (continuous) random variable $T \in \mathbb{R}^+$. From this, several describing functions can be derived, with the most intuitive one being the *survival function* $S(t)$. This monotonically decreasing function gives the probability of surviving beyond a time point t and is defined as $S(t) = \mathrm{P}(T > t) \in [0, 1]$ for $0 \leq t < \infty$ (Kalbfleisch and Prentice, 2002). With the survival function, quantification of, e.g., the 30-day survival probability is possible. Moreover, the survival function can be used to derive *summary measures* of T, such as (restricted) mean and median survival times, which can conveniently be used for risk quantifica-

tion. The median survival time is defined by $t_{\text{med}} = \min\{t \mid S(t) \leq 0.5\}$ and represents the time at which 50 % of the patients have experienced the event of interest (Klein and Moeschberger, 2003). The quantification of life expectancy is enabled by the mean survival time given as $\mu = \mathrm{E}(T) = \int_0^\infty S(t)dt$. The restricted mean survival time (RMST) describes the life expectancy in $[0, \tau]$ with a time-horizon $\tau > 0$ and is given by $\mu(\tau) = \mathrm{E}(\min(T, \tau)) = \int_0^\tau S(t)dt$ (Klein and Moeschberger, 2003). Unlike the survival function, the *hazard function* $h(t)$ does not represent a probability but describes the non-negative instantaneous event rate at time t. The hazard function $h(t) = \lim_{\Delta \to 0^+} \frac{1}{\Delta} \mathrm{P}(t \leq T < t + \Delta \mid T \geq t)$ directly affects the slope of the survival function: higher event rates result in a strongly decreasing survival function, while lower event rates correspond to a more gradual decrease. The survival function can be expressed in terms of the cumulative hazard function $H(t)$ as $S(t) = \exp(-H(t))$, with $H(t) = \int_0^t h(u)du$ representing the accumulated hazard until a time point t (Kalbfleisch and Prentice, 2002).

2.2 Semi- and non-parametric approaches for analyzing time-to-event outcomes

In the following, a time-to-event dataset $\mathcal{D} = \{(\tilde{T}_i, \delta_i, X_i^T) \mid i = 1, \ldots, n\}$ containing n independent patients is considered. The covariate vector of patient i is denoted by $X_i \in \mathbb{R}^p$. The survival (censoring) time of patient i is denoted by T_i (C_i) and the observed time is given as $\tilde{T}_i = \min(T_i, C_i)$. Further, δ_i indicates whether patient i has been censored ($\delta_i = 0$) or whether the event of interest has been observed ($\delta_i = 1$) at $\tilde{T}_i$. Common examples to describe the survival time in a parametric fashion include the Exponential, Weibull, and Lognormal distributions. Those distributions provide closed formulas for the survival and the hazard functions (Kalbfleisch and Prentice, 2002). However, in some cases, none of the existing distributions may provide an adequate description of the available data $\mathcal{D}$. The most popular *non-parametric* estimator for the survival function is the Kaplan-Meier estimator, assuming independent censoring for a consistent estimation of $S(t)$ (Klein and Moeschberger, 2003). Here, independent censoring refers to the assumption that the additional information on censoring does not change the instantaneous event rate, i.e., $h(t)\Delta = \mathrm{P}(t \leq T < t + \Delta, T \leq C \mid \min(T, C) \geq t)$. The Kaplan-Meier estimator is constructed by relating the number of events

at time t to the number of patients still at risk at t without making any parametric assumptions on the distribution of T. While this estimator can be applied in different subgroups separately to investigate group effects, it does not account for other covariates in relation to survival times. The most popular approach to consider multiple covariates is the semi-parametric Cox proportional hazards (PH) model (Cox, 1972). Cox assumes the hazard to be given as $h(t \mid X_i) = h_0(t) \exp(\beta^T X_i)$ implying the hazard ratio $\mathrm{HR} = h(t \mid X_i)/h(t \mid X_j)$ of two patients $i \neq j$ being time-constant (PH assumption). The vector $\beta \in \mathbb{R}^p$ defines covariate effects which can be estimated consistently by maximizing the partial likelihood function under the assumption of independent censoring (Cox, 1972). The baseline hazard $h_0(t)$ is shared by all patients and independent of the covariates, so it does not contribute to the partial likelihood function. Instead of being estimated via the partial likelihood function, the cumulative baseline hazard function $H_0(t) = \int_0^t h_0(u)du$ can be estimated using the Breslow estimator $\hat{H}_0(t)$ (Klein and Moeschberger, 2003). The Cox model is considered semi-parametric because it assumes a parametric form for the covariates while treating the baseline hazard function $h_0(t)$ as a nuisance parameter. In contrast to the Cox model, accelerated failure time (AFT) models assume a direct relationship between the covariates and the logarithmic survival time expressed by $\ln(T_i) = \gamma^T X_{i.} + \varepsilon_i$ with $X_{i.}^T = \left(1, X_i^T\right)$ (Kalbfleisch and Prentice, 2002). Given this direct relationship, the acceleration factor $\exp(\gamma^T X_{i.})$ describes the multiplicative effect of X_i on T_i. While AFT models do not necessarily assume the HR to be time-constant, an assumption on the distribution of the error term ε implies a distributional assumption on T. For example, an extreme-value (normal) distribution for ε results in a Weibull (Lognormal) distribution for T. Given this assumption, γ can be estimated by maximizing the full likelihood function (Kalbfleisch and Prentice, 2002).

2.3 Pseudo-value regression for time-to-event outcomes

The modeling approaches described in 2.2 are well-established in time-to-event analysis but impose restrictive assumptions about the survival process and/or the underlying survival time distribution. Pseudo-value regression offers a less restrictive alternative for ana-

lyzing covariate effects on censored time-to-event outcomes (Andersen and Pohar Perme, 2010). In general, the outcome of interest can be expressed in terms of T as $\theta = \mathrm{E}\left[f(T)\right]$, $f : \mathbb{R}^+ \to A$, $A \subset \mathbb{R}^k$, $k \in \mathbb{N}^+$. For example, survival probabilities at t can be written as $S(t) = \theta(t) = \mathrm{E}\left[\mathcal{I}(T > t)\right]$, while the RMST at τ is given by $\mu(\tau) = \theta(\tau) = \mathrm{E}\left[\min(T, \tau)\right]$. Without censoring, θ could be easily estimated by $\frac{1}{n}\sum_{i=1}^{n} f(T_i)$, as $f(T_i)$ is observable for all patients. However, censoring prevents the full observation of $f(T_i)$ and the idea is to impute $f(T_i)$ with continuous pseudo-values for censored and uncensored patients. Given a consistent estimator $\hat{\theta}$ of θ (e.g., Kaplan-Meier), the pseudo-value for patient i out of $\mathcal{D}$ is calculated as $\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$, where $\hat{\theta}^{-i}$ denotes the leave-one-out estimator of θ based on $\mathcal{D}$ but excluding patient i (Andersen et al., 2003). Intuitively, $\hat{\theta}_i$ can be viewed as the contribution of patient i to $\hat{\theta}$ derived on $\mathcal{D}$. Pseudo-values provide a fully observed, unconditional imputation of the outcome values of interest and can thus be used as outcome in conventional modeling techniques for continuous data (e.g., linear regression) to assess the direct relationship between covariates and the outcome of interest. The direct relationship can be expressed as $\hat{\theta}_i = g^{-1}(\gamma^T X_{i\cdot}) + \varepsilon_i$ with a suitable link function g (e.g., cloglog-link for survival probabilities or log-link for the RMST) (Andersen and Pohar Perme, 2010). The most common approach for estimating the covariate effects γ is the generalized estimation equation (GEE) method (Andersen et al., 2003; Klein and Andersen, 2005; Graw et al., 2009; Andersen and Pohar Perme, 2010). It can be shown that replacing the outcome values by pseudo-values enables a consistent estimation of covariate effects under conditional random censoring, i.e., assuming that T and C are independent random variables given the covariates (Graw et al., 2009; Overgaard et al., 2017).

2.4 The role of machine learning in time-to-event analysis

Apart from the modeling approaches presented in 2.2 and 2.3, there are numerous (supervised and unsupervised) machine learning approaches for time-to-event data available, such as random survival forests or support vector machines (van Belle et al., 2007; Ishwaran et al., 2008). However, pseudo-value regression offers a completely new perspective on ma-

chine learning in time-to-event modeling, as the pseudo-values can be treated as continuous outcome values, allowing the application of machine learning approaches designed for continuous outcomes. One supervised machine learning technique discussed in this dissertation is component-wise gradient boosting, referred to as *gradient boosting* in the following (Bühlmann and Hothorn, 2007; Hofner et al., 2014). The aim of gradient boosting is to obtain an optimal prediction of the outcome of interest (e.g., pseudo-values) given a set of covariates. This is achieved by iteratively minimizing a risk function over a prediction function. In the first step of each iteration, pre-specified regression estimators (base-learners) are related separately to the negative gradient of the risk function. These base-learners can comprise, e.g., univariable linear regression functions or splines. Secondly, the base-learner with the best fit is selected and used to update the prediction function. The update is performed by adding the chosen base-learner, scaled by a shrinkage factor, to the current prediction function. These steps are repeated until a finite number of iterations is achieved (early stopping). When using simple univariable linear base-learners, the optimal prediction function is an additive combination of a selected subset of covariates, automatically implying variable selection. Another well-established method discussed in this dissertation is model-free recursive partitioning by regression trees (Breiman et al., 1984; Hothorn et al., 2006). The aim of recursive partitioning is to derive local estimates of a continuous outcome (e.g., pseudo-values) by an iterative splitting of the covariate space into mutually exclusive subspaces. At each step, the algorithm selects the optimal covariate and a corresponding binary split rule fulfilling a split criterion (Breiman et al., 1984; Hothorn et al., 2006). This iteratively forms a tree structure that separates patient groups (nodes) into smaller subgroups (daughter nodes) until a stopping criterion, like maximum tree depth, is met. The final estimates are obtained by averaging the outcome values of the patients within each terminal node (Breiman et al., 1984). Regression trees provide built-in variable selection (by selecting one covariate at each split), are able to model interactions and non-linear relationships without pre-specification in a model equation, and remain interpretable, especially with small tree depths. However, regression trees often suffer from overfitting if the tree structure becomes too complex and perfectly fits the data including

possible noise. A common approach to mitigate this issue is to use an ensemble of regression trees (e.g., 500 or 1,000 trees) grown on different random subsets of the data and to average their estimates. This technique, termed random forest regression, is a well-established supervised machine learning method (Breiman, 2001). While random forest regression constitutes a model-free algorithm allowing for variable selection and requiring no prior assumptions on the structure of the covariate effects, it belongs to the group of *black-box* machine learning models that are known to be hardly interpretable. In clinical research, model interpretability and the importance of covariates for the estimated outcome values are crucial for risk communication. Thus, numerous interpretable machine learning (IML) measures, such as feature importance or Shapley values, have been proposed (Molnar, 2022). These methods aim for the post-hoc explanation of individual (local) or population-based (global) estimates.

2.5 Datasets

Two large longitudinal clinical studies provided the data basis for this dissertation, serving as representative examples of observational studies and randomized controlled trials (RCTs).
The first dataset originates from the Peri-Interventional Outcome Study in the Elderly (POSE), a European multi-center, prospective observational trial (NCT03152734) (POSE study group, 2021). POSE included 9,497 patients aged 80 years or older undergoing any kind of surgical or non-surgical procedure under anesthesia. The primary outcome of the study was the time from intervention until death from any cause (overall survival). Patients were censored at the last date on which they were known to be alive. In total, 388 patients have been reported dead within 30 days after intervention. Data collected included patient-specific characteristics (e.g., age, sex, BMI), medical history, frailty, as well as details on the conducted intervention (surgical or non-surgical, urgency, severity).
The second dataset is drawn from the multi-center randomized phase III SUCCESS-A trial (NCT02181101) (de Gregorio et al., 2020). SUCCESS-A included 3,754 female patients with primary invasive breast cancer and a high risk of recurrence. Patients were randomized equally to one of two treatment arms (control or interventional group). The primary outcome

of the study was the time from randomization to the earliest disease progression or death from any cause (disease-free survival) within a 5-year follow-up period. Patients were censored at the last date on which they were known to be disease-free. SUCCESS-A aimed to compare the two treatment arms with respect to disease-free survival. In total, 458 patients have experienced the event of interest. Data collected included patient-specific (e.g., age, BMI) and tumor-specific covariates (e.g., stage, grade) as well as the treatment group.

2.6 Thesis outline

This dissertation aimed to develop, evaluate, and apply new modeling approaches for clinical time-to-event outcomes, presenting the development of a Cox based scoring system (Publication 1) and investigating the combination of pseudo-value regression and machine learning techniques (Publications 2 and 3) regarding performance and applicability. The development of new modeling approaches aimed to ensure flexibility in clinical applications by (i) maintaining interpretability for clinical use, (ii) reducing restrictive assumptions, and (iii) enabling the automatic selection of relevant patient information, including main covariate effects, interactions, time-varying effects, and more complex covariate structures.

2.6.1 A semi-parametric scoring system for modeling 30-day mortality

Publication 1 illustrates the development of an easy-to-use scoring system based on a Cox model to assess 30-day probability of death in elderly patients ($\geq$80 years) derived on the POSE data (POSE study group, 2021; Schenk et al., 2023). Potential risk factors were ranked and clustered by their ease of availability before intervention and their simplicity and usability in clinical practice. A number of Cox models fitted to different sets of risk factors was evaluated with respect to their predictive accuracy. The final set of risk factors, selected as a balance between predictive accuracy, ease of availability and simplicity, included severity (minor/intermediate, major), urgency (elective, non-elective), and living conditions (independent, assisted). A key component of the development process to ensure direct and simple interpretability and clinical applicability was the conversion of the estimated coefficients from

the final Cox model into a scoring system (Sullivan et al., 2004). This scoring system, named Pre-Interventional Risk Assessment in The Elderly (PIRATE), assigns risk points to risk factor categories and sums them up to a total score ranging from 0 to 5. The corresponding 30-day probability of death can be extracted from a look-up table. With just three binary risk factors, PIRATE is a simple, user-friendly tool for identifying high-mortality risk in elderly patients.

2.6.2 A semi-parametric approach for modeling survival probabilities using pseudo-values

Although PIRATE is easy to use and readily interpretable, its development relied on Cox model assumptions and manual risk factor selection without considering interactions among risk factors. In this context, the aim on an alternative modeling approach was to achieve a higher flexibility in modeling survival probabilities while maintaining interpretability of the estimates. Flexibility can be increased by reducing model assumptions and by a data-driven variable selection. To this end, Publication 2 developed pseudo-value regression trees (PRT) for modeling survival probabilities on a grid of K time points (Schenk et al., 2024). This semi-parametric extension to the GEE approach is characterized by building a multivariate regression tree with pseudo-value outcome and by successively fitting regularized additive models using gradient boosting to the data in each node of the tree. Using PRT, all available covariates are considered as potential risk factors but only the most informative ones are selected by the multivariate regression tree and the gradient boosting. Potential time-varying (treatment) effects can be modeled by including a spline base-learner for the K time points in the node-wise boosting models. The regression tree and the boosting models are able to perform variable selection and to capture interactions or more complex covariate structures. Interpretability of the estimated survival probabilities is maintained by limiting the maximum depth of the regression tree, resulting in a reasonable number of terminal nodes defining patient subgroups. Finally, gradient boosting assigns interpretable additive models to each subgroup. Publication 2 conducted a simulation study on PRT's properties and performance and applied it to the SUCCESS-A study data to demonstrate applicability and interpretability.

2.6.3 A pseudo-value random forest approach for modeling restricted mean survival times and treatment effects

In clinical research, treatment effects are often quantified using HRs from Cox models. Since the interpretation of HRs is only valid if the PH assumption holds, treatment effects have been recommended to be reported using summary measures like RMST differences, which offer simple interpretations and enable causal estimation (Uno et al., 2014; Stensrud and Hernán, 2020). Standard methods for estimating RMSTs and RMST differences, such as the integration of estimated survival functions, constitute indirect modeling approaches and often face limitations. For example, Kaplan-Meier does not account for covariates, while the validity of the estimated RMST values derived by Cox or AFT models strongly depend on the correctness of the underlying model. This leads to the need of an alternative modeling approach with the aims to (i) directly model RMST values (and their differences), (ii) have less restrictive assumptions, and (iii) enable data-driven selection of main, interaction and time-varying effects. To address the aims in (i)-(iii), Publication 3 developed a pseudo-value random forest (PVRF) approach for the estimation of RMSTs (Schenk et al., 2025). Beyond, g-computation is applied, allowing for the estimation of causal treatment effects, represented by RMST differences (keeping measured confounding variables constant) (Snowden et al., 2011). The PVRF method extends standard GEE modeling by eliminating the need for prior knowledge to include interactions or complex covariate structures. Interpretability was ensured through the calculation of permutation feature importance and local Shapley values. Publication 3 conducted comprehensive simulation studies on the performance of PVRF in estimating RMST values and treatment effects, along with an application to the SUCCESS-A data for illustration.

2.7 References

Andersen PK, Klein JP, Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. In: Biometrika 2003; 90 (1): 15–27

Andersen PK, Pohar Perme M. Pseudo-observations in survival analysis. In: Statistical Methods in Medical Research 2010; 19 (1): 71–99

Breiman L. Random forests. In: Machine Learning 2001; 45 (1): 5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton: ChapmanHall/CRC, 1984

Bühlmann P, Hothorn T. Boosting algorithms: Regularization, prediction and model fitting. In: Statistical Science 2007; 22 (4): 477–505

Cox DR. Regression models and life-tables. In: Journal of the Royal Statistical Society. Series B (Methodological) 1972; 34 (2): 187–220

de Gregorio A, Häberle L, Fasching PA, Müller V, Schrader I, Lorenz R, Forstbauer H, Friedl TWP, Bauer E, de Gregorio N, Deniz M, Fink V, Bekes I, Andergassen U, Schneeweiss A, Tesch H, Mahner S, Brucker SY, Blohmer JU, Fehm TN, Heinrich G, Lato K, Beckmann MW, Rack B, Janni W. Gemcitabine as adjuvant chemotherapy in patients with high-risk early breast cancer – results from the randomized phase III SUCCESS-A trial. In: Breast Cancer Research 2020; 22 (1): 111

Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. In: Lifetime Data Analysis 2009; 15 (2): 241–255

Hofner B, Mayr A, Robinzonov N, Schmid M. Model-based boosting in R: A hands-on tutorial using the R package mboost. In: Computational Statistics 2014; 29 (1): 3–35

Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. In: Journal of Computational and Graphical Statistics 2006; 15 (3): 651–674

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. In: The Annals of Applied Statistics 2008; 2 (3): 841–860

Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. Hoboken: Wiley, 2002

Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. In: Biometrics 2005; 61 (1): 223–229

Klein JP, Moeschberger ML. Survival analysis: Techniques for censored and truncated data. New York: Springer, 2003

Molnar C. Interpretable machine learning (second edition). A guide for making black box models explainable. Independently published, 2022

Overgaard M, Parner ET, Pedersen J. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. In: The Annals of Statistics 2017; 45 (5): 1988–2015

POSE study group. Peri-interventional outcome study in the elderly in Europe: A 30-day prospective cohort study. In: European Journal of Anaesthesiology 2021; 39 (3): 198–209

Schenk A, Basten V, Schmid M. Modeling the restricted mean survival time using pseudo-value random forests. In: Statistics in Medicine 2025; 44 (5): e70031

Schenk A, Berger M, Schmid M. Pseudo-value regression trees. In: Lifetime Data Analysis 2024; 30 (2): 439–471

Schenk A, Kowark A, Berger M, Rossaint R, Schmid M, Coburn M, the POSE study group. Pre-Interventional Risk Assessment in The Elderly (PIRATE): Development of a scoring system to predict 30-day mortality using data of the Peri-Interventional Outcome Study in the Elderly. In: PLoS One 2023; 18 (12): e0294431

Snowden JM, Rose S, Mortimer KM. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. In: American Journal of Epidemiology 2011; 173 (7): 731–738

Stensrud MJ, Hernán MA. Why test for proportional hazards? In: Journal of the American Medical Association 2020; 323 (14): 1401–1402

Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. In: Statistics in Medicine 2004; 23 (10): 1631–1660

Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, Schrag D, Takeuchi M, Uyama Y, Zhao L, Skali H, Scott S, Jacobus S, Hughes M, Packer M, Wei LJ. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. In: Journal of Clinical Oncology 2014; 32 (22): 2380–2385

van Belle V, Pelckmans K, Suykens J, Huffel SV. Support vector machines for survival analysis. In: Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007) 2007: 1–8

3 Publications

3.1 Publication 1: Pre-Interventional Risk Assessment in The Elderly (PIRATE): Development of a scoring system to predict 30-day mortality using data of the Peri-Interventional Outcome Study in the Elderly

Schenk A*, Kowark A*, Berger M, Rossaint R, Schmid M∓, Coburn M∓, the POSE study group. Pre-Interventional Risk Assessment in The Elderly (PIRATE): Development of a scoring system to predict 30-day mortality using data of the Peri-Interventional Outcome Study in the Elderly. In: PLoS ONE 2023; 18 (12): e0294431

*These authors are joint first authors.

∓These authors are joint last authors.

Link to publication and supplementary information:
https://doi.org/10.1371/journal.pone.0294431

The implementation of the PIRATE web application can be found at:
https://schenkalina.shinyapps.io/pirate_app

PLOS ONE

RESEARCH ARTICLE

Pre-Interventional Risk Assessment in The Elderly (PIRATE): Development of a scoring system to predict 30-day mortality using data of the Peri-Interventional Outcome Study in the Elderly

Alina Schenk[1☯*], **Ana Kowark**[2,3☯], **Moritz Berger**[1], **Rolf Rossaint**[3], **Matthias Schmid**[1‡], **Mark Coburn**[2‡], **the POSE Study group**[¶]

1 Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Bonn, Germany, **2** Department of Anaesthesiology and Intensive Care Medicine, University Hospital Bonn, Bonn, Germany, **3** Department of Anaesthesiology, Medical Faculty University Hospital RWTH Aachen, Aachen, Germany

☯ These authors contributed equally to this work.
‡ MS and MC also contributed equally to this work.
¶ POSE Study Group: Collaborators are listed in S1 File.
* schenk@imbie.uni-bonn.de

Citation: Schenk A, Kowark A, Berger M, Rossaint R, Schmid M, Coburn M, et al. (2023) Pre-Interventional Risk Assessment in The Elderly (PIRATE): Development of a scoring system to predict 30-day mortality using data of the Peri-Interventional Outcome Study in the Elderly. PLoS ONE 18(12): e0294431. https://doi.org/10.1371/journal.pone.0294431

Editor: Pasquale Abete, Universita degli Studi di Napoli Federico II, ITALY

Received: May 10, 2023

Accepted: November 1, 2023

Published: December 21, 2023

Data Availability Statement: The data is not publicly available due to privacy and ethical restrictions. As the data are personally identifiable, the General Data Protection Regulation of the European Union prevents us from making them publically available. The authors must be the points of contact for fielding data access requests because they include Prof. Coburn who is the principle investigator of the POSE study. Also, our author list contains the whole POSE study group,

Abstract

Risk assessment before interventions in elderly patients becomes more and more vital due to an increasing number of elderly patients requiring surgery. Existing risk scores are often not tailored to marginalized groups such as patients aged 80 years or older. We aimed to develop an easy-to-use and readily applicable risk assessment tool that implements pre-interventional predictors of 30-day mortality in elderly patients (≥80 years) undergoing interventions under anesthesia. Using Cox regression analysis, we compared different sets of predictors by taking into account their ease of availability and by evaluating predictive accuracy. Coefficient estimates were utilized to set up a scoring system that was internally validated. Model building and evaluation were based on data from the Peri-Interventional Outcome Study in the Elderly (POSE), which was conducted as a European multicenter, observational prospective cohort study. Our risk assessment tool, named PIRATE, contains three predictors assessable at admission *(urgency*, *severity* and *living conditions*). Discriminatory power, as measured by the concordance index, was 0.75. The estimated prediction error, as measured by the Brier score, was 0.036 (covariate-free reference model: 0.043). PIRATE is an easy-to-use risk assessment tool that helps stratifying elderly patients undergoing interventions with anesthesia at increased risk of mortality. PIRATE is readily available and applies to a wide variety of settings. In particular, it covers patients needing elective or emergency surgery and undergoing in-hospital or day-case surgery. Also, it applies to all types of interventions, from minor to major. It may serve as a basis for multidisciplinary and informed shared decision-making.

meaning that there are no other possible points of contact for fielding data access requests. Also note that the POSE Study investigators have established a procedure to gain access to the original data on reasonable request. For this, researchers need to submit a proposal for a secondary analysis to the steering committee of the study, please see the guideline and the list of approved secondary analyses at https://pose-trial.org/secondary-analyses. The authors can be contacted via email at the following addresses: Prof. Dr. Mark Coburn (Mark.Coburn@ukbonn.de) PD Dr. Ana Kowark (Ana.Kowark@ukbonn.de) Prof. Dr. Rolf Rossaint (RRossaint@ukaachen.de) Patient data were collected on paper-based case report forms (SDC 3, http://links.lww.com/EJA/A657) and entered into an electronic database (OpenClinica, Boston, Massachusetts, USA) pseudonymised. In addition to automatic database completion, consistency and plausibility checks, and manual multilevel data validation were performed. Discrepancies were clarified with local investigators.

Funding: This study was supported by the European Society of Anaesthesiology and Intensive Care (ESAIC) as an ESAIC Research Group. This constituted the advertising of the POSE study and POSE meetings on the annual Euroanaesthesia congress, the indirect use of the ESAIC members contact lists, and the financial support for holding of three steering committee meetings at the ESAIC Secretariat in Brussels. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: AK, MB, MC, MS and RR report grants from ESAIC, during the conduct of the study. AS: No competing interests declared. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Introduction

According to the World Health Organization (WHO) World *Report on Aging and Health* significant impairment in the elderly population is reported. The number of elderly people in Europe will double by 2050 and thus the number of elderly patients requiring surgery [1]. In consequence, there is increasing need for pre-interventional risk assessment and outcome prediction focusing on elderly patients.

The key challenge of any pre-interventional risk assessment in the elderly is to identify and stratify patients at increased risk of mortality and morbidity, accounting for characteristics that are of particular importance to elderly people, like functional status, level of independence and frailty. Pre-interventional risk assessments may thus contribute to informed decision making, helping both, the patients and possible authorized representatives of the elderly patients, to better evaluate the trade-off between the medical necessity of a (non-) surgical intervention and patient specific outcomes [2]. Moreover, they may be employed to guide clinical planning and decision making, in particular by customising (non-)surgical interventions. In this respect, the updated *Pre-Operative Evaluation of Adults Undergoing Elective Noncardiac Surgery* guideline of the European Society of Anaesthesiology and Intensive Care recommends in its section on geriatric patients to assess *pre*-interventional functional status, level of independence, comorbidities and frailty [3]. Further, the guideline on *Perioperative Care in Adults* published by the National Institute for Health and Care Excellence (NICE) in 2020 recommends to use validated risk stratification tools to supplement clinical assessment when planning surgery [4].

Despite these recommendations, there are thus far no risk assessment tools specifically developed on elderly patients (≥80 years). To the best of our knowledge, no performance evaluations of existing risk assessment scores in the subgroup of elderly patients exist. Commonly used scores such as e.g. the *Preoperative Score to Predict Postoperative Mortality* (POSPOM), the *Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity* (POSSUM), the *Portsmouth-POSSUM* (P-POSSUM), the *Surgical Outcome Risk Tool* (SORT), the *National Surgical Quality Improvement Program* (NSQIP) *Universal Surgical Risk Calculator*, the *Estimation of Physiologic Ability and Surgical Stress* (E-Pass), and the *Surgical Risk Scale* (SRS) have all been developed on data referring to a wider age range and employing a number of risk factors that are, to some extent, not assessable before intervention [5–13].

Therefore, the aim of this analysis was to develop a pre-interventional risk calculation tool that is tailored to the assessment of post-interventional mortality in elderly patients (≥ 80 years). Using prospectively collected data from the Peri-interventional Outcome Study in the Elderly (POSE), we derived and internally validated a user-friendly scoring system, named **P**re-**I**nterventional **R**isk **A**ssessment in **T**he **E**lderly (PIRATE) [14]. As described in detail in the Results section below, PIRATE resulted from a stepwise predictor selection procedure taking into account

i. simplicity and usability of the scoring system in daily clinical practice (avoiding complex and time-consuming calculations),

ii. ease of availability of predictors *before* intervention (in particular, by using unambiguously defined risk categories), and

iii. prediction accuracy.

Reporting of the PIRATE tool will be based on the *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis* (TRIPOD) statement [15].

Methods

Study population

The step-by-step development of our scoring system was based on the POSE database (exported on 17th of February 2020). POSE was conducted as a European multicenter, observational prospective cohort study to investigate mortality rates and other outcomes in the elderly population. Patients were eligible, if aged 80 years or older and undergoing surgical or non-surgical interventions under anesthesia. The study period lasted from October 2017 to December 2018. Each center recruited patients for 30 consecutive days within the study period. Interventions were classified as either surgical or non-surgical, elective or non-elective, and inpatient or outpatient. In total, POSE enrolled 9,862 patients from 177 study centers in 20 different countries, of which 9,497 patients were eligible for analysis. The reasons for exclusion of 365 patients comprised death before intervention (n = 20), intervention postponed/ cancelled (n = 301), missing patient records (n = 22), and not collected data (n = 22). Of 9,497 patients, 388 experienced the event of interest (i.e., death within 30 days after intervention) and 9,109 did not experience the event of interest ("controls"), resulting in a post-interventional mortality rate of 4.2% (95% CI 3.8%-4.7%) [14]. POSE was approved by the University Hospital RWTH Aachen, Germany (EK 162/17). Mandatory research ethics board (REB) approval or a waiver was granted at each center. Written informed consent was obtained from all subjects participating in the trial. POSE was registered prior to patient enrollment at clinicaltrials.gov (NCT03152734, Chief coordinating investigator: Mark Coburn, Date of registration: May 15, 2017). The development of PIRATE was approved by the POSE Steering Committee as a secondary analysis (https://pose-trial.org/secondary-analyses). A data transfer agreement between the University Hospital RWTH Aachen and the Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn was established. AS, MB and MS had no access to information that could identify individual patients during or after data collection. It is not precluded that AK, RR and MC could have identified patients from their respective study site in the course of their work as treating physicians.

Outcome definition

The outcome of interest was the time after intervention until death from any cause. Patients potentially having an event after 30 days were censored. The survival status of patients discharged before day 30 was enquired using telephone interviews [14].

Definition and choice of predictors

The aim of this secondary analysis was the pre-interventional risk assessment of post-interventional mortality of elderly patients ($\geq$ 80 years), i.e., the prediction of 30-day mortality after intervention.

The basis of the stepwise development of the PIRATE tool was the complete POSE cohort (9,497 patients). We considered 15 potential predictors (seven binary, six categorical and two continuous predictors). Of these, ten predictors (four binary, four categorical and two continuous predictors) fulfilled the requirement of being assessable *before* intervention (see the POSE statistical analysis plan [14] for details on all available predictors and their categories, see Table 1 for details on included predictors and their categories). These ten predictors, including *age [years]*, *bmi [kg/m^2]*, *sex*, *severity (minor*, *intermediate*, *major)* and *urgency (elective*, *urgent*, *emergent)* of intervention, *type of intervention*, *multimorbidity* and *referring facility* of the patients as well as *frailty* and a test for patients' *mobility* (timed up and go [TUG] test) were considered in the development process of the scoring system. In POSE, a patient was

Table 1. Patient characteristics of the POSE [14] cohort used for the development of PIRATE. Values are mean (SD) or number (proportion).

Variable	All n = 9497 n (%)	Cases n = 388 n (%)	Controls n = 9109 n (%)
Age [years] (mean, sd)	84.32 (3.8)	85.78 (4.8)	84.26 (3.8)
BMI [kg/m^2] (mean, sd)	25.94 (4.33)	25.15 (4.53)	25.98 (4.32)
Missing	148 (1.6%)	11 (2.8%)	137 (1.5%)
Sex			
male	4485 (47.2%)	192 (49.5%)	4293 (47.1%)
female	5012 (52.8%)	196 (50.5%)	4816 (52.9%)
Severity			
minor	1947 (20.5%)	38 (9.8%)	1909 (21.0%)
intermediate	3612 (38.0%)	107(27.6%)	3505 (38.5%)
major	3938 (41.5%)	243 (62.6%)	3695 (40.6%)
Urgency			
elective	7176 (75.6%)	146 (37.6%)	7030 (77.2%)
emergent	479 (5.0%)	87 (22.4%)	392 (4.3%)
urgent	1842 (19.4%	155 (39.9%)	1687 (18.5%)
Frailty			
frail	1336 (14.1%)	180 (46.4%)	1156 (12.7%)
not frail	8161 (85.9%)	208 (53.6%)	7953 (87.3%)
Type of intervention			
abdominal	1149 (12.1%)	89 (22.9%)	1060 (11.6%)
cardiovascular and thoracic	896 (9.4%)	60 (15.5%)	836 (9.2%)
ENT; ophthalmologic	1594 (16.8%)	8 (2.1%)	1586 (17.4%)
gynaecologic and urologic	1437 (15.1%)	21 (5.4%)	1416 (15.5%)
interventional	1026 (10.8%)	29 (7.5%)	997 (10.9%)
neurosurgery	196 (2.1%)	22 (5.7%)	174 (1.9%)
orthopaedic, trauma and plastic	2860 (30.1%)	142 (36.6%)	2718 (29.8%)
transplant or other surgery	339 (3.6%)	17 (4.4%)	322 (3.5%)
Living conditions (Facility)			
Home	8220 (86.6%)	254 (65.5%)	7966 (87.5%)
Other hospital	184 (1.9%)	31 (8.0%)	153 (1.7%)
Rehabilitation	60 (0.6%)	2 (0.5%)	58 (0.6%)
Nursing home	670 (7.1%)	65 (16.8%)	605 (6.6%)
other	360 (3.8%)	36 (9.3%)	324 (3.6%)
missing	3 (0.03%)	0 (0%)	3 (0.03%)
Multimorbidity			
yes	7334 (77.2%)	359 (92.5%)	6975 (76.6%)
no	2163 (22.8%)	29 (7.5%)	2134 (23.4%)
Mobility (TUG test)			
limited	6461 (68.0%)	316 (81.4%)	6145 (67.5%)
normal	1910 (20.1%)	16 (4.1%)	1894 (20.8%)
missing	1126 (11.9%)	56 (14.4%)	1070 (11.7%)

Abbreviation

BMI = Body Mass Index

ENT = Ear, Nose and Throat

POSE = Peri-Interventional Outcome Study in the Elderly

PIRATE = Pre-Interventional Risk Assessment in The Elderly

SD = Standard deviation

TUG = Timed up and go

https://doi.org/10.1371/journal.pone.0294431.t001

classified as *frail* if at least 4 of 6 criteria (mini-cog score of $\leq$ 3 points, albumin level of $\leq$ 3.3 g/d, more than 1 fall in the last 6 months, haematocrit level of $<$ 35%, preoperative functional status is partially dependent or totally dependent, $\geq$3 comorbidities) were fulfilled [14]. Following the definition by the WHO, *multimorbidity* was defined as the presence of at least two chronic conditions [1, 14]. The TUG test was performed to assess mobility of patients. The patients were asked to stand up from a chair, to walk three metres, to turn around and to walk back and sit down again. The test result was evaluated as normal mobility if the patient was able to perform the TUG test in 12 seconds or less. If the patient was not able to perform the TUG test or took more than 12 seconds to perform the test, the test result was evaluated as limited mobility.

Development of the scoring system

Development of the scoring system was based on a stepwise procedure that accounted for the trade-off between prediction accuracy and simplicity, focussing on the predictors' ease of availability in daily clinical routine. In each step of the development process, we fitted a Cox proportional hazards regression model containing different subsets or combinations of the ten initially available predictors (described above). In order to internally validate the developed scoring system at each step, we repeatedly divided the entire study cohort on the center level into a derivation cohort and a validation cohort (100 replications). Specifically, each derivation cohort provided a training data set comprising a set of randomly chosen study centers that included approximately two thirds of the patients in POSE. The patients of the remaining study centers were allocated to the respective validation cohort providing the test data set. Prediction accuracy was measured using the concordance index (C-index) averaged across the 100 validation cohorts [16]. Variable importance was measured by the loss in C-index when permuting the respective predictor. To assess calibration, we generated calibration plots that compared predicted 30-day survival probabilities to their respective Kaplan-Meier estimates. Prediction error of the final model was measured using the Brier score [17]. The various model building steps will be described in detail in the Results section. After model building, we developed a scoring system based on the final Cox proportional hazards regression model, assigning risk points to each category of the included risk factors (predictors) [18]. With this system (entitled **P**re-**I**nterventional **R**isk **A**ssessment in **T**he **E**lderly [PIRATE]), users can simply add all risk points and extract the respective estimated 30-day mortality from a look-up table.

Handling of missing data

Missing data were imputed using multiple imputation (fully conditional specification with all ten initially available predictors [19, 20]). We generated 12 imputed data sets, following the POSE trial statistical analysis plan [14].

A sensitivity analysis composed of the application of the development process on each of the 12 imputed data sets revealed only marginal differences in the results (on the third decimal place of C-index values) that are less relevant for the final conclusions. Thus, the development is illustrated for one single imputed data set in the following. The majority of missing values was present in *mobility*, which is, as explained in the Results section, not considered in the final scoring system. Thus, changes across the imputed datasets for *mobility* were negligible.

All calculations were performed using the R language and environment for statistical computing (version 4.1.0).

Results

Patient characteristics of the 9,497 POSE patients (without imputation of missing values) are presented in Table 1. In the following, we will give a detailed description of each model

building step, weighing simplicity, usability, availability of predictors and prediction accuracy. The C-index value presented in each step represents the mean value averaged across 100 replications.

Step 0: Model with all available predictors

The model including all initially available predictors (*age*, *bmi*, *sex*, *facility*, *type of intervention*, *severity*, *urgency*, *multimorbidity*, *timed up go*, *frailty*) reached a mean C-index of 0.818.

Step 1: Grouping of predictors based on availability at the time of admission

Based on expert discussions with members of the POSE study team, we grouped the predictors according to the following criteria:

- Very easy to gather: *age*, *sex*, *facility*,
- Easy to gather: *bmi*, *urgency*, *type of intervention*,
- Hard to gather: *severity*, *multimorbidity*,
- Very hard to gather: *frailty (as assessed in POSE)*, *timed up go*.

Based on this grouping, we considered the following set of models:

- Model 0: *Null model (without any predictors)*,
- Model 1: *age*, *sex*, *facility*,
- Model 2: *age*, *sex*, *facility*, *bmi*, *urgency*, *type of intervention*,
- Model 3: *age*, *sex*, *facility*, *bmi*, *urgency*, *type of intervention*, *severity*, *multimorbidity*,
- Model 4 (from Step 0): *age*, *sex*, *facility*, *bmi*, *urgency*, *type of intervention*, *severity*, *multimorbidity*, *frailty*, *timed up go*.

Fig 1(A) presents the mean C-index values that were obtained from applying the above models to the 100 different training data sets. It is seen that there was an upwards trend in prediction accuracy as the number of predictors increased. On the other hand, the differences in C-index values between models 2, 3 and 4 were considerably smaller than the respective difference between models 1 and 2. Based on this result and keeping the ease of availability of the predictors in mind, model 2 (including *age*, *sex*, *facility*, *bmi*, *urgency* & *type of intervention*, *and excluding four predictors from Step 0)* seemed to be a reasonable compromise between prediction accuracy and usability. The mean C-index of model 2 was 0.785.

Step 2: Statistical importance of the predictors (permutation importance)

In the next step, we analyzed the individual contributions of the ten predictors to the prediction accuracy of the models. To this purpose, we ranked the predictors according to their (statistical) permutation importance. This was done by randomly permuting the training data of the ten available predictors, considering one predictor at a time. Full models with all ten predictors were then fitted to the training data (one model per permuted predictor, each time leaving the training data of the other eight predictors unchanged) and the C-indices were calculated on the (non-permuted) test data. For each predictor, we calculated its permutation importance, which was defined as the difference between the C-index values obtained from the full model with original data and the model(s) with permuted data. The ranking of the

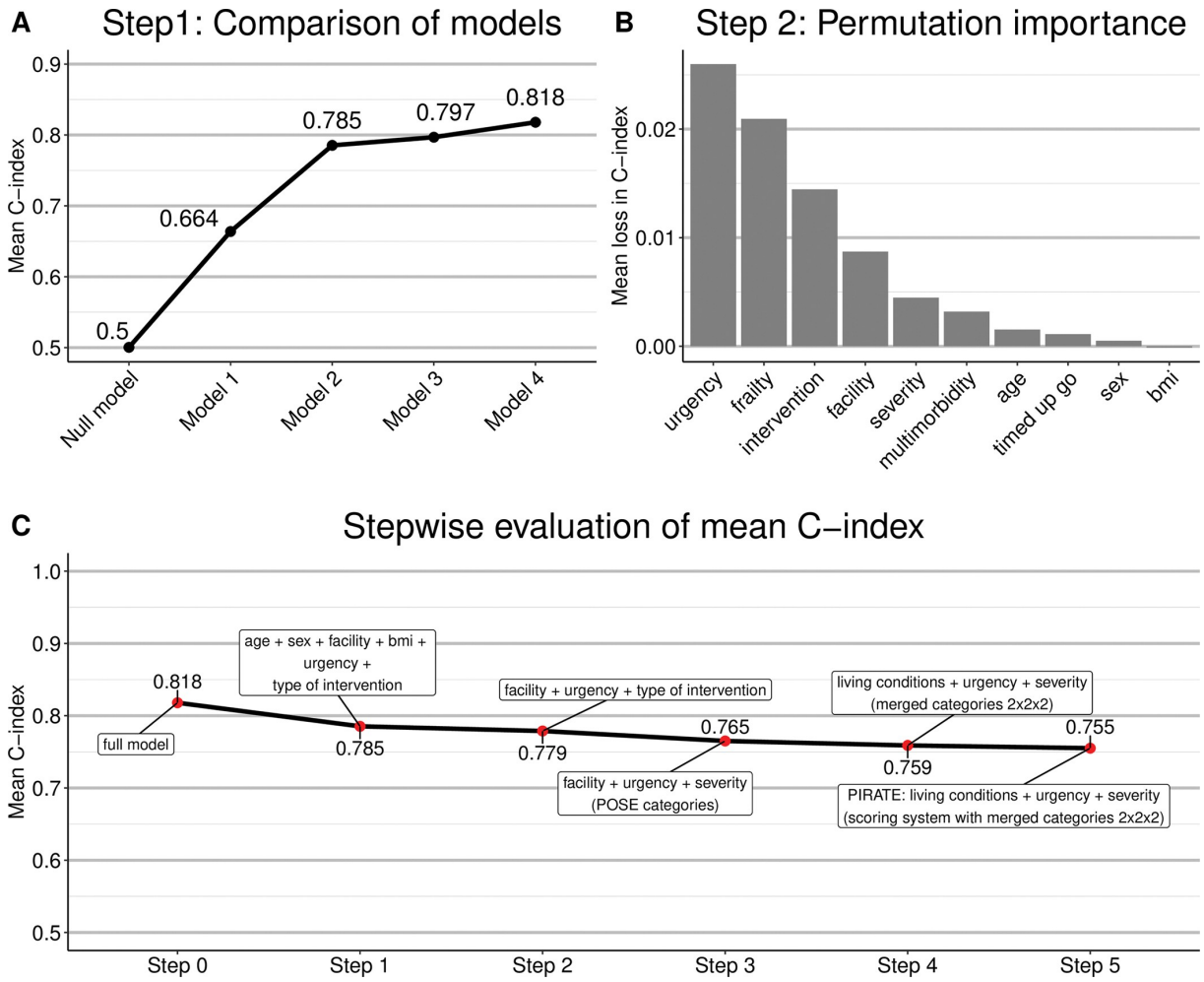


Fig 1. Development of PIRATE. (A) Mean C-index values that were obtained from applying the models in Step 1 to the 100 different training data sets. There was an upwards trend in prediction accuracy as the number of predictors increased in each model. (B) Permutation importance of the ten initially available predictors in Step 2. Permutation importance was defined as the difference between the C-index values obtained from the full model (from Step 0) with original data and the model(s) with permuted data. (C) Stepwise evaluation of the mean C-index from Step 0 (full model) to Step 5 (PIRATE).

https://doi.org/10.1371/journal.pone.0294431.g001

importance values of the ten predictors is presented in Fig 1(B). Statistically, the most important predictor was *urgency* followed by *frailty*, *type of intervention* and *facility*. Including these four predictors in the model, we obtained a mean C-index of 0.807 (on 100 different test data sets containing one-third of the complete cohort). *Urgency* as well as *type of intervention* and *facility* matched the set of predictors contained in our favored model in Step 1. *Frailty*, however, was not considered for inclusion in this model, as it is rather hard to assess in clinical routine when using the definition of frailty in POSE (comprising six individual items, see [14]). Further, the inclusion of *age*, *bmi* and *sex* (and *timed up go*) did not result in a gain in the mean C-index compared to the model excluding those predictors. Additionally, the inclusion of *frailty* in Step 1 (Model 3 vs. Model 4) did not increase the C-index appreciably (0.797 compared to 0.818). Thus, excluding *frailty*, *age*, *bmi* and *sex*, we fitted a model solely containing *urgency*, *type of intervention* and *facility*. This model resulted in a mean C-index of 0.779

(compared to 0.785 for the model from Step 1 containing *age*, *sex*, *facility*, *bmi*, *urgency* and *type of intervention*; see Fig 1(C)).

Step 3: Replacement of type of intervention by severity

The model from Step 2 containing *urgency*, *type of intervention* and *facility* consists of three categorical predictors with, in total, 3 x 8 x 5 = 120 combinations of categories. Regarding the simplicity and usability of the score in clinical routine, differentiating eight categories for *type of intervention* seems impractical given that the score should be calculated as quickly as possible. On the other hand, the *severity* of an intervention (coded by three categories) is strongly associated with the *type of intervention*: Once the *type of the intervention* is known, the *severity* of an intervention can simply be evaluated (Chi-Squared test, $p<10^{-16}$). The replacement of *type of intervention* by *severity* in our model lead to a slightly lower mean C-index (0.765 compared to 0.779 from step 2 [including *urgency*, *type of intervention* and *facility*], Fig 1(C)) but tremendously facilitates the application of the score.

Step 4: Merging categories

The model resulting from Step 3 containing *urgency*, *severity* and *facility* included three categorical predictors with 3 x 3 x 5 = 45 combinations of categories. In order to further simplify calculation of the score, we reduced the number of categories of each predictor to two. More specifically, we collapsed two of the three categories of *urgency* (*elective*, *urgent* and *emergency*), obtaining a binary predictor that indicated whether an intervention was planned (*elective*) or not. Analogously, rather than distinguishing between *minor*, *intermediate* and *major severity*, we generated a binary predictor indicating whether the intervention to be performed was *major* or not. *Referring facility* was transformed into the two categories *independently living* or *(medically) assisted*. Here, the categories *rehabilitation*, *other hospital* and *nursing home* were summarized to *(medically) assisted* while *home* was considered as *independently living*, since the respective field in the case report form was originally *home/independent*. Further, regarding the category *other* in *referring facility*, free text answers were manually screened and assigned to one of the two aforementioned categories. More specifically, free text answers (indicated as *other* in Table 1) referring to *religious community*, *monastery*, *hostel* and *homeless* were allocated to *independently living* while all other text answers indicated help from a family member or a trained nurse and were therefore allocated to *(medically) assisted*. In the remainder, we will use the term *living conditions* consisting of the two aforementioned aggregated categories instead of *facility* which refers to the covariate with the initial five categories as in POSE. The simplified score containing the three binary predictors reached a mean C-index of 0.759 (Fig 1(C)).

Step 5: Transferring the score to a scoring system

To facilitate the application and interpretation of the score in the clinical practice, we transferred the model derived in Step 4 to a scoring system that is based on the assignment of *risk points*. Following the approach described in Sullivan et al. [18], we fitted Cox regression models to the data of the 100 derivation cohorts, incorporating the three binary predictors derived in Step 4. Based on the estimated coefficients obtained from the Cox regression models, the scoring system was set up in each of the derivation cohorts, and the respective estimated 30-day probabilities of death were calculated for the patients in the validation cohorts. Reference categories for each risk factor were chosen according to the strength of risk association, assigning zero points to the groups with the lowest risk and higher numbers of points to groups with higher risk (for details, see [18]). Thus, an increasing score is related to an

increased estimated 30-day probability of death. We termed the resulting system **P**re-**I**nterventional **R**isk **A**ssessment in **T**he **E**lderly (PIRATE). Note that the methodology proposed by Sullivan et al. involves a constant *B* reflecting the number of regression units corresponding to one point [18]. For PIRATE, we set *B* equal to the regression coefficient of *severity*, as estimated from the Cox regression model. Thus, the constant reflects the increase in 30-day mortality risk associated with a major intervention [18].

Compared to the Cox regression model in Step 4, the C-index of the scoring system decreased only slightly (from 0.759 to 0.755, see Fig 1(C) and below).

PIRATE: The final risk assessment tool

The final scoring system (complete cohort with 9,497 observations) is presented in Table 2. Using the data in Table 2, the individual risk of a patient can be calculated by summing up all points belonging to the values of the patient's risk factors. The respective estimated 30-day probability of death can be extracted from the "look-up" Table 3. Total score values in the full POSE cohort ranged between 0 and 5 (see S1 Table for example calculations of the risk score).

The scoring system showed good discrimination ability with the mean estimated C-index across all validation cohorts of 0.755 (min = 0.708, max = 0.797). Prediction error was also small, with mean estimated Brier score of 0.036 (min = 0.026, max = 0.046) across all validation cohorts (compared to 0.043 obtained from a reference model not containing any predictor information). Fig 2 shows exemplary calibration plots for six validation cohorts.

Fig 3(A) presents the distribution of the score values in the complete cohort (9,497 observations). The grey bars represent the relative frequencies of the score values in the full POSE cohort, the black line represents the respective estimated 30-day probabilities of death, and the blue line refers to the Kaplan-Meier estimates of 30-day mortality in patients having the respective score value. As seen from the figure, the scores in the POSE cohort mainly ranged between 0 and 3, with only few observations having a score higher than 3. Fig 3(A) shows that PIRATE-based probability estimates (black line) and the Kaplan-Meier estimates of 30-day mortality (blue line) matched well for almost all score values in the full POSE cohort.

Stratified Kaplan-Meier estimates in subgroups defined by the 25%, 50% and 75% percentiles of the score values are shown in Fig 3(B). Together with Fig 3(A), the non-overlapping survival curves in Fig 3(B) reflect the score's ability to discriminate between high-risk and low-risk patients.

Table 2. PIRATE scoring system, as derived from the coefficient estimates of the Cox regression model in Step 4. The constant B is given as $B = 0.5986$ [18].

Risk factor	Coefficient estimate	Risk points
Severity		
minor/intermediate	-	+0
major	0.5986	+1
Urgency		
elective	-	+0
non-elective	1.3912	+2
Living conditions		
independent	-	+0
(medically) assisted	0.8985	+2

Abbreviation

PIRATE = Pre-Interventional Risk Assessment in The Elderly

https://doi.org/10.1371/journal.pone.0294431.t002

Table 3. Look-up table for the predicted 30-day probability of death after intervention.

Total points	Estimated probability of death [%] (within 30 days after intervention)
0	1.29%
1	2.33%
2	4.20%
3	7.51%
4	13.24%
5	22.78%

https://doi.org/10.1371/journal.pone.0294431.t003

Discussion

Using the POSE cohort, we were able to derive a new mortality risk assessment tool (PIRATE) that is based on three fast and simply to gather pre-interventional predictors. Starting with a multivariable Cox model containing ten predictors, our modeling approach balanced between

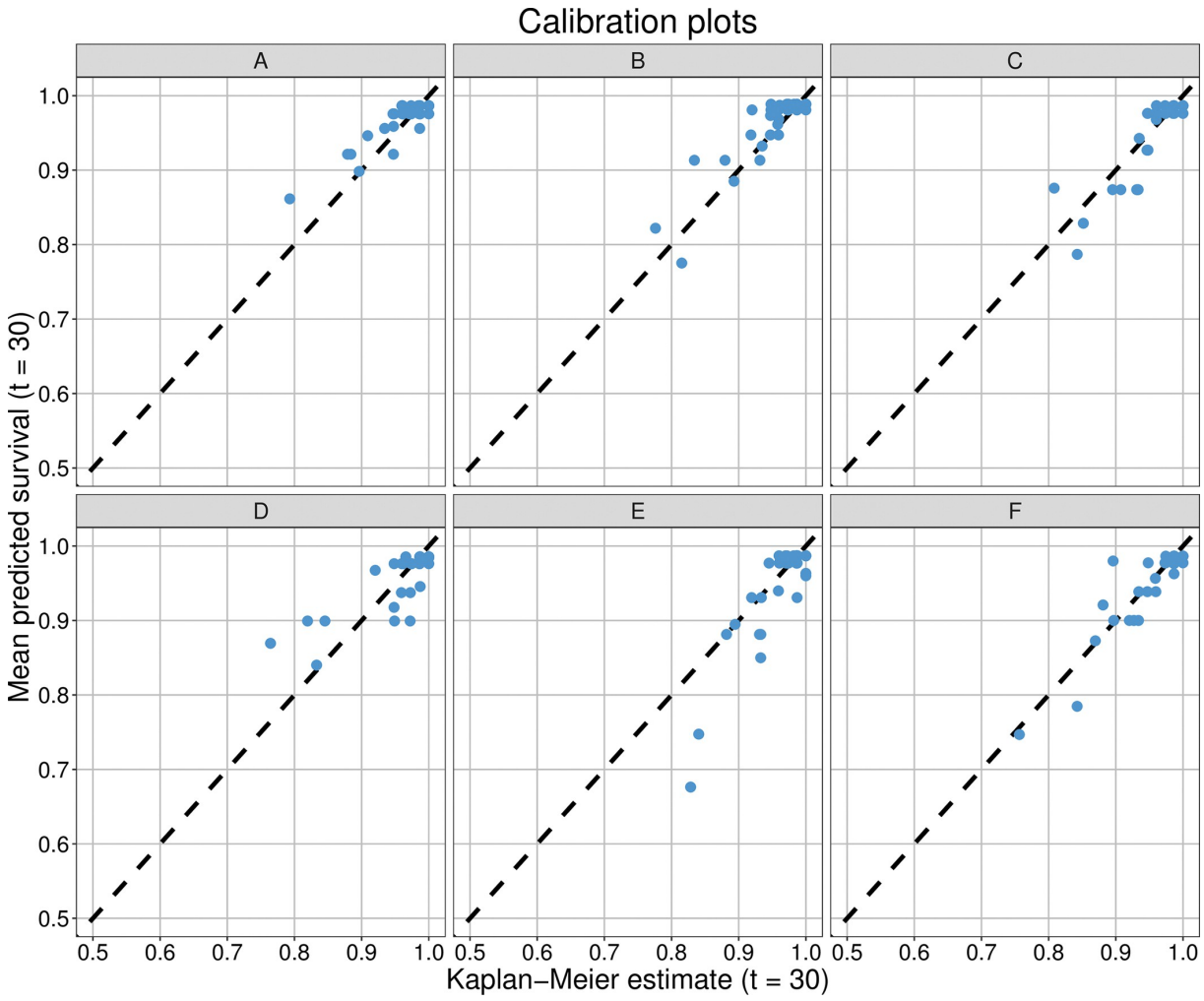


Fig 2. Calibration plots. Calibration plots for six exemplary validation cohorts. The plots depict the predicted probabilities based on the scoring system versus the Kaplan-Meier estimates in subgroups.

https://doi.org/10.1371/journal.pone.0294431.g002

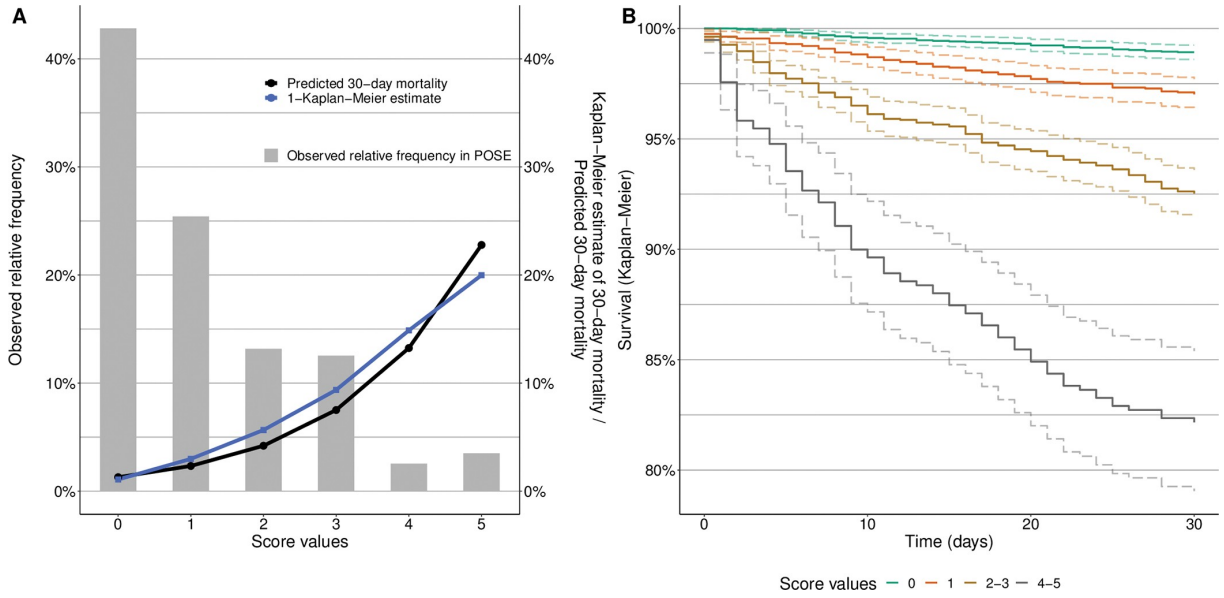


Fig 3. Evaluation of the PIRATE tool. (A) Distribution of the risk score values. The grey bars represent the relative frequencies of the risk score values in the full POSE cohort, the black line represents the respective estimated probabilities obtained from PIRATE, and the blue line refers to the death probabilities (one minus Kaplan-Meier estimates) for patients having the respective score. (B) Stratified Kaplan-Meier estimates in subgroups. Groups were defined by the 25%, 50% and 75% percentiles of the risk score values in POSE. The non-overlapping survival curves reflect the score's ability to distinguish among high risk and low risk patients.

https://doi.org/10.1371/journal.pone.0294431.g003

(i) simplicity and usability of the score in clinical routine, (ii) availability of predictors (and the speed in gathering those) and (iii) prognostic accuracy of the score. The resulting PIRATE system demonstrates that an easy-to-use score solely based on readily available pre-interventional patient characteristics can be a powerful tool in predicting the post-interventional 30-day probability of death in elderly patients. In our internal validation analysis of the POSE data set, the three-predictor PIRATE system was able to identify patients with an increased mortality risk and discriminated well between high- and low-risk patients, thereby offering the possibility to improve both risk communication (based on easily understandable patient characteristics) and post-interventional treatment optimization. In particular, PIRATE highlights the markedly different prognoses for urgent (non-elective) and scheduled (elective) interventions. This is seen, for example, by considering the group of patients living medically assisted and undergoing a severe intervention (patients 3 and 4 in S1 Table): In this group, the predicted 30-day mortality risk is almost three times higher (22.78%) if the intervention is non-elective (patient 4) than if the intervention is elective (30-day mortality risk 7.51%, patient 3).

Comparison to existing scores

Previously developed scores (e.g. POSSUM, P-POSSUM, POSPOM) used a logistic regression model with a binary outcome (dead vs. alive) for score development not accounting for censoring. In contrast to these scores, PIRATE is based on a Cox regression model that accounts for the characteristics of the survival and censoring processes during the post-interventional 30-day period [5–7]. Further, compared to other scores, we solely included readily available *pre*-interventional predictors, focussing on a quick and easy risk assessment before intervention [6, 7]. Similar to POSPOM, we derived a user-friendly scoring system that is applicable in

daily clinical routine [5]. Of note, PIRATE was derived using data exclusively collected in the elderly target population.

As part of our project, we evaluated the predictive performance of the POSPOM scoring system in the POSE study cohort, mapping the categories in POSE to the risk factors used in POSPOM [5]. While POSPOM showed excellent performance and calibration on its original validation cohort extracted from the French National Hospital Discharge Database (C-index: 0.929), it reached a C-index of 0.76 in our study population containing elderly patients, which is, in fact, very similar to the C-index obtained from our PIRATE system (C-index: 0.755). In this respect, it is important to note that POSPOM was not developed exclusively for elderly patients, using a derivation cohort with mean age 54.6 years (SD = 17.9 years) and a slightly different outcome definition (all-cause mortality, regardless of whether in-hospital or not) [5].

Thus, our results demonstrate that, by optimising our system on data containing elderly patients only, and focussing on three simple pre-interventional factors, we were able to obtain essentially the same discriminatory power as the more complex POSPOM system.

Prognostic predictors not included in PIRATE

The recently published updated guideline from the European Society of Anaesthesiology and Intensive Care Medicine recommends to assess pre-interventional functional status, level of independence, comorbidity and frailty in the geriatric patient [3]. The PIRATE easy-to-use characteristic *living conditions* is in line with this guideline. While developing PIRATE, we additionally analyzed several pre-interventional patient specific characteristics recommended in the guideline such as *frailty*, and the *type of the planned intervention* whose inclusion in a scoring system might lead to an even more accurate prediction of the post-interventional 30-day probability of death in elderly patients. Although increasing the prognostic power, which is in line with the recommendations of the guideline, those characteristics were not considered for PIRATE for different reasons as outlined in the Results section (i.e. ease of pre-interventional availability and the speed in gathering those) but have been described in previous risk prediction tools [5–13]. Regarding the assessment of frailty, it should be noted that several novel tools with a high accuracy and feasibility have become available during the past years [21]. These include, among others, the clinical frailty scale (CFS) [22], which has been systematically reviewed and recommended for use when predicting mortality and non-home discharge after surgery [23]. Since the CFS and its properties had not been studied in detail at the time POSE was planned, and since it was not possible to gather the CFS data retrospectively, we considered the original POSE frailty score for potential inclusion in PIRATE. The relatively large number of variables needed for the calculation of this score (both clinical and laboratory, see Methods section) led us to the decision to classify frailty as *very hard to gather*. In future studies involving the CFS, frailty will likely be much easier to assess.

Comprehensive geriatric assessment of elderly patients is generally considered to be important for the prognosis of post-interventional 30-day mortality. This has been demonstrated, for instance, by Abete et al. [24], who investigated the impact of surgical scores (e.g., POSSUM), living conditions, disabilities, cognitive function (evaluated by Mini-Mental State Examination, MMSE), depressive symptoms and the severity of comorbidities on 30-day mortality. In line with our results, they demonstrated that POSSUM (developed for patients undergoing emergency and elective surgical procedures, similar to PIRATE) and living conditions (included in the final PIRATE tool) were significantly associated with the 30-day mortality in patients aged 65 years or older [24]. While POSE also collected information on cognitive function (e.g. via the mini-cog test), we did not include these predictor variables in PIRATE, as we aimed to consider only those predictors that are readily available in emergency settings (see

above). In this respect, it should be noted that the study setting considered by Abete et al. differed from POSE not only by the wider age range but also by the exclusion of patients with indication for emergency surgery. The evaluation procedures recommended by Abete et al. could thus be used as a tool to refine PIRATE in non-emergency cases.

Another important risk factor for post-interventional death is sarcopenia [25]. As sarcopenia is characterized by age-related loss of muscle mass and strength, it has been suggested to collect information on falls in elderly study populations and investigate the association between muscle mass, strength, and the prevalence of falls. In a comprehensive evaluation of non-institutionalized people, Curcio et al. [25] demonstrated a strong relationship between the Tinetti Mobility Test (TMT, being an indicator of fall risk) and muscle mass and strength, concluding that TMT represents a tool to detect sarcopenia in elderly patients [25]. In POSE, the mobility of elderly patients was evaluated by the history of falls, and also by the TUG test (both used in the *frailty* assessment). While we considered *frailty* in the development process of PIRATE, we eventually excluded this variable from the set of predictors, as it would be hard to gather the respective information in non-elective interventional settings (please see Step 2, and also the above discussion).

Strengths

The development of the PIRATE scoring system is based on POSE, which was a prospective European multicenter study involving 177 hospitals across 20 countries. As a consequence, PIRATE refers to a broad study population while, at the same time, benefiting from quality-controlled data at the individual patient level collected in a highly standardized setting. We believe that this setting greatly improved estimation and prediction accuracy of the developed scores, even in view of a relatively moderate sample size (at least compared to often-used electronic health record databases involving more patients but employing less standardized methods for data capture).

Generally, the Cox regression model used in the development of PIRATE involves meaningful regression coefficients that have an intuitive interpretation in terms of hazard ratios, relating estimates to established formulas for the derivation of death probabilities. In particular, the use of Cox regression enabled us to translate the estimated regression coefficients into the proposed scoring system [18]. We acknowledge that the prediction accuracy of PIRATE might be improved further by replacing Cox regression with a machine-learning-(ML)-based technique. For example, recent work by Kwon et al. [26] and Seki et al. [27] indicated a strong performance of deep neural networks, random forests, multilayer perceptron and gradient boosting decision trees when used for the prediction of (in-hospital) mortality. However, while increasing prediction accuracy, ML-based predictions often rely on a multitude of predictor variables, which might–or might not–be assessable at the time of surgery. Also, they typically result in "black-box predictions", complicating the interpretation of the predictors' effects and requiring additional electronic support to make predictions on unseen data (e.g., through an online calculator). In contrast, PIRATE has the advantage of being readily applicable without having to use supplementary electronic tools.

By construction of the scoring system, PIRATE allows clinicians to assign risk points to the values of predictors at the individual patient level, including an immediate interpretation of which predictor indicates a worse outcome (e.g. a non-elective surgery leads to a higher probability of post-interventional death within 30 days than an elective one). Basing risk assessment on the scoring system instead of directly computing probabilities of death from the underlying Cox regression model may thus help to improve clinical utility and to establish the tool in daily clinical routine.

Common issues in score development are the transferability to and the external validation on different cohorts. These issues may become a problem when there are non-overlapping sets of risk factors in the derivation and validation cohorts, caused e.g. by different definitions or categorizations of predictors in the respective databases. These problems clearly do not apply to PIRATE, which guarantees a high degree of transferability due to its small number of unambiguously defined and easy-to-determine predictors.

Limitations

Although the PIRATE tool has a number of distinct strengths, there are several limitations to consider. Compared to the development of POSPOM, for instance, which was based on data of 2,717,902 patients with 12,786 in hospital deaths (derivation cohort), the sample size and especially the number of events in the POSE cohort is relatively small [5]. On the other hand, as mentioned earlier, POSE provides prospectively collected data as part of a multicenter study ensuring high data quality compared to routinely collected data.

Importantly, we highlight the need for an external validation of the proposed scoring system. Although we performed an in-depth internal assessment of discrimination and calibration by repeatedly dividing the original POSE cohort on center level into a derivation and validation cohort, we acknowledge that selecting a prediction model based on comparisons of a performance measure (such as the C-index) is not guaranteed to be entirely free of some remaining "optimistic bias". In this respect, external validation studies involving future or unseen data will provide further important insight in the generalization properties of PIRATE. We expect the collaborative network established for the POSE study (involving more than 170 study sites all over Europe) to facilitate the planning and conduct of such studies.

Conclusions

In summary, the proposed PIRATE system constitutes a user-friendly tool to identify patients aged 80 years and older at increased risk of mortality after surgical intervention under anesthesia. PIRATE is readily available and applies to a wide variety of settings. In particular, it covers patients in need for elective or emergency surgery and undergoing in-hospital or day-case surgery. Also, it applies to all types of interventions, from minor to major. Further, PIRATE is in line with recent guidelines, which recommend to apply risk stratification tools to guide anesthesia care in the elderly patient. The scoring system could be used by physicians to evaluate patients' individual risk in order to adapt and customize treatment strategies and post-interventional health care. Future research needs to include an external validation of the scoring system.

Supporting information

S1 File. POSE study group.
(DOCX)

S1 Table. Example application of the PIRATE tool.
(DOCX)

Acknowledgments

Assistance with the study: POSE Study group collaborators substantially contributed to the patient recruitment and acquisition and processing of data. All authors participated in the critical revision and final approval of this manuscript. All authors agree to be accountable for all

aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Author Contributions

Conceptualization: Alina Schenk, Moritz Berger, Matthias Schmid, Mark Coburn.

Data curation: Ana Kowark.

Formal analysis: Alina Schenk, Moritz Berger, Matthias Schmid.

Funding acquisition: Rolf Rossaint, Mark Coburn.

Methodology: Alina Schenk, Moritz Berger, Matthias Schmid, Mark Coburn.

Project administration: Ana Kowark, Rolf Rossaint, Mark Coburn.

Software: Alina Schenk, Moritz Berger, Matthias Schmid.

Visualization: Alina Schenk.

Writing – original draft: Alina Schenk, Moritz Berger, Matthias Schmid.

Writing – review & editing: Ana Kowark, Moritz Berger, Rolf Rossaint, Matthias Schmid, Mark Coburn.

References

1. World Health Organization. World report on ageing and health. World Health Organization. 2015
2. Oresanya LB, Lyons WL, Finlayson E. Preoperative assessment of the older patient: a narrative review. JAMA. 2014; 311(20):2110–2120. https://doi.org/10.1001/jama.2014.4573 PMID: 24867014
3. De Hert S, Staender S, Fritsch G, Hinkelbein J, Afshari A, Bettelli G, et al. Pre-operative evaluation of adults undergoing elective noncardiac surgery: Updated guideline from the European Society of Anaesthesiology. Eur J Anaesthesiol. 2018; 35(6):407–465. https://doi.org/10.1097/EJA.0000000000000817 PMID: 29708905
4. Perioperative care in adults. London: National Institute for Health and Care Excellence (NICE); August 19, 2020.
5. Le Manach Y, Collins G, Rodseth R, Le Bihan-Benjamin C, Biccard B, Riou B, et al. Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and Validation. Anesthesiology. 2016; 124 (3):570–579. https://doi.org/10.1097/ALN.0000000000000972 PMID: 26655494
6. Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. Br J Surg. 1991; 78 (3):355–360. https://doi.org/10.1002/bjs.1800780327 PMID: 2021856
7. Tyagi A, Nagpal N, Sidhu DS, Singh A, Tyagi A. Portsmouth physiological and operative severity score for the Enumeration of Mortality and morbidity scoring system in general surgical practice and identifying risk factors for poor outcome. J Nat Sci Biol Med. 2017; 8(1):22–25. https://doi.org/10.4103/0976-9668.198342 PMID: 28250670
8. Wong DJN, Oliver CM, Moonesinghe SR. Predicting postoperative morbidity in adult elective surgical patients using the Surgical Outcome Risk Tool (SORT). Br J Anaesth. 2017; 119(1):95–105. https://doi.org/10.1093/bja/aex117 PMID: 28974065
9. Golan S, Adamsky MA, Johnson SC, Barashi NS, Smith ZL, Rodriguez MV, et al. National Surgical Quality Improvement Program surgical risk calculator poorly predicts complications in patients undergoing radical cystectomy with urinary diversion. Urol Oncol. 2018; 36(2):77.e1–77.e7. https://doi.org/10.1016/j.urolonc.2017.09.015 PMID: 29033195
10. Haga Y, Ikei S, Ogawa M. Estimation of Physiologic Ability and Surgical Stress (E-PASS) as a new prediction scoring system for postoperative morbidity and mortality following elective gastrointestinal surgery. Surg Today. 1999; 29(3):219–225. https://doi.org/10.1007/BF02483010 PMID: 10192731
11. Sutton R, Bann S, Brooks M, Sarin S. The Surgical Risk Scale as an improved tool for risk-adjusted analysis in comparative surgical audit. Br J Surg. 2002; 89(6):763–768. https://doi.org/10.1046/j.1365-2168.2002.02080.x PMID: 12027988

12. Capuano AW, Shah RC, Blanche P, Wilson RS, Barnes LL, Bennett DA, et al. Derivation and validation of the Rapid Assessment of Dementia Risk (RADaR) for older adults. PLoS One. 2022; 17(3): e0265379. https://doi.org/10.1371/journal.pone.0265379 PMID: 35299231

13. Piccininni M, Rohmann JL, Huscher D, Mielke N, Ebert N, Logroscino G, et al. Correction: Performance of risk prediction scores for cardiovascular mortality in older persons: External validation of the SCORE OP and appraisal. PLoS One. 2020; 15(5):e0233051. https://doi.org/10.1371/journal.pone.0233051 PMID: 32374778

14. POSE-Study group. Peri-interventional outcome study in the elderly in Europe: A 30-day prospective cohort study. Eur J Anaesthesiol. 2022; 39(3):198–209. https://doi.org/10.1097/EJA.0000000000001639 PMID: 34799496

15. Moons KG, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015; 162(1):W1–W73. https://doi.org/10.7326/M14-0698 PMID: 25560730

16. Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Stat Med. 2013; 32(13):2173–2184. https://doi.org/10.1002/sim.5681 PMID: 23172755

17. Kvamme H., & Borgan Ø. The Brier Score under Administrative Censoring: Problems and a Solution. J Mach Learn Res. 2023; 24(2):1–26.

18. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med. 2004; 23(10):1631–1660. https://doi.org/10.1002/sim.1742 PMID: 15122742

19. Rubin DB. Multiple Imputation for nonresponse in surveys. New York, NY: Wiley; 1987.

20. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 2011; 45(3), 1–67.

21. Alkadri J, Hage D, Nickerson LH, Scott LR, Shaw JF, Aucoin SD, et al. A Systematic Review and Meta-Analysis of Preoperative Frailty Instruments Derived From Electronic Health Data. Anesth Analg. 2021; 133(5):1094–1106. https://doi.org/10.1213/ANE.0000000000005595 PMID: 33999880

22. Guidet B, de Lange DW, Boumendil A, Leaver S, Watson X, Boulanger C, et al. The contribution of frailty, cognition, activity of daily life and comorbidities on outcome in acutely admitted patients over 80 years in European ICUs: the VIP2 study. Intensive Care Med. 2020; 46(1):57–69.

23. Aucoin SD, Hao M, Sohi R, Shaw J, Bentov I, Walker D, et al. Accuracy and Feasibility of Clinically Applied Frailty Instruments before Surgery: A Systematic Review and Meta-analysis. Anesthesiology. 2020; 133(1):78–95. https://doi.org/10.1097/ALN.0000000000003257 PMID: 32243326

24. Abete P, Cherubini A, Di Bari M, Vigorito C, Viviani G, Marchionni N, et al. Does comprehensive geriatric assessment improve the estimate of surgical risk in elderly patients? An Italian multicenter observational study. Am J Surg. 2016; 211(1):76–83.e2. https://doi.org/10.1016/j.amjsurg.2015.04.016 PMID: 26116322

25. Curcio F, Basile C, Liguori I, Della-Morte D, Gargiulo G, Galizia G, et al. Tinetti mobility test is related to muscle mass and strength in non-institutionalized elderly people. Age (Dordr). 2016; 38(5–6):525–533. https://doi.org/10.1007/s11357-016-9935-9 PMID: 27566307

26. Kwon JM, Kim KH, Jeon KH, Lee SE, Lee HY, Cho HJ, et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PLoS One. 2019; 14(7):e0219302. https://doi.org/10.1371/journal.pone.0219302 PMID: 31283783

27. Seki T, Kawazoe Y, Ohe K. Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data. PLoS One. 2021; 16(2):e0246640. https://doi.org/10.1371/journal.pone.0246640 PMID: 33544775

3.2 Publication 2: Pseudo-value regression trees

Schenk A, Berger M, Schmid M. Pseudo-value regression trees. In: Lifetime Data Analysis 2024; 30 (2): 439–-471

Link to publication and supplementary information:
https://doi.org/10.1007/s10985-024-09618-x

Implementations are available at:
https://www.imbie.uni-bonn.de/cloud/index.php/s/5oZDBSJjW4pLjtb

Lifetime Data Analysis (2024) 30:439–471
https://doi.org/10.1007/s10985-024-09618-x

Pseudo-value regression trees

Alina Schenk[1] · **Moritz Berger**[1] · **Matthias Schmid**[1]

Received: 23 January 2023 / Accepted: 19 January 2024 / Published online: 25 February 2024

Abstract

This paper presents a semi-parametric modeling technique for estimating the survival function from a set of right-censored time-to-event data. Our method, named pseudo-value regression trees (PRT), is based on the pseudo-value regression framework, modeling individual-specific survival probabilities by computing pseudo-values and relating them to a set of covariates. The standard approach to pseudo-value regression is to fit a main-effects model using generalized estimating equations (GEE). PRT extend this approach by building a multivariate regression tree with pseudo-value outcome and by successively fitting a set of regularized additive models to the data in the nodes of the tree. Due to the combination of tree learning and additive modeling, PRT are able to perform variable selection and to identify relevant interactions between the covariates, thereby addressing several limitations of the standard GEE approach. In addition, PRT include time-dependent effects in the node-wise models. Interpretability of the PRT fits is ensured by controlling the tree depth. Based on the results of two simulation studies, we investigate the properties of the PRT method and compare it to several alternative modeling techniques. Furthermore, we illustrate PRT by analyzing survival in 3,652 patients enrolled for a randomized study on primary invasive breast cancer.

Keywords Gradient boosting · Interactions · Model trees · Pseudo-values · Survival probabilities

Mathematics Subject Classification 62N01 · 62N02 · 62P10

✉ Alina Schenk
schenk@imbie.uni-bonn.de

[1] Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Bonn, Germany

1 Introduction

The estimation of individual-specific survival probabilities is a common task in time-to-event analysis. A plethora of methods has been developed to address this issue, including, among many other approaches, group-wise Kaplan-Meier estimation, Cox regression (Cox 1972), parametric accelerated failure time models (Kalbfleisch and Prentice 2002), and inverse-probability-of-censoring-(IPC)-weighted regression models (Molinaro et al. 2004). Although these approaches are widely used in many disciplines, they often rely on restrictive assumptions limiting their utility. A notable example is the Cox regression model, which requires careful interpretation when the proportional hazards assumption is violated (e.g. Stensrud and Hernán 2020). Similarly, parametric accelerated failure time models may produce invalid results when the underlying distributional assumptions are not met, and IPC-based methods are biased if the working model for the censoring process is misspecified (van der Laan and Robins 2003). Invalid findings may also occur when the complexity of the data-generating process is not fully captured by the model, for instance when relevant covariates are excluded or when interactions between covariates remain undetected (e.g. Vatcheva et al. 2015). In some cases, model misspecification can be avoided by employing methods from the machine learning field (e.g. survival random forests, Ishwaran et al. 2008, or deep neural networks, Lee et al. 2018; Zhao and Feng 2020); however, application of these techniques is often infeasible due to small sample sizes or limitations in the interpretability of the estimated predictor-response relationships. For these reasons, it remains a challenging task to specify time-to-event models yielding accurate and interpretable estimates of individual-specific survival probabilities.

In this paper we propose a novel model building technique named *pseudo-value regression trees* (PRT). Our method is based on pseudo-value regression (Klein and Andersen 2005), which provides a direct modeling framework to estimate the survival function from a set of right-censored time-to-event data. Unlike Cox regression, pseudo-value regression is not based on a statistical model for the hazard function (from which the survival function can subsequently be derived by application of a suitable transformation); instead, it defines a direct link between the survival function and the covariate values on a grid of pre-specified time points $t_1, \dots, t_K$. Usually, K is set to a moderate number, e.g. $K = 5$ or $K = 10$ (see Andersen and Pohar 2010). Given data from a set of n independent individuals with survival times $T_i \in \mathbb{R}^+$ and time-independent baseline covariates $X_i \in \mathbb{R}^p$, $i = 1, \dots, n$, the key idea of pseudo-value regression is to approximate the survival probabilities $S(t_k|X_i) = \mathrm{P}(T_i > t_k|X_i) = \mathrm{E}[\mathbb{1}_{\{T_i > t_k\}}|X_i]$, $k = 1, \dots, K$, by a set of jackknife *pseudo-values*. The latter are defined as

$$\hat{\theta}_i(t_k) = n \cdot \hat{S}_{\mathrm{KM}}(t_k) - (n-1) \cdot \hat{S}_{\mathrm{KM}}(t_k)^{-i}, \quad i = 1, \dots, n, \tag{1}$$

where $\hat{S}_{\mathrm{KM}}(t_k)$ and $\hat{S}_{\mathrm{KM}}(t_k)^{-i}$ denote the Kaplan-Meier estimators based on the complete data and the reduced data (without individual i), respectively. Since it can be shown that $\mathrm{E}[\hat{\theta}_i(t_k)|X_i] \xrightarrow{p} \mathrm{E}[\mathbb{1}_{\{T_i > t_k\}}|X_i]$ as $n \to \infty$ (provided that the censoring mechanism is independent of the event times and the covariates, Graw et al. 2009;

Overgaard et al. 2017), consistent estimates of $S(t_k|X_i)$ can be obtained by fitting a statistical model that regresses the pseudo-values on the covariates (Andersen and Pohar 2010). Unlike IPC-based methods, which often discard the covariate information of censored individuals when used in combination with (weighted) regression techniques (Molinaro et al. 2004), pseudo-value regression is based on a "pseudo" complete data set that includes all available values $X_1, \ldots, X_n$ in the estimation equation (Andersen and Pohar 2010).

The standard approach to fit a pseudo-value model is to specify a monotonically increasing link function $g(\cdot)$ and to use $\hat{\theta}_i(t_k)$ as outcome variable in the regression model

$$g\big(S(t_k|X_i)\big) = g\big(\,\mathrm{E}\left[\mathbb{1}_{\{T_i > t_k\}}|X_i\right]\big) = \alpha_k + \gamma^T X_i\,, \quad k = 1, \ldots, K\,, \tag{2}$$

where $(\alpha_1, \ldots, \alpha_K, \gamma^\top)^\top \in \mathbb{R}^{K+p}$ is a vector of unknown coefficients. Estimation of the coefficients is usually based on generalized estimating equations (GEE, Liang and Zeger 1986), setting $g(\cdot)$ equal to the complementary log-log link function (Andersen et al. 2003; Andersen and Pohar 2010). While the GEE approach accounts for possible dependencies between the pseudo-values $\hat{\theta}_i(t_1), \ldots, \hat{\theta}_i(t_K)$ obtained from the same individual, it is limited by the restrictive definition of the predictor $\eta_{ik} = \alpha_k + \gamma^T X_i$. In particular, η_{ik} does not allow for modeling time-dependent effects (since γ is assumed to be constant in time), and it is restricted to modeling main covariate effects only. Although more flexible effect terms (representing e.g. interactions with time or between the covariates) could be included in (2), we are not aware of any algorithm to identify these terms in a data-driven way. On the other hand, pre-specification of the interaction terms is often infeasible, as it would require detailed knowledge on the, usually hidden, interaction structure in the data-generating process. Another limitation of the standard regression model in (2) is that the intercept terms $\alpha_1, \ldots, \alpha_K$ (representing the "baseline" risk function) are estimated in an unrestricted fashion. As a consequence, the fitted survival probabilities are not guaranteed to decrease with time.

To address these limitations, we extend the standard model in (2) by a semi-parametric approach for the estimation of survival probabilities via pseudo-value regression. Our proposed PRT method is inspired by *logistic model trees* (LMT, Landwehr et al. 2005), which is a popular classification method combining the strengths of tree learning and binary regression by fitting a series of regularized logistic models to the data in the nodes of a classification tree. In order to adapt LMT to pseudo-value regression, we propose to replace the classification tree by a multivariate conditional inference tree (Hothorn et al. 2006) and to use a novel GEE-type optimization criterion for modeling the pseudo-values in the nodes. The proposed PRT method does not require pre-specification of any main or interaction effects, neither among the covariates nor between the covariates and time.

Briefly, the PRT method is characterized by the following steps: First, in order to identify the most important interactions between the covariates, we build a multivariate conditional inference tree (Hothorn et al. 2006) using the pseudo-values as K-dimensional continuous response variable. In the second step, we apply a gradient boosting algorithm with linear base-learners (Bühlmann and Hothorn 2007; Hofner

et al. 2014) to the data in each node of the tree. Our node-wise boosting algorithm is based on the aforementioned GEE-type optimization criterion, including a pre-specified link function to ensure that survival probability estimates are bounded between 0 and 1. Following the idea of LMT, the fitted values of boosting models in higher-level nodes are used as offset values to refine models in lower-level nodes, leading to the stabilization of estimates along single paths. In each node, the fitting of boosting models is stopped early, enabling the selection of relevant covariates. Furthermore, to model interactions between time and the covariates used to build the tree, we include a time-dependent monotonic base-learner (Hofner et al. 2011) in each node-wise model. This base-learner also ensures that survival probability estimates decrease with time.

The result of our model building technique is a set of pseudo-value regression models, each corresponding to a single path from the root node to a terminal node of the conditional inference tree. Due to the combination of tree learning and model-based boosting, the node-wise models include a mixture of interaction and time-dependent effects, all of which are identified in a data-driven way. Furthermore, PRT guarantees interpretability of the node-wise boosting fits by additively combining linear and monotonic base-learners (Hofner et al. 2014). Estimates of individual survival probabilities are obtained by dropping the covariate values down the tree and by evaluating the pseudo-value regression model in the respective terminal node.

The rest of the paper is organized as follows: In Sect. 2.1, we will start with the definition and properties of pseudo-values, including a description of the standard GEE approach for pseudo-value regression. Section 2.2 provides a brief introduction to logistic model trees (Landwehr et al. 2005). Section 3 contains a detailed description of the PRT method, including definitions of the multivariate recursive partitioning and model-based boosting techniques. In Sect. 4 we will present two simulation studies investigating the properties of the PRT method. Furthermore, we will present a comparison to established methods for survival probability estimation. In Sect. 5, we will apply the PRT method to data from the randomized phase III SUCCESS-A trial (de Gregorio et al. 2020), demonstrating that PRT are able to identify subgroups and predictors of disease-free survival in patients with non-metastatic breast cancer. The main findings of the paper are summarized and discussed in Sect. 6, along with a brief overview and discussion of related approaches. Further results and illustrations, as well as details on the implementation of the PRT method, are provided in the Supplementary Material.

2 Prerequisites

2.1 Pseudo-values for survival probability estimation

Consider a set of n independent individuals with survival times T_i and covariate values $X_i = (X_{i1}, \ldots, X_{ip})^\top$, $i = 1, \ldots, n$, that are subject to right-censoring. Denote the censoring times and the observed survival times by C_i and $\tilde{T}_i = \min(T_i, C_i)$, respectively. The status variable Δ_i indicates whether the i-th individual is censored ($\Delta_i = 0$) or whether the event of interest has been observed ($\Delta_i = 1$).

Following Graw et al. (2009), we assume that the censoring times are independent of both the covariates and the event times.

The aim of pseudo-value regression is to model the expectation of a function $\psi(T_i)$ conditional on X_i (Andersen and Pohar 2010). A special case, which will be considered in this paper, is the conditional survival probability $S(t_k|X_i) = \mathrm{E}\left[\mathbb{1}_{\{T_i>t_k\}}|X_i\right]$ for time points t_k, $k = 1, \ldots, K$, with $\psi(T_i) = \mathbb{1}_{\{T_i>t_k\}}$. In order to fit a regression model for $\mathrm{E}\left[\psi(T_i)|X_i\right] = \mathrm{E}\left[\mathbb{1}_{\{T_i>t_k\}}|X_i\right]$, knowledge about the values $\mathbb{1}_{\{T_i>t_k\}}$ is required. In the absence of censoring, $\mathbb{1}_{\{T_i>t_k\}}$ is observable for all individuals: As $T_i = \tilde{T}_i$, it is simply given by $\mathbb{1}_{\{\tilde{T}_i>t_k\}}$. In this case, the Kaplan-Meier estimator is precisely one minus the empirical cumulative distribution function, implying that the pseudo-value $\hat{\theta}_i(t_k)$ (as defined in (1)) coincides with $\mathbb{1}_{\{\tilde{T}_i>t_k\}}$. In the presence of censoring, $\mathbb{1}_{\{T_i>t_k\}}$ is not observable for all individuals; in this case the idea is to replace $\mathbb{1}_{\{T_i>t_k\}}$ by pseudo-values for both, censored and uncensored individuals (Andersen and Pohar 2010).

Figure 1 (A) provides an illustration of pseudo-values in a censoring-free data set (left panel) and in a set of right-censored data (middle and right panels, adapted from Andersen and Pohar 2010). The figure shows that the values $\hat{\theta}_i(t_k)$ are not bounded between 0 and 1 in the presence of censoring. In particular, when focusing on single time points (Fig. 1 (B)), it appears hard to approximate the empirical distribution of pseudo-values by a parametric distribution (as it strongly depends on both the time point and the censoring pattern).

As outlined in Sect. 1, the standard approach to pseudo-value regression is to use the unconditional values $\hat{\theta}_i(t_k)$ as outcome variable in a GEE model of the form (2). Defining the *response function* by $h(\cdot) := g^{-1}(\cdot)$, it is convenient to rewrite Equation (2) as

$$S(t_k|X_i) = \mathrm{E}\left[\mathbb{1}_{\{T_i>t_k\}}|X_i\right] = g^{-1}(\beta^T X_{i,k}) = h(\beta^T X_{i,k})\,, \tag{3}$$

where the augmented covariate vector $X_{i,k} = (0, \ldots, 0, 1, 0, \ldots, 0, X_i^\top)^\top \in \mathbb{R}^{K+p}$ contains an additional set of K binary indicators that are all zero except for the k-th one. The coefficient vector $\beta = (\alpha_1, \ldots, \alpha_K, \gamma^\top)^\top \in \mathbb{R}^{K+p}$ comprises both the baseline risk function and the covariate effects. Common choices for the response function are $h(\beta^T X_{i,k}) = \exp(\beta^T X_{i,k})/(1 + \exp(\beta^T X_{i,k}))$ (corresponding to the logit link) and $h(\beta^T X_{i,k}) = 1 - \exp(-\exp(-\beta^T X_{i,k}))$ (corresponding to the complementary log-log link, Klein and Andersen 2005). Both functions ensure that the survival probabilities $S(t_k|X_i)$ in (3) are bounded between 0 and 1.

Denoting $h(\beta^T X_{i,\cdot}) = (h(\beta^T X_{i,1}), \ldots, h(\beta^T X_{i,K}))^T$ and $\hat{\theta}_i = (\hat{\theta}_i(t_1), \ldots, \hat{\theta}_i(t_K))^\top$, the GEE estimate of β is given by the solution to

$$\sum_i \left\{ \frac{\partial}{\partial \beta} h(\beta^T X_{i,\cdot}) \right\}^T V_i^{-1} \left\{ \hat{\theta}_i - h(\beta^T X_{i,\cdot}) \right\} = 0\,, \tag{4}$$

where $V_i \in \mathbb{R}^{K\times K}$ defines a working covariance matrix accounting for possible dependencies between pseudo-values obtained from the same individual. In practice, V_i is often set to a diagonal matrix (corresponding to an independent correlation structure), as Klein and Andersen (2005) found no advantage of using more

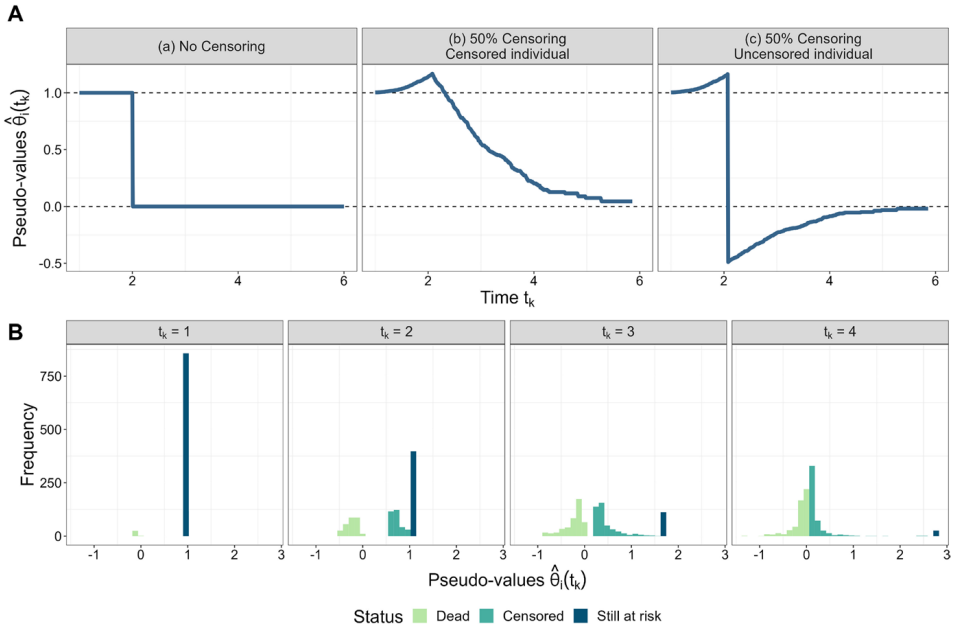


Fig. 1 **A** Illustration of pseudo-values obtained from two data sets with $n = 1000$ individuals each ($0 \leq t_k \leq 6$, adapted from Andersen and Pohar 2010). Panel (a) refers to an individual with $\tilde{T}_i = T_i = 2$ in a censoring-free data set, whereas the other panels refer to a censored individual with $\tilde{T}_i = 2, \Delta_i = 0$ (Panel (b)) and an uncensored individual with $\tilde{T}_i = 2, \Delta_i = 1$ (Panel (c)) in a data set with 50% right-censored survival times. In the censoring-free scenario (a), the pseudo-value at time t_k is simply a binary function indicating whether the individual is still event-free at t_k ($\hat{\theta}_i(t_k) = 1$) or not ($\hat{\theta}_i(t_k) = 0$). In the scenario with 50% censoring, the individuals in (b) and (c) have exactly the same pseudo-values up to their common observed survival time ($\tilde{T}_i = 2$), showing a monotonically increasing pattern. After $\tilde{T}_i = 2$, the pseudo-values of the two individuals differ: While the censoring of the individual in (b) caused $\hat{\theta}_i(t_k)$ to become monotonically decreasing after $\tilde{T}_i = 2$, the observed event in (c) caused $\hat{\theta}_i(t_k)$ to drop to a negative value at $\tilde{T}_i = 2$ and to increase afterwards. **B** Histograms of pseudo-values at different time points in the data set with 50% right-censored survival times from (A). The colors indicate the status of the individuals at the respective time points (dead, censored, still at risk). Pseudo-values of individuals that were observed to experience the event of interest before t_k are negative, whereas pseudo-values are ≥ 1 in individuals that are still at risk at t_k. Obviously, the distribution of the pseudo-values is strongly dependent on both the censoring pattern and the time point of interest

complex versions. As shown by Graw et al. (2009), solving (4) yields a consistent ($n \to \infty$) and asymptotically normal estimator for β, provided that the model is specified correctly.

2.2 Logistic model trees (LMT)

To address the issues described in Sect. 1 and to extend Model (2) to more complex situations containing interactions and time-dependent effects, we propose to build pseudo-value regression models using an adaptation of the LMT method (Landwehr et al. 2005). Originally, LMT have been proposed to develop classification models with a binary outcome. The method consists of two main steps: First, relevant subgroups and interactions are detected by growing a classification tree on the complete

data set (see Sect. 3.1 for details on the tree construction). Second, binary logistic regression models are fitted to the data in each node of the tree, resulting in the estimation of covariate-dependent (node-wise) class probabilities. Unlike earlier approaches to combining tree learning with regression modeling (Quinlan 1992), Landwehr et al. did not fit standard regression models (based on maximum likelihood estimation) but used the LogitBoost method with simple regression functions (Friedman et al. 2000) to build regularized main-effects logistic models (see Sect. 3.2 for details on boosting). Of note, LogitBoost avoids overfitting the data by identifying subsets of the covariates that are most relevant to the node-wise fits. As a consequence, the LMT method performs variable selection at two levels: First, the classification tree selects the covariates that are most relevant to creating subgroups of the data; second, LogitBoost selects the covariates that are most relevant to the node-wise models.

An important characteristic of LMT is the successive refinement of the boosting fits in each tree level, which is achieved by node-wise updates of the LogitBoost coefficients: Starting at the root node of the tree and descending down to the terminal nodes, the LogitBoost coefficients in each daughter node are constructed as updated versions of the coefficients in the respective parent node (Landwehr et al. 2005). Thus, information from higher-level nodes (closer to the root) is incorporated in the models at lower levels, leading to a stabilization of the model fits in the terminal nodes. The estimated class probability for an individual is obtained by dropping the respective covariate values down to a terminal node and evaluating the logistic model associated with that node.

As LMT are a combination of logistic regression and tree learning, they are considerably more flexible than either of the two methods alone, covering both simple main-effects logistic models and standard classification trees as special cases. More specifically, a classification tree of depth 0 (no splits) with a LogitBoost procedure in the root node represents the simple (main-effects-only) logistic model whereas a classification tree of any depth > 0 and no covariates selected by LogitBoost is equivalent to a standard classification tree (Landwehr et al. 2005).

3 Pseudo-value regression trees (PRT)

Given the limitations of the standard GEE approach, and considering the flexibility of LMT in dealing with complex interaction structures, we propose to build pseudo-value regression models by extending the LMT methodology to the estimation of survival probabilities. Briefly, the idea of our *pseudo-value regression trees (PRT)* approach is to replace the binary classification tree by a conditional inference tree with multivariate pseudo-value outcome (accounting for possible dependencies between pseudo-values from the same individual, Sect. 3.1), to replace LogitBoost by a component-wise gradient boosting algorithm (including a time-dependent monotonic base-learner and a novel GEE-type optimization criterion, Sect. 3.2), and to use the successively refined boosting models for the estimation of individual-specific survival probabilities (Sect. 3.3).

3.1 Tree building

The first step of PRT is to grow a regression tree on all available data, replacing the binary outcome of LMT by the pseudo-values $\hat{\theta}_i(t_k) \in \mathbb{R}$, $i = 1, \ldots, n$, $k = 1, \ldots, K$. Generally, the PRT method is not restricted to a specific algorithm for tree building, see e.g. Greenwell (2022) for an overview of the many available options. However, one needs to account for possible dependencies between the pseudo-values obtained for the same individual. To address this issue, we consider the *multivariate conditional inference framework* (Hothorn et al. 2006), which allows for building regression trees with a K-dimensional outcome.

3.1.1 Conditional inference trees

The general idea of tree building is to derive local estimates of the outcome variable by partitioning the covariate space into a set of mutually exclusive subspaces (Breiman et al. 1984; Hothorn et al. 2006; Greenwell 2022). Starting at the *root node* of the tree (comprising all individuals), tree building is done recursively by applying a set of decision rules to the available data. Usually, the decision rules are binary, implying that each node is followed by two *daughter nodes* (each containing a subgroup of the individuals). Tree building is terminated when a pre-defined stopping criterion is reached, resulting in a set of *terminal nodes* from which the local estimates of the outcome are derived. In case of PRT, the local estimates are given by the node-wise boosting fits (see Sect. 3.3).

During tree building, all decision rules are derived locally from the individuals in the respective node. Each rule is characterized by a *split variable* that is selected in a data-driven way from the covariate set. In case of a continuous split variable x^*, the decision rule is defined by $x^* > \xi$ vs. $x^* \leq \xi$, where $\xi \in \mathbb{R}$ is a threshold estimated from the data. In case of a categorical split variable, the decision rule is obtained by dividing the set of categories into two mutually exclusive subsets.

Within this framework, the conditional inference approach (Hothorn et al. 2006) is a method for tree construction that accounts for the distributional properties of the covariates (thereby avoiding a selection bias towards covariates with many possible splits). Decision rules are derived as follows: Given a node with individuals $\mathcal{N} \subseteq \{1, \ldots, n\}$ and data $\mathcal{L} = \{(\hat{\theta}_i(t_1), \ldots, \hat{\theta}_i(t_K), X_{i1}, \ldots, X_{ip}), i \in \mathcal{N}\}$, the first step is to determine the covariate showing the strongest association with the outcome variable. In PRT, this is done by evaluating the generalized correlation coefficients

$$T_j(\mathcal{L}) = \text{vec}\left(\sum_{i \in \mathcal{N}} \tilde{g}_j(X_{ij}) \cdot (\hat{\theta}_i(t_1), \ldots, \hat{\theta}_i(t_K))^\top\right) \in \mathbb{R}^{\tilde{p}_j \times K}, \quad j = 1, \ldots, p, \tag{5}$$

where $\tilde{g}_j(\cdot) \in \mathbb{R}^{\tilde{p}_j}$, $j = 1, \ldots, p$, is a set of transformation functions depending on the measurement scales of the covariates. For the purposes of PRT, we set $\tilde{g}_j(X_{ij}) = X_{ij}$ if the j-th covariate is measured on a continuous scale. For unordered and ordered factors, the functions $\tilde{g}_j(X_{ij})$ are given by a set of dummy variables or some other coding. Next, the elements of $T_j(\mathcal{L})$ are standardized (assuming conditional

independence of the covariates and the outcome, see Hothorn et al. 2006) and transformed using the absolute value function. By this, the standardized and transformed elements of $T_j(\mathcal{L})$ can be interpreted as absolute correlations between the j-th covariate and each of the K pseudo-value elements. Specifically, a separate correlation coefficient is computed at each t_k, $k \in \{1, \ldots, K\}$, so that the dependency between the pseudo-values and time (which is possibly non-monotonic, see Fig. 1) does not affect these calculations. For each j, the maximum value of the absolute correlations is then used to measure the association between the j-th covariate and the K-dimensional pseudo-value outcome and to test the null hypothesis of independence. Altogether, there are p maximum values, resulting in p hypothesis tests. Again, by definition, each of the p maximum values refers to only one time point t_k, $k \in \{1, \ldots, K\}$, so that the tree building step of PRT does not depend on the functional form of the relationship between the pseudo-value outcome and time. Using the default specification in the R package **partykit**, we employ 9,999 permutations to determine the conditional distributions of the maximum values under the null. Finally, the covariate with minimum p-value in the permutation tests is selected as split variable. By definition of this procedure, both the construction of the coefficients in (5) and the implementation of the subsequent hypothesis tests (permuting individuals instead of single pseudo-values) account for the multivariate structure of the vectors $(\hat{\theta}_i(t_1), \ldots, \hat{\theta}_i(t_K))$.

The second step is to derive the actual decision rule associated with the selected covariate. This is done by determining either a threshold ξ (if the selected covariate is continuous) or a grouping of the categories (if the selected covariate is a factor), such that the daughter nodes become maximally dissimilar with respect to the outcome variables. Denoting the set of possible decision rules by $\mathcal{S}$, each decision rule $s \in \mathcal{S}$ is characterized by two mutually exclusive sets of individuals $\mathcal{N}_{\text{left},s}$ and $\mathcal{N}_{\text{right},s}$, referring to the daughter nodes. In order to determine the optimal decision rule, the idea is to maximize

$$\max_{k \in \{1,\ldots,K\}} \left| \frac{\sum_{i \in \mathcal{N}} \mathbb{1}_{\{i \in \mathcal{N}_{\text{right},s}\}} \cdot \hat{\theta}_i(t_k) - \mu_{k,s}}{\sigma_{k,s}} \right| \tag{6}$$

over all decision rules $s \in \mathcal{S}$, where $\mu_{k,s}$ and $\sigma_{k,s}$, denote the conditional means and standard deviations, respectively, of $\sum_{i \in \mathcal{N}} \mathbb{1}_{\{i \in \mathcal{N}_{\text{right},s}\}} \cdot \hat{\theta}_i(t_k)$, $k = 1, \ldots, K$ (computed in the same way as above, cf. Hothorn et al. 2006). By definition, the coefficients in (6) measure the association between node membership and the outcome values; hence, maximizing (6) ensures that the sets of individuals in the daughter nodes become maximally dissimilar with respect to the outcome. Note that each of the decision rules $s \in \mathcal{S}$ depends on the selected covariate; for ease of notation we did not indicate this dependency in (6).

3.1.2 Tuning of the tree

Generally, the partitioning steps described in Sect. 3.1.1 could be applied until each terminal node contains exactly one individual. In case of PRT, this situation would

be clearly undesirable, as a large number of terminal nodes would compromise the interpretability of the tree. Furthermore, the tree tends to overfit the data if the node sizes are too small, leading to numerical issues with the fitting of the boosting models.

To ensure interpretability of the PRT model, we propose to fix the depth D of the regression tree at a small number. In our experiments (Sects. 4.1 and 4.2) we used $D \leq 5$, noting that $D = 5$ (referring to five-way interactions) is already a large value regarding interpretability. Also note that, in some cases, tree building could be terminated before the value D is reached. For instance, the current implementation of the conditional inference tree method in the R package **partykit** stops tree building if all p-values of the permutation tests are larger than a pre-specified threshold. For the purposes of PRT, we set this threshold to 0.05. In addition to restricting the depth of the tree, we require a pre-specified minimum number of observations in each terminal node. In our experiments we set this number to $5 \cdot K$, i.e. to five times the number of time points.

3.2 Component-wise gradient boosting

After having grown the regression tree, the next step is to apply a gradient boosting procedure to the data in each node. Here we propose to consider *component-wise gradient boosting*, as described in Bühlmann and Hothorn (2007) and Hofner et al. (2014).

For gradient boosting it is convenient to organize the data in long format: Since the pseudo-values differ between time points, the idea is to create an augmented data matrix representing each individual by K rows (one per time point, resulting in an overall number of $n \cdot K$ rows). Furthermore, the augmented data matrix includes an additional ID column, as well as a continuous covariate containing the time values $t_1, \ldots, t_K$. In the following, we will refer to the rows of the augmented data matrix as

Table 1 Augmented data of two exemplary individuals, assuming three time points $t_1 = 0.3$, $t_2 = 1.5$, and $t_3 = 3.8$. Each individual is represented by $K = 3$ observations (= rows), each referring to one of the time points. The *ID* column is a factor identifying the individuals, and the covariate values (which are assumed to be time-independent) are replicated K times each (columns $x_1, \ldots, x_p$). The x_0 column refers to an intercept term that is needed for technical reasons in the gradient boosting algorithm

ID	Time	Pseudo-value	x_0	x_1	.	x_p
1	0.3	1.003	1	0.46	.	− 0.27
1	1.5	0.805	1	0.46	.	− 0.27
1	3.8	0.359	1	0.46	.	− 0.27
2	0.3	1.003	1	− 0.18	.	0.14
2	1.5	1.141	1	− 0.18	.	0.14
2	3.8	− 0.822	1	− 0.18	.	0.14

observations (denoted by $\tilde{\imath} \in \{1, \ldots, n \cdot K\}$), in contrast to individuals. Table 1 presents the augmented data of two exemplary individuals.

3.2.1 Details on the algorithm

We first describe the procedure that is applied locally to the data in each node. Given a node with $\tilde{n}$ individuals and $\tilde{n} \cdot K$ observations (denoted by $\mathcal{M} \subseteq \{1, \ldots, n \cdot K\}$), the input of the component-wise gradient boosting procedure is an augmented data set of the form $\{(\hat{\theta}_{\tilde{\imath}}, t_{\tilde{\imath}}, X_{\tilde{\imath}0}, X_{\tilde{\imath}1}, \ldots, X_{\tilde{\imath}p}), \tilde{\imath} \in \mathcal{M}\}$, where $\hat{\theta}_{\tilde{\imath}}$ and $t_{\tilde{\imath}}$ refer to the pseudo-values and the time points, respectively, of the $\tilde{\imath}$-th observation (*Pseudo-value* and *Time* columns in Table 1). Correspondingly, the values $X_{\tilde{\imath}0}, X_{\tilde{\imath}1}, \ldots, X_{\tilde{\imath}p}$ refer to the $x_0, x_1, \ldots, x_p$ columns in Table 1.

The aim of gradient boosting is to estimate an "optimal" prediction function $f^* \in \mathbb{R}$ by minimizing the empirical risk function $\mathcal{R} = \sum_{\tilde{\imath} \in \mathcal{M}} \rho(\hat{\theta}_{\tilde{\imath}}, f_{\tilde{\imath}})$ over the vector $f = \{f_{\tilde{\imath}}\}_{\tilde{\imath} \in \mathcal{M}} = \{f(\mathcal{X}_{\tilde{\imath}})\}_{\tilde{\imath} \in \mathcal{M}}$, where $\mathcal{X}_{\tilde{\imath}}$ is a subset of $\{t_{\tilde{\imath}}, X_{\tilde{\imath}0}, X_{\tilde{\imath}1}, \ldots, X_{\tilde{\imath}p}\}$ and $\rho \in \mathbb{R}$ is a loss function measuring the "deviation" between the outcome and some prediction function $f \in \mathbb{R}$. Note that f^* is not required to depend on all available covariates; instead, the idea is to select the relevant covariates in a data-driven way (hence the term "component-wise", which will be omitted in the following sections for the sake of brevity). The loss function will be described in more detail in Sect. 3.2.2.

Estimation of f^* is performed in an iterative fashion. Starting with some offset values $\hat{f}^{[0]} = \{\hat{f}_{\tilde{\imath}}^{[0]}\}_{\tilde{\imath} \in \mathcal{M}}$, the idea is to minimize the empirical risk function by repeating the following steps: (i) Compute the negative gradient vector $u^{[m]} = -\{\partial \rho / \partial f_{\tilde{\imath}}(\hat{f}_{\tilde{\imath}}^{[m-1]})\}_{\tilde{\imath} \in \mathcal{M}}$ (with m denoting the iteration number), (ii) relate $u^{[m]}$ to the time values and the covariates by a set of univariable regression estimators (denoted by $b_t(t_{\tilde{\imath}}), b_0(X_{\tilde{\imath}0}), b_1(X_{\tilde{\imath}1}), \ldots, b_p(X_{\tilde{\imath}p})$ and fitted separately to the negative gradient vector $u^{[m]}$), (iii) select the regression estimator with the best fit, and (iv) update $\hat{f}^{[m]} = \hat{f}^{[m-1]} + \nu \cdot \hat{u}^{[m]}$, where ν is a step length factor and $\hat{u}^{[m]}$ is the vector of fitted values obtained from the selected regression estimator. For the purposes of PRT, we set $\nu = 0.01$. More details on the algorithm are given in Hofner et al. (2014).

Usually, the boosting algorithm is not run until convergence but "stopped early", implying that the stopping iteration (denoted by m_{stop}) becomes the main tuning parameter of the algorithm (see Sect. 3.2.4). By early stopping, the estimate of f^* is shrunken towards zero, with ν serving as a shrinkage factor. Importantly, as each of the regression estimators is linked to exactly one of the covariates or time, early stopping, together with the selection step in (iii), results in the selection of a subset of relevant covariates. Note that a regression estimator is not removed from the set of candidate estimators after being selected in step (iii), so that the same regression estimator (= covariate) might be selected in multiple iterations.

Generally, the specification of the regression estimators (hereinafter termed *base-learners*) determines the shape of the estimated function $\hat{f}^{[m_{\text{stop}}]}$. In the literature, many types of base-learners have been proposed, including smoothing splines and trees of various depths (Friedman 2001; Bühlmann and Yu 2003; Hofner et al. 2014). To increase the interpretability of the PRT model, we propose to specify simple linear base-learners for the covariates, implying that the estimators $b_j(X_{\tilde{\imath}j})$,

$j = 1, \ldots, p$, refer to a set of simple linear regression models. Following Hofner et al. (2014), we propose to exclude the intercept terms from these models and to specify a separate simple linear model $b_0(X_{\tilde{\imath}0})$ for the constant terms $X_{\tilde{\imath}0} \equiv 1$. Regarding the choice of $b_t(t_{\tilde{\imath}})$, we propose to use a P-spline estimator that is constrained to increase with time, thereby ensuring monotonicity of the baseline risk (see below). For the experiments in Sect. 4 we used the default implementation of monotonic P-splines in the R package **mboost**; details are given in Hofner et al. (2011).

With these specifications, and due to the additive updates in step (iv), the boosting fit at iteration m_{stop} can be written as an additive combination of the covariates and time. More specifically, the estimated values of f^* become equal to

$$\hat{f}_{\tilde{\imath}}^{[m_{\text{stop}}]} = \hat{f}_{\tilde{\imath}}^{[0]} + \sum_{j=0}^{p} \gamma_j X_{\tilde{\imath}j} + \alpha(t_{\tilde{\imath}}), \quad \tilde{\imath} = 1, \ldots, \tilde{n} \cdot K, \tag{7}$$

where the intercept γ_0 and the slope coefficients γ_j, $j = 1, \ldots, p$, are defined by ν times the sum of the coefficient estimates of the simple linear models at the iterations at which the respective base-learners b_j, $j = 0, 1, \ldots, p$, were selected. Analogously, the function $\alpha(\cdot)$ is defined by ν times the sum of the monotonic P-spline functions at the iterations at which b_t was selected. Since b_t is monotonically increasing in time (and since the same holds for the offset values $\hat{f}_{\tilde{\imath}}^{[0]}$, see below), the baseline risk (represented by $\hat{f}_{\tilde{\imath}}^{[0]} + \gamma_0 + \alpha(t_{\tilde{\imath}})$) is also guaranteed to be monotonically increasing in time. This, in turn, leads to a monotonically decreasing survival function for a given set of covariate values, see Sect. 3.3.

In the preceding paragraphs we implicitly assumed that all covariates are measured on a continuous scale. Generally, base-learners for categorical covariates can be specified analogously (e.g. by linear models based on dummy variables or some other coding). Note, however, that care has to be taken when a categorical covariate is binary and when the same covariate has been selected as split variable in some higher-level node of the tree. In this case, the covariate will have zero variance, implying that the respective base-learner has to be excluded from the boosting algorithm. Similar adaptions have to be made for multi-categorical split variables.

Remark: We emphasize that restricting the base-learners to a set of simple main-effects models does not preclude the inclusion of interactions in the final PRT model. This is because gradient boosting is applied node-wise, introducing interactions between the split variable(s) and the variables selected by the boosting algorithm. In particular, the selection of the time base-learner b_t defines a time-dependent effect of the split variable on survival. An illustration of the ability of PRT to model time-dependent effects is given in Section S3 in the supplementary material.

Having defined the node-wise boosting procedure, it remains to (i) specify the loss function $\rho(\hat{\theta}_{\tilde{\imath}}, f_{\tilde{\imath}})$, (ii) define the offset values $\hat{f}_{\tilde{\imath}}^{[0]}$ in Equation (7), and (iii) conceive a strategy for the optimization of m_{stop}. We will elaborate on these issues in the following sections.

3.2.2 Specification of the loss function

Boosting algorithms with a continuous outcome often employ the squared error loss, implicitly assuming normality of $\hat{\theta}_{\tilde{\imath}}$ ("L_2 boosting", Bühlmann and Yu 2003). In case of PRT, this assumption is clearly not appropriate, as the distribution of the pseudo-values is far from normal (see Fig. 1), and as the predicted survival probabilities are constrained to lie in the interval [0, 1]. We therefore propose to use a novel loss function defined by

$$\rho(\hat{\theta}_{\tilde{\imath}}, f_{\tilde{\imath}}) = \left(\hat{\theta}_{\tilde{\imath}} - (1 - \exp(-\exp(-f_{\tilde{\imath}})))\right)^2 = \left(\hat{\theta}_{\tilde{\imath}} - h(f_{\tilde{\imath}})\right)^2, \tag{8}$$

which is inspired by the loss function underlying the GEE approach (assuming a complementary log-log link with $h(f_{\tilde{\imath}}) = 1 - \exp(-\exp(-f_{\tilde{\imath}}))$, see Sect. 2.1). The derivative of this loss function, which is needed to compute the negative gradient vector $u^{[m]}$, is derived as

$$\begin{aligned} \frac{\partial \rho}{\partial f_{\tilde{\imath}}} &= 2 \cdot \exp(-\exp(-f_{\tilde{\imath}})) \cdot \exp(-f_{\tilde{\imath}}) \cdot \left(\hat{\theta}_{\tilde{\imath}} - (1 - \exp(-\exp(-f_{\tilde{\imath}})))\right) \\ &= -2 \cdot \frac{\partial h}{\partial f_{\tilde{\imath}}} \cdot \left(\hat{\theta}_{\tilde{\imath}} - h(f_{\tilde{\imath}})\right). \end{aligned} \tag{9}$$

Under the assumption that $V_i = \text{diag}(1, \ldots, 1) \in \mathbb{R}^{K \times K}$ (corresponding to an independent correlation structure), the derivative in (9) is equivalent to the criterion in (4).

3.2.3 Definition of the offset values

Following the original LMT approach by Landwehr et al. (2005), we propose to refine the node-wise boosting models by passing the characteristics of higher-level boosting fits down to the models in lower-level nodes. The general idea is to incorporate these characteristics in the node-specific offset values $\hat{f}^{[0]}$, also accounting for the time-dependency of the pseudo-values $\hat{\theta}_{\tilde{\imath}}$.

More specifically, given a node with $\tilde{n}$ individuals, $\tilde{n} \cdot K$ observations (denoted by $\mathcal{M} \subseteq \{1, \ldots, n \cdot K\}$), and augmented data $\{(\hat{\theta}_{\tilde{\imath}}, t_{\tilde{\imath}}, X_{\tilde{\imath}0}, X_{\tilde{\imath}1}, \ldots, X_{\tilde{\imath}p}),\ \tilde{\imath} \in \mathcal{M}\}$, we define the offset value for some observation $\tilde{\imath}^* \in \mathcal{M}$ by

$$\hat{f}^{[0]}_{\tilde{\imath}^*} = \frac{1}{\tilde{n}} \sum_{\tilde{\imath} \in \mathcal{M}} \hat{f}^P_{\tilde{\imath}} \cdot \mathbb{1}_{\{t_{\tilde{\imath}} = t_{\tilde{\imath}^*}\}}, \tag{10}$$

where $\hat{f}^P_{\tilde{\imath}}$ denotes the fitted value of the $\tilde{\imath}$-th observation in the parent node. (Note that all observations $\tilde{\imath} \in \mathcal{M}$ contained in the current node are also part of the observations in the respective parent node.) Thus, the offset values $\hat{f}^{[0]} \in \mathbb{R}^{\tilde{n} \cdot K}$ are given by the time-dependent average of the fitted values of $\tilde{\imath} \in \mathcal{M}$ in the respective parent node. Regarding the root node (for which no parent node is available), the offset values $\hat{f}^{[0]} \in \mathbb{R}^{n \cdot K}$ are given by the time-dependent average of the pseudo-values in the complete data. Conceptually, Equation (10) implies that offset values in lower-level

nodes depend on the covariates selected by the boosting models in higher-level nodes. We will elaborate on this dependency in Sect. 3.3.

Remark: The offset calculation described above implies that, in each node, there is a common "average" time trend from which the node-wise boosting model starts iterating. Doing this, the re-calculation of the offset in each node corresponds to "shifts" of the node-wise time trends, followed by the addition of individual-specific effects (via the node-wise boosting models). The rationale of this approach is that the performance of boosting algorithms is known to strongly depend on the choice of a suitable offset value. Often, a good choice is the average value of the predictor (calculated from the data at hand and resulting in a common offset for all individuals, see e.g., Bühlmann and Hothorn 2007). The current implementation of PRT follows this idea. Alternatively, node-wise offset values could be calculated using the observation-wise predictions from the respective parent nodes. When developing the PRT method, we found that the use of average time-dependent offsets (as described above) resulted in better model fits than the "observation-wise" strategy, presumably because taking averages stabilized the fits (in the sense that residual variability in the parent models was better controlled). This is why we eventually decided to implement the "average" strategy in PRT.

3.2.4 Tuning of the gradient boosting models

For the original LMT method, Landwehr et al. (2005) proposed a heuristic to tune the number of LogitBoost iterations in the nodes of the classification tree. Instead of optimizing the iteration number separately in each node ("inner cross-validation", which would have resulted in a high computational effort), the authors determined the optimal value of m_{stop} in the root node (using cross-validation) and applied this value to all other (lower-level) boosting models as well. This approach significantly sped up the algorithm and worked surprisingly well in approximating the node-wise-optimized LMT model. On the other hand, the optimal m_{stop} value determined in the root node is likely too large for the boosting models in the terminal nodes, as these models require more regularization due to the smaller node sizes. Also, cross-validation tends to show a high variability when the node size becomes small. The same is true for a modified strategy that would use the same m_{stop} in all boosting models but would optimize this value across the whole tree (i.e. by minimizing the cross-validated sum of the loss values in Equation (8) computed from the predictions in the terminal nodes). Analogous to Landwehr et al. (2005), we therefore propose a heuristic that avoids the computational burden of optimizing m_{stop} in every node of the tree. The main idea of our strategy is to optimize a single tuning parameter for the whole PRT model (in the following denoted by $m_{\text{stop}}(1)$) and to use this parameter for the calculation of the node-wise iteration numbers. More specifically, for a given value of $m_{\text{stop}}(1)$, we propose to determine the number of boosting iterations in some node $\mathcal{N}$ by

$$m_{\text{stop}}(\mathcal{N}) = \frac{\tilde{n}_{\mathcal{N}}}{n} \cdot m_{\text{stop}}(1)\,, \tag{11}$$

where $\tilde{n}_{\mathcal{N}}$ is the number of individuals in $\mathcal{N}$. By definition, Equation (11) links the node-specific iteration numbers to the numbers of individuals contained in the respective nodes. It also implies that the tuning parameter $m_{\text{stop}}(1)$ becomes equal to the number of boosting iterations in the root node (where $\tilde{n}_{\mathcal{N}} = n$). Unlike the tuning approach of LMT, Eq. (11) does not assign the same iteration number to all nodes; instead, the node-specific values $m_{\text{stop}}(\mathcal{N})$ are forced to decrease as the tree depth increases. By this, our strategy incorporates the well-established result that boosting models applied to smaller data sets (found in lower-level nodes) require more regularization (i.e. smaller values of m_{stop}) than models applied to larger data sets (found in higher-level nodes). In line with this result, the largest number of iterations is assigned to the root node. In our experiments (Sect. 4) we determined the optimal value of m_{stop} by five-fold cross-validation, minimizing the loss function (8) in the terminal nodes (averaged across all observations in the test folds).

Remark: In order to avoid overoptimism in the cross-validation procedure, we computed separate sets of pseudo-values in each of the training and test folds. For the same reason, we grew a new tree on every training fold.

3.3 Calculation of the estimated survival probabilities

After having determined the optimal value of $m_{\text{stop}}(1)$, the gradient boosting models (with iteration numbers $m_{\text{stop}}(\mathcal{N})$) are successively fitted to the data in each node of the conditional inference tree. The last step of PRT is to calculate the individual-specific survival probabilities at all time points $t_1, \ldots, t_K$. This is done as follows: First, each observation $\tilde{\imath} \in \{1, \ldots, n \cdot K\}$ with covariate values $X_{\tilde{\imath}0}, X_{\tilde{\imath}1}, \ldots, X_{\tilde{\imath}p}$ and time point $t_{\tilde{\imath}} \in \{t_1, \ldots, t_K\}$ is dropped down to a terminal node. Note that, by construction of the tree in Sect. 3.1, all K observations (= time points) referring to one individual are part of the same terminal node. Next, the fitted values $\{\hat{f}_{\tilde{\imath}}^{[m_{\text{stop}}(\mathcal{N})]}\}_{\tilde{\imath}=1,\ldots,n\cdot K}$ are calculated from the boosting models in the terminal nodes. Finally, the estimated survival probabilities $\hat{S}_{\tilde{\imath}}(t_{\tilde{\imath}}|X_{\tilde{\imath}})$ are obtained by transforming the fitted values using the response function $h(\cdot)$, giving

$$\begin{aligned}\hat{S}_{\tilde{\imath}}(t_{\tilde{\imath}}|X_{\tilde{\imath}}) &= \sum_{\mathcal{N}\in\mathcal{N}^{\mathcal{T}}} \mathbb{1}_{\{\tilde{\imath}\in\mathcal{N}\}} \cdot h\Big(\hat{f}_{\tilde{\imath}}^{[m_{\text{stop}}(\mathcal{N})]}\Big) \\ &= \sum_{\mathcal{N}\in\mathcal{N}^{\mathcal{T}}} \mathbb{1}_{\{\tilde{\imath}\in\mathcal{N}\}} \cdot \Big(1 - \exp\Big(-\exp\Big(-\hat{f}_{\tilde{\imath}}^{[m_{\text{stop}}(\mathcal{N})]}\Big)\Big)\Big), \quad \tilde{\imath} = 1, \ldots, n \cdot K,\end{aligned} \tag{12}$$

where $\mathcal{N}^{\mathcal{T}}$ indicates the set of terminal nodes. Note that the above procedure also applies to any set of new individuals (possibly not contained in the available data), see Section S2 in the supplementary material for an illustration.

We emphasize that the node membership (indicated by $\mathbb{1}_{\{\tilde{\imath}\in\mathcal{N}\}}$) is determined by a set of binary decision rules depending on at most D split variables. Thus, the fitted survival probabilities in (12) contain (at most) D-way interactions between the split variables and each of the covariates selected by the boosting algorithm. Moreover, the multiplication of $\mathbb{1}_{\{\tilde{\imath}\in\mathcal{N}\}}$ with the baseline risk (contained

PRT in a nutshell

1. Input and data pre-processing
- Input data $\{(X_{i1}, \ldots, X_{ip}, \tilde{T}_i, \Delta_i), i = 1, \ldots, n\}$.
- Specify time points $t_1, \ldots, t_K$.
- Calculate pseudo-values $\hat{\theta}_i(t_k)$, $k = 1, \ldots, K$ (Equation (1)).

2. Determination of the tuning parameter $m_{\text{stop}}(1)$
- Specify the tree depth D and the minimum number of observations in the terminal nodes.
- Create the augmented data matrix with observations $\tilde{\imath} = 1, \ldots, n \cdot K$.
- Specify linear base-learners $b_0(X_{\tilde{\imath}0}), b_1(X_{\tilde{\imath}1}), \ldots, b_p(X_{\tilde{\imath}p})$ and monotonic P-spline base-learner $b_t(t_{\tilde{\imath}})$.
- Carry out five-fold cross-validation. For $\ell = 1, \ldots, 5$:
  - Build a multivariate conditional inference tree with pseudo-value outcome on the ℓ-th training data.
  - Fit node-wise gradient boosting models to the tree (with time-dependent offset values as defined in (10)).
- Determine the optimal $m_{\text{stop}}(1)$ as described in Section 3.2.4.

3. Model building
- Grow the multivariate conditional inference tree on the complete data set.
- Given the optimal $m_{\text{stop}}(1)$, calculate the node-wise stopping iterations $m_{\text{stop}}(\mathcal{N})$ using Formula (11).
- Fit node-wise gradient boosting models to the tree (with time-dependent offset values as defined in (10)).

4. Calculation of estimated survival probabilities
- Drop observations down the tree and evaluate gradient boosting models in the terminal nodes.
- Calculate the estimated survival probablities $\hat{S}_{\tilde{\imath}}(t_{\tilde{\imath}}|X_{\tilde{\imath}})$ using Formula (12).

Fig. 2 Schematic overview of the PRT method

in $\hat{f}_{\tilde{\imath}}^{[m_{\text{stop}}(\mathcal{N})]}$, see Equation (7)), defines a time-dependent effect in each terminal node. Of note, this effect does not only depend on the split variables but also on the covariates selected by higher-level boosting models (which are incorporated in the offset value $\hat{f}_{\tilde{\imath}}^{[0]}$).

Figure 2 presents a schematic overview of the PRT method, summarizing Sects. 3.1 to 3.3.

4 Experiments

To investigate the properties of the PRT method, we carried out two simulation studies. In the first study (Sect. 4.1), the data-generating process was characterized by a tree with lognormal survival models in the terminal nodes (reflecting the true structure of a PRT model). The aim of this study was to analyze whether PRT was able to identify relevant covariates and subgroups defined by the interaction effects. In the second study (Sect. 4.2), the data-generating process was based on a lognormal model with an additive mixture of main and interaction effects. Here,

the aim was to analyze the performance of PRT in the presence of model misspecification (since the additive model did not have a tree structure). Both studies were based on 100 Monte Carlo replications and $K = 5$ time points (following the suggestions by Andersen and Pohar 2010). In each replication, we generated a training data set for model building and a test data set for model evaluation. The sample sizes of all data sets were set to $n = 1000$. The time points were chosen as the mean empirical 10%, 30%, 50%, 70%, and 90% quantiles of $\tilde{T}_i, i = 1, \ldots, 1000$, computed from 1000 additional individuals (generated independently using the data-generating processes of the simulation studies). Five-fold cross-validation on the training data was used to optimize the stopping parameter $m_{\text{stop}}(1)$.

Accuracy was measured by applying the fitted models to the test data set. We computed the mean squared error (MSE) defined by

$$\text{MSE} = \frac{1}{n \cdot K} \sum_{\tilde{\imath}=1}^{n \cdot K} \left(\hat{S}_{\tilde{\imath}}(t_{\tilde{\imath}} | X_{\tilde{\imath}}) - S_{\tilde{\imath}}(t_{\tilde{\imath}} | X_{\tilde{\imath}}) \right)^2, \tag{13}$$

where $\hat{S}_{\tilde{\imath}}(t_{\tilde{\imath}}|X_{\tilde{\imath}})$ and $S_{\tilde{\imath}}(t_{\tilde{\imath}}|X_{\tilde{\imath}})$ denote the estimated and the true survival probabilities, respectively, of observation $\tilde{\imath}$ at time $t_{\tilde{\imath}}$. To evaluate the error on the scale of the survival probabilities, we further computed the root mean squared error (RMSE), defined as the square root of (13). Additionally, we calculated the bias of the estimated survival probabilities (defined as the average deviation of the estimated survival probabilities from their respective true survival probabilities) and the Brier score (Kvamme and Borgan 2023). Discrimination ability was measured using the concordance index (C-index, Gerds et al. 2013). Bias and Brier score values were averaged across the five time points. The time horizon for the C-index was set equal to the largest time point t_5.

4.1 Simulation study 1

We considered a model with ten covariates $x = (x_1, \ldots, x_{10})^T$ that followed a multivariate standard uniform distribution. The correlation matrix of this distribution was generated randomly and was the same in each replication (see Section S4 in the supplementary material). We used the method by Demirtas (2004) to sample the covariate values, restricting the pairwise Pearson correlations between the covariates to 0.5 in absolute value.

Imitating the structure of a tree with $D = 2$, we first formed two subgroups of individuals defined by the decision rule $x_1 \leq \xi_1$ vs. $x_1 > \xi_1$ with $\xi_1 = \text{median}(x_1)$. Afterwards, the groups of individuals with $x_1 \leq \xi_1$ and $x_1 > \xi_1$ were split according to the decision rules $x_2 \leq \xi_2$ vs. $x_2 > \xi_2$ and $x_3 \leq \xi_3$ vs. $x_3 > \xi_3$, with $\xi_2 = \text{median}(x_2 \,|\, x_1 \leq \xi_1)$ and $\xi_3 = \text{median}(x_3 \,|\, x_1 > \xi_1)$, respectively (see Fig. 3). This resulted in four terminal nodes (indicated by the node numbers $m = 3, 4, 6, 7$ in Fig. 3). In each of the terminal nodes, we generated lognormal survival times using different sets of informative and non-informative covariates. Denoting the node-wise linear predictors by

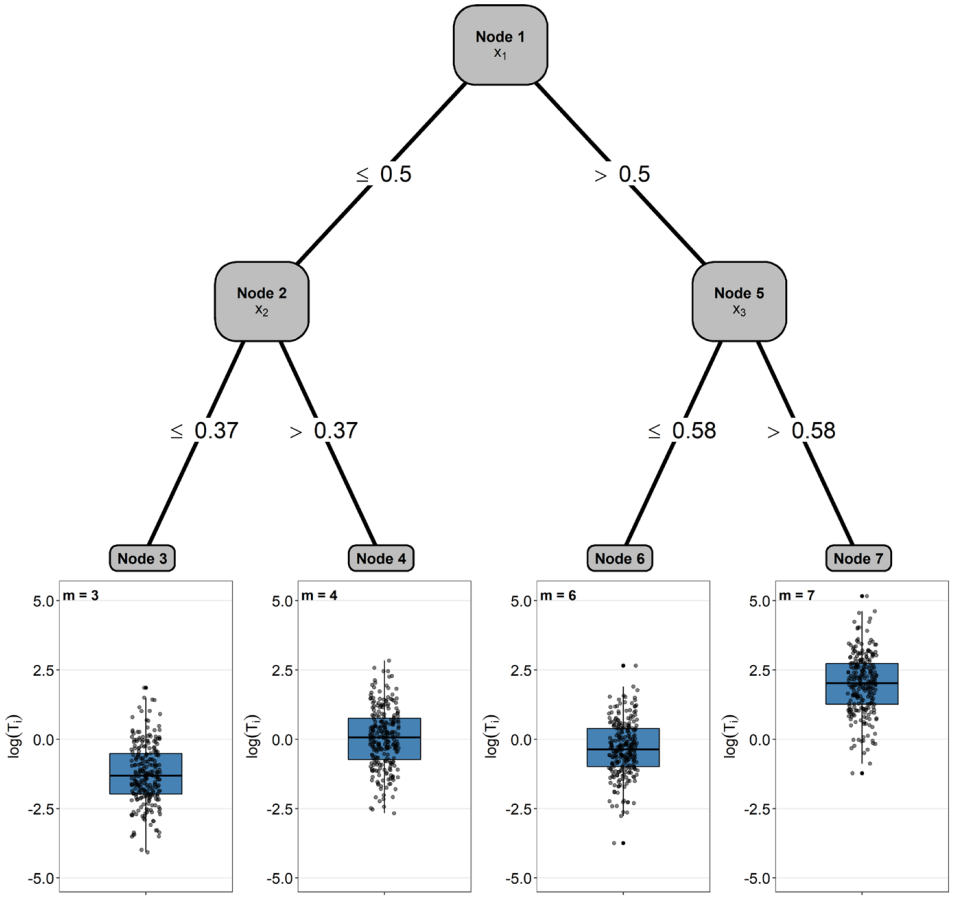


Fig. 3 The plot illustrates the data-generating process of the first simulation study. The boxplots below the terminal nodes were generated from a random sample of size $n = 1000$. They present the distributions of the survival times on the log scale

$$\eta_{im} = \sum_{j=1}^{10} \gamma_j^{(m)} X_{ij}\,, \quad i = 1, \ldots, n, \quad m \in \{3, 4, 6, 7\}\,, \tag{14}$$

the models in the terminal nodes were defined by

$$\begin{aligned} \log(T_i) = &\mathbb{1}_{\{X_{i1} \le \xi_1\}} \cdot \mathbb{1}_{\{X_{i2} \le \xi_2\}} \cdot \eta_{i3} + \\ &\mathbb{1}_{\{X_{i1} \le \xi_1\}} \cdot \mathbb{1}_{\{X_{i2} > \xi_2\}} \cdot \eta_{i4} + \\ &\mathbb{1}_{\{X_{i1} > \xi_1\}} \cdot \mathbb{1}_{\{X_{i3} \le \xi_3\}} \cdot \eta_{i6} + \\ &\mathbb{1}_{\{X_{i1} > \xi_1\}} \cdot \mathbb{1}_{\{X_{i3} > \xi_3\}} \cdot \eta_{i7} \; + \; \epsilon_i\,, \end{aligned} \tag{15}$$

with $\epsilon_i \sim N(0, 1)$, $i = 1 \ldots, n$. In each terminal node, we set the coefficients $\gamma_j^{(m)}$ of five randomly selected covariates to zero. The other coefficients were sampled from continuous uniform distributions with supports $[-1.25, 1.25]$ $(m = 3)$, $[-1, 1]$

Table 2 Covariate effects $\gamma_j^{(m)}$, $j = 1, \ldots, 10$, in the first simulation study. The numbers $m \in \{3, 4, 6, 7\}$ indicate the terminal nodes of the tree in Fig. 3

m	$\gamma_1^{(m)}$	$\gamma_2^{(m)}$	$\gamma_3^{(m)}$	$\gamma_4^{(m)}$	$\gamma_5^{(m)}$	$\gamma_6^{(m)}$	$\gamma_7^{(m)}$	$\gamma_8^{(m)}$	$\gamma_9^{(m)}$	$\gamma_{10}^{(m)}$
3	0	− 1.19	− 0.94	− 0.82	− 0.94	0	− 0.95	0	0	0
4	− 0.63	− 0.47	− 0.43	0.67	0	− 0.54	0	0	0	0
6	0	− 0.47	− 0.39	0	− 0.64	0	0	− 0.09	0	0.01
7	0.50	0.48	0	0.24	0	0	0.68	0.77	0	0

$(m = 4)$, $[-1, 1]$ $(m = 6)$, and $[0, 1]$ $(m = 7)$, see Table 2. Note that the values of the coefficients were the same in each replication. The censoring times C_i were generated independently by the same process, resulting in a censoring rate of 50%.

We first investigated the effect of the tree depth on the performance of PRT. To this purpose, we evaluated the RMSE, the bias, the Brier score and the C-index on a grid of tree depths $D \in \{0, 1, 2, 3, 4, 5\}$. As expected, the absolute bias of PRT decreased with increasing D and remained nearly constant for $D > 2$ (see Fig. 4). The smallest RMSE, the smallest Brier score and the highest C-index values were

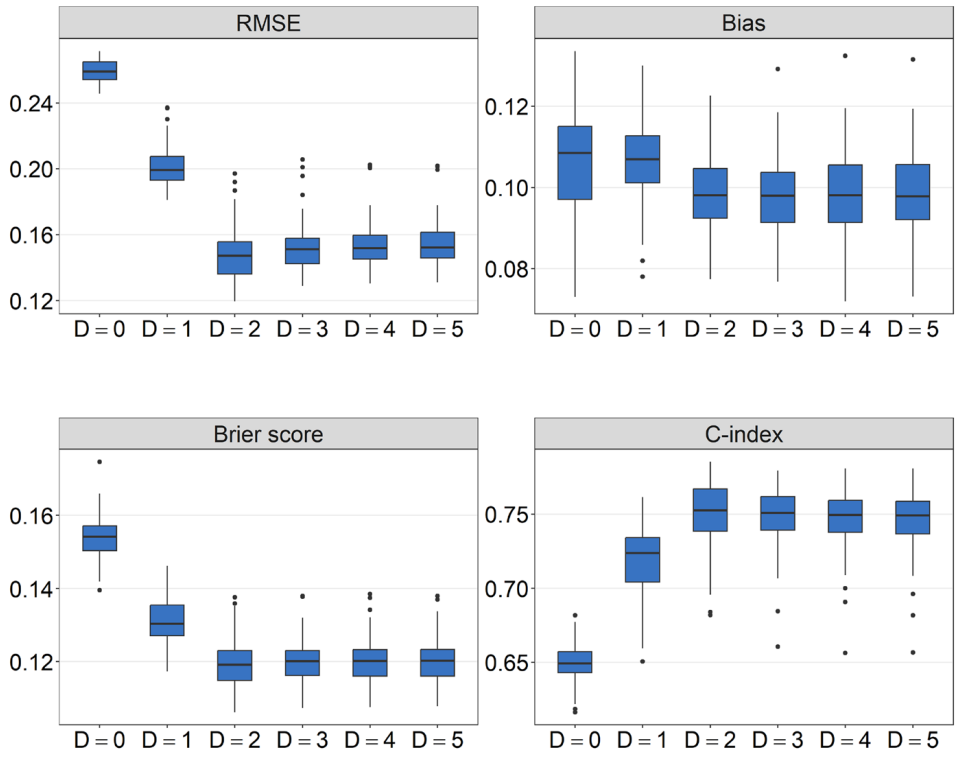


Fig. 4 Results of the first simulation study. The boxplots present the RMSE, bias, Brier score, and C-index values that were obtained by applying the PRT method with varying tree depths $(D \in \{0, 1, 2, 3, 4, 5\})$ to the training data and by evaluating the resulting model fits on the test data. Note that $D = 2$ corresponds to the true tree depth, as defined by the data-generating process

obtained from the PRT model with $D = 2$, matching the true depth defined by the data-generating process. Figure 4 also shows that the RMSE and Brier score values strongly increased and the C-index values strongly decreased when the tree depth was chosen too low. Specifically, the highest RMSE, the highest Brier score and the lowest C-index values were obtained from the PRT model with $D = 0$, which corresponds to a component-wise gradient boosting algorithm in the root node. These results demonstrate the negative effects obtained by ignoring relevant interactions between the covariates. By contrast, increasing D beyond the true depth $D > 2$ did not prove to be particularly harmful with regard to the RMSE, the Brier score and the C-index. Note, however, that large values of D tend to have a negative effect on the interpretability of PRT, increasing both the interaction depth and the number of terminal nodes (see Sect. 3.1.2).

Second, we investigated the ability of PRT to identify the split variables and the informative covariates in the node-wise boosting models. To this purpose, we summarized the selection rates and the coefficient estimates of the PRT fits, setting $D = 2$. As seen from the node labels in Fig. 5, the PRT fits identified the true underlying tree (first split at x_1 in the root node, second splits at x_2 and x_3 in Nodes 2 and 5, respectively) in 80% of the Monte Carlo replications. By contrast, 2% of the fits did only use two of the covariates x_1, x_2, x_3 for splitting; 18% of the fits used the split variables x_1, x_2, x_3 but did not identify the correct order.

Regarding the coefficient estimates of the node-wise boosting models, we observed that informative covariates (defined by $\gamma_j^{(m)} \neq 0$ in the present or in any of the lower-

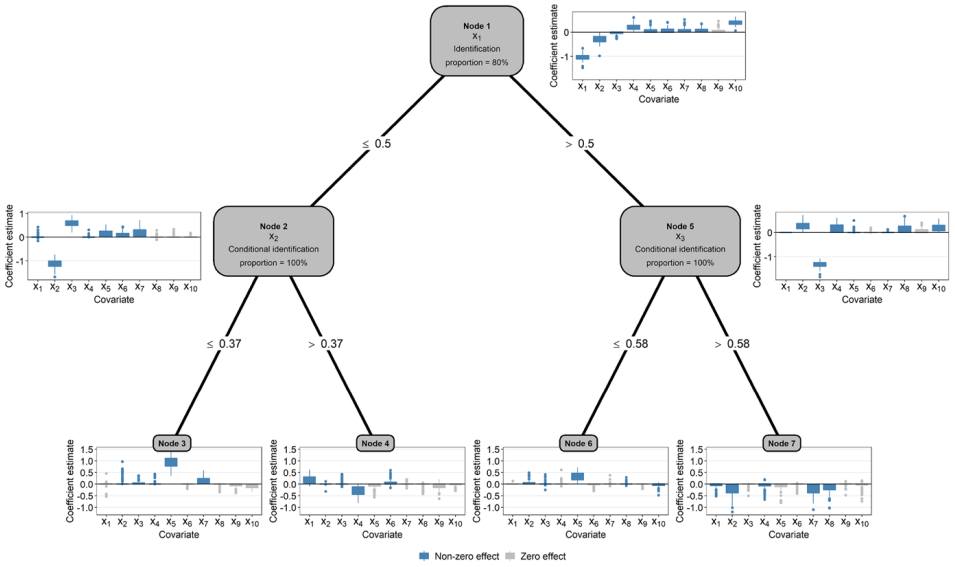


Fig. 5 Results of the first simulation study ($D = 2$, 100 Monte Carlo replications). The plot presents the percentages of correctly identified split variables as well as boxplots of the coefficient estimates obtained from the node-wise boosting fits. In Nodes 2–7, the percentages and coefficient estimates are conditional on having identified the split variable in the parent node. Blue and gray boxplots refer to informative covariates (defined by a non-zero effect in the present or in any of the lower-level nodes) and non-informative covariates, respectively. Coefficient estimates are zero if the respective base-learners were not selected by the gradient boosting algorithm

level nodes) were preferably selected and had higher coefficient estimates in absolute value than non-informative covariates. For example, x_9 (which did not have an effect on survival in any of the terminal nodes, Table 2) had a mean coefficient estimate of only 0.05 (upper right gray boxplot in Fig. 5). An important characteristic of PRT can be observed when comparing the estimated coefficients at various levels of the tree: Obviously, as the tree depth increased, the coefficient estimates of informative covariates became smaller in absolute value. This result is due to the time-dependent offset values in the boosting fits, which, by construction of the PRT method, capture the information of the model fits in higher-level nodes. Consider, for instance, the boosting fits in terminal Node 3 (lower left panel in Fig. 5): Although the coefficient estimates of the informative covariate x_2 appear to be small in this node, the strong negative effect of x_2 is clearly captured by the boosting fits in the parent node (see the blue boxplot to the left of Node 2). The respective coefficient estimates are passed to the daughter nodes via the time-dependent offset values, implying that they are also included in the boosting fits in terminal Node 3. This example demonstrates how the PRT method successively refines the coefficient estimates, passing relevant information from higher levels to the baseline risk in lower-level nodes.

4.2 Simulation study 2

In the second simulation study, we considered a model with an interaction structure that did not match the tree structure of the PRT model. The aim of this study was to evaluate the accuracy of the survival probability estimates in the presence of model misspecification. Furthermore, we compared the PRT method to several alternative modeling techniques.

We considered a model with 30 covariates $x = (x_1, \ldots, x_{30})^\top$ that followed a multivariate normal distribution with mean zero and a randomly generated covariance matrix (see Section S4 in the supplementary material). All covariates had unit variance. For the main covariate effects we defined the linear predictor

$$\eta_{i,1} = \sum_{j=1}^{30} \gamma_j X_{ij}\,, \quad i = 1, \ldots, n. \tag{16}$$

Five out of the 30 covariates were informative, having non-zero coefficients γ_j. Furthermore, we considered all two-way interactions $x_j \cdot x_l$, $j \neq l$, $1 \leq j < l \leq 30$, and defined an interaction-only predictor by

$$\eta_{i,2} = \sum_{1 \leq j < l \leq 30} \gamma_{jl} X_{ij} X_{il}\,, \quad i = 1, \ldots, n. \tag{17}$$

The coefficients γ_{jl} were set to zero if at least one of the covariates x_j or x_l was non-informative. All non-zero coefficients γ_j, γ_{jl} were sampled from a continuous uniform distribution with support $[-1, 1]$. They remained the same in each Monte Carlo replication. The combined predictor (including both the main effects and the interaction effects) was defined by

$$\eta_i = \lambda \cdot \frac{\eta_{i,1} - \text{mean}(\eta_{i,1})}{\text{sd}(\eta_{i,1})} + (1-\lambda) \cdot \frac{\eta_{i,2} - \text{mean}(\eta_{i,2})}{\text{sd}(\eta_{i,2})}, \quad i = 1, \ldots, n, \tag{18}$$

where mean(·) and sd(·) denote the empirical mean and standard deviation, respectively, and $\lambda \in [0, 1]$ is a weighting factor that was included in (18) to analyze the impact of different weightings of the main and interaction effects on the performance of the PRT method. By definition, the predictor in (18) contained only main effects if $\lambda = 1$. Decreasing the value of λ put more weight on the interaction terms, resulting in an interaction-only model if $\lambda = 0$. In our simulation study, we considered $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. Finally, we generated the survival times from a lognormal model defined by

$$\log(T_i) = \frac{\eta_i - \text{mean}(\eta_i)}{\text{sd}(\eta_i)} + \epsilon_i, \quad \epsilon_i \sim N(0, 1), \quad i = 1, \ldots, n. \tag{19}$$

The censoring times were generated in the same way as in the first simulation study, resulting in a censoring rate of 50%.

In addition to analyzing the performance of PRT, we compared our method to the following alternative approaches: (i) a regression tree built using model-based recursive partitioning (MOB, Zeileis et al. 2008, *MOB*), (ii) L_2 boosting with a pseudo-value outcome and tree base-learners of depth two (Friedman 2001, *BoostedTree*), (iii) a survival random forest (not relying on pseudo-values but on the untransformed data $(\tilde{T}_i, \Delta_i, X_i^\top)$, Ishwaran et al. 2008, *SRF*), (iv) a multivariate conditional inference tree without node-wise gradient boosting (built in the same way as in Sect. 3.1 using the multivariate pseudo-value outcome, *TreeOnly*), (v) component-wise gradient boosting with pseudo-value outcome (fitted to the data in the root node only, in the same way as in Sect. 3.2.1, *BoostingOnly*), (vi) the standard GEE approach with complementary log-log link and main effects only (cf. Sect. 2.1, *GEE*), (vii) an inverse-probability-of-censoring-(IPC)-weighted least squares model using log-transformed event times (with main effects only, Molinaro et al. 2004, *IPCW-LS*), (viii) a Cox proportional hazards model with main effects only (*Cox*), (ix) a parametric accelerated failure time model (based on log-transformed survival times and assuming normally distributed errors, *Lognormal*) and (x) the Kaplan-Meier estimator (*KaplanMeier*). For *MOB* and *TreeOnly* we used the same tree depths and minimum numbers of observations in the terminal nodes as for PRT. Along the same lines, we fitted main-effects Weibull models (not using pseudo-values but the untransformed data $(\tilde{T}_i, \Delta_i, X_i^\top)$) in the terminal nodes of the *MOB* tree. These models were based on the proportional hazards assumption, analogous to the complementary log-log link used in PRT. Note that the *BoostedTree* method did not include a time base-learner, effectively ignoring the dependency between pseudo-values of the same individual. Further note that the lognormal model (*Lognormal*) did not include any main and/or interaction terms with zero coefficients. Consequently, the structure of the lognormal model matched the definition of the data-generating process, and we expected this model to be superior to all other methods, providing lower reference values for the RMSE and the Brier score and an upper reference value for

the *C*-index. In contrast, the KM method served as a covariate-free "null" model providing upper reference values for the RMSE and the Brier score and a lower reference value for the *C*-index. Further details on the specification and tuning of the methods are given in Section S4 in the supplementary material.

The RMSE, Brier score and *C*-index values obtained from the second simulation study are presented in Fig. 6. Note that Fig. 6 includes the results for $D = 2$ only, as this tree depth reflects the two-way interaction effects in (17). The results for $D > 2$ were very similar (see Section S5 in the supplementary material). Estimates of the bias of PRT (mostly ranging between −0.05 and 0.05 for all tree depths) are presented in Section S6 in the supplementary material.

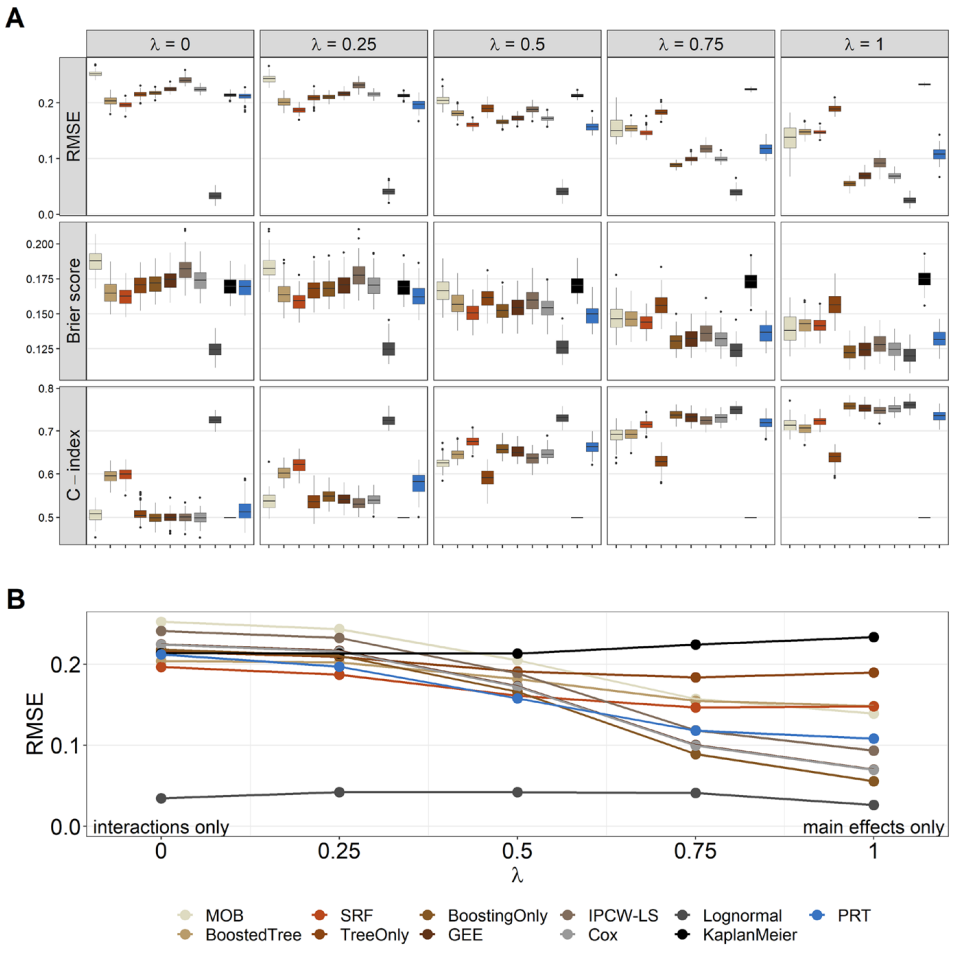


Fig. 6 Results of the second simulation study ($D = 2$, 100 Monte Carlo replications). **A** Boxplots of the RMSE, Brier score and *C*-index values, as obtained by evaluating the model fits on the 100 test data sets. **B** Mean RMSE values (across the replications). Note that *MOB* did not converge in some of the replications (failure rates = 2%, 1%, 2%, 0%, and 1% for $\lambda = 0, 0.25, 0.5, 0.75$, and 1, respectively). The results of these models were excluded from the plots

As expected, the *Lognormal* method outperformed the other methods in all settings, resulting in the smallest RMSE values, lowest Brier score values, and highest *C*-index values by far. As described above, this was because the *Lognormal* model matched the structure of the true data-generating process. For $\lambda = 0$ (no main effects, two-way interactions only), PRT outperformed all other methods except *SRF*, *BoostedTree* and *Lognormal*. Again, this is a plausible result, as the *BoostedTree* algorithm was defined by tree base-learners of depth two, resulting in an additive combination of two-way interactions (thereby matching the true structure of the predictor in case $\lambda = 0$). Similarly, *SRF* is a tree ensemble that is expected to outperform single-tree methods like PRT in terms of prediction accuracy. The *MOB* approach resulted in rather high RMSE values, high Brier score values and low *C*-index values, which was likely due to the high variability of the Weibull fits in the terminal nodes. Note that, unlike PRT, *MOB* does not perform variable selection in the terminal nodes and is not stabilized by offset values containing information from higher-level nodes. This increases the variability of coefficient estimates when the number of covariates is large relative to the node size. For the same reason, *MOB* could not even be fitted in some of the Monte Carlo replications, see the caption of Fig. 6. We also observed that the simple Kaplan-Meier estimator (serving as a covariate-free null model) performed quite well for $\lambda = 0$. This result might be explained by the large numbers of zero main effects (30 out of 30 when $\lambda = 0$) and zero interaction terms (425 out of 435), making it hard for any modeling technique to approximate the true model structure.

When increasing the value of λ to 0.25 (corresponding to models with non-zero main effects but 25% weight on the interaction terms), the PRT method performed better in terms of RMSE, Brier score and *C*-index than all other methods (except *SRF* and *Lognormal*, see above). When main effects and interactions were weighted equally ($\lambda = 0.5$), the PRT method performed best with respect to both RMSE and Brier score (except *Lognormal*, as expected). *C*-index values were highest for *Lognormal* (as expected), followed by *SRF* and PRT with only minor differences between the latter two. Note, in particular, that PRT performed better than *TreeOnly* and *BoostingOnly* in the scenarios with $\lambda \leq 0.5$. This result clearly demonstrates the benefit of combining the two methods if both main and (relevant) interaction effects are present.

In the scenarios with $\lambda \in \{0.75, 1\}$, all tree-based approaches (PRT, *MOB*, *BoostedTree*, *SRF*, *TreeOnly*) were outperformed by the main-effects models *BoostingOnly*, *GEE*, *IPCW-LS*, *Cox* and *Lognormal*. This is another plausible result, as the data-generating process either put a small weight on the interaction terms ($\lambda = 0.75$) or completely excluded the interaction terms ($\lambda = 1$) in these scenarios. Among the main-effects models, *BoostingOnly* performed best (except *Lognormal*, see above), demonstrating the benefit of variable selection and shrinkage in scenarios with a larger number of non-informative covariates. The standard approaches (*GEE*, *Cox*) also resulted in RMSE values that were substantially smaller than those of the tree-based methods. On the other hand, we note that PRT performed best among the tree-based methods, coming closest to the RMSE values of the main-effects models when $\lambda \in \{0.75, 1\}$. Of note, among the single-tree methods (PRT, *MOB* and *TreeOnly*),

PRT was the only method that was able to outperform the ensemble method *SRF* in these settings (with respect to all considered performance measures).

5 Application

To illustrate the PRT method, we analyzed data from the SUCCESS-A trial (NCT02181101), which was a multicenter randomized phase III study that enrolled 3,754 patients with a primary invasive breast cancer between September 2005 and March 2007. All patients had a high recurrence risk, meaning that the SUCCESS-A study population did not constitute a random sample from the general population; for details on the inclusion/exclusion criteria and the design of the study see de Gregorio et al. (2020). The study had two treatment arms, with patients either receiving one of the standard chemotherapy regimens (control group) or standard chemotherapy with the addition of gemcitabine (experimental group). The randomization ratio was 1:1. The aims of SUCCESS-A were to compare the two groups with respect to disease-free survival (DFS) and overall survival (OS) within a five-year follow-up period. Here we focus on DFS, which, according to the STEEP system, was defined as the period from the date of randomization to the earliest date of disease progression (distant metastases, local and contralocal recurrence, and secondary primary tumors or death from any cause, de Gregorio et al. 2020). Since the definition of DFS included death from any cause, we did not consider death as a competing event.

Patients were censored at the last date on which they were known to be disease-free, resulting in an event rate of 12.2% (458 events in 3,754 patients). The maximum observation time was 5.5 years (6 months of chemotherapy followed by 5 years of follow-up; median 5.2 years, first quartile 3.7 years, third quartile 5.5 years). In addition to the survival times, the study collected data on several established prognostic factors, including age at randomization (*age*, measured in years), body mass index (*BMI*, measured in kg/m^2), tumor stage (*stage*, four categories, pT1/pT2/pT3/pT4), tumor grade (*grade*, three categories, G1/G2/G3), lymph node status (*nodal status*, two categories, pN0/pN+), tumor type (*type*, three categories, ductal/lobular/other), menopausal status (*meno*, two categories, pre-/post-menopausal), and receptor status for estrogen (*ER*), progesterone (*PR*), and *HER2* (two categories each, negative/positive), see de Gregorio et al. (2020). A descriptive summary of the prognostic variables is given in Table S3 in Section S7 in the supplementary material.

A key issue in the development of treatment guidelines for breast cancer is the identification of patient subgroups with possibly different risks of disease progression (Coates et al. 2015; Senkus et al. 2015). To illustrate the PRT method, we investigated the existence of such subgroups in the SUCCESS-A data, noting that tree-based methods have a long tradition in medical risk assessment (including, among other techniques, univariate survival trees, LeBlanc and Crowley 1992; Bacchetti and Segal 1995, tree-structured classification and regression, Ciampi et al. 1995; Puth et al. 2020, and mixtures of survival trees, Jia et al. 2022). In addition to using the aforementioned prognostic factors as covariates, we included the group status (*group*, two categories, control/experimental) and *time* (monotonic P-spline base-learner) in our model. Pseudo-values for DFS were computed at 1, 2, 3, 4, and

5 years ($K = 5$). The depth of the regression tree was fixed at $D = 3$. Our rationale for choosing this number was that it allowed for capturing interaction effects while, at the same time, resulting in a tree with a reasonably simple interpretation (analogous to the specification of the interaction order in linear regression). Patients with missing values in any of the variables (102 patients, 2.7%) were excluded from analysis, resulting in an analysis data set with $n = 3{,}652$ patients. The accuracy of the PRT model was evaluated by computing five-fold cross-validated values of the concordance index (C-index, Uno et al. 2011, with five-year time horizon) and the Brier score (Kvamme and Borgan 2023, averaged across the five time points).

As mentioned above, the aim of applying PRT to the SUCCESS-A data was to illustrate our method but not to optimize it with respect to prediction accuracy. For sensitivity analysis, we additionally present the Brier score and C-index values obtained from PRT with tree depths fixed at $D \in \{2, 4, 5\}$, corresponding to maximum numbers of 4, 16 and 32 terminal nodes, respectively. We also compared PRT

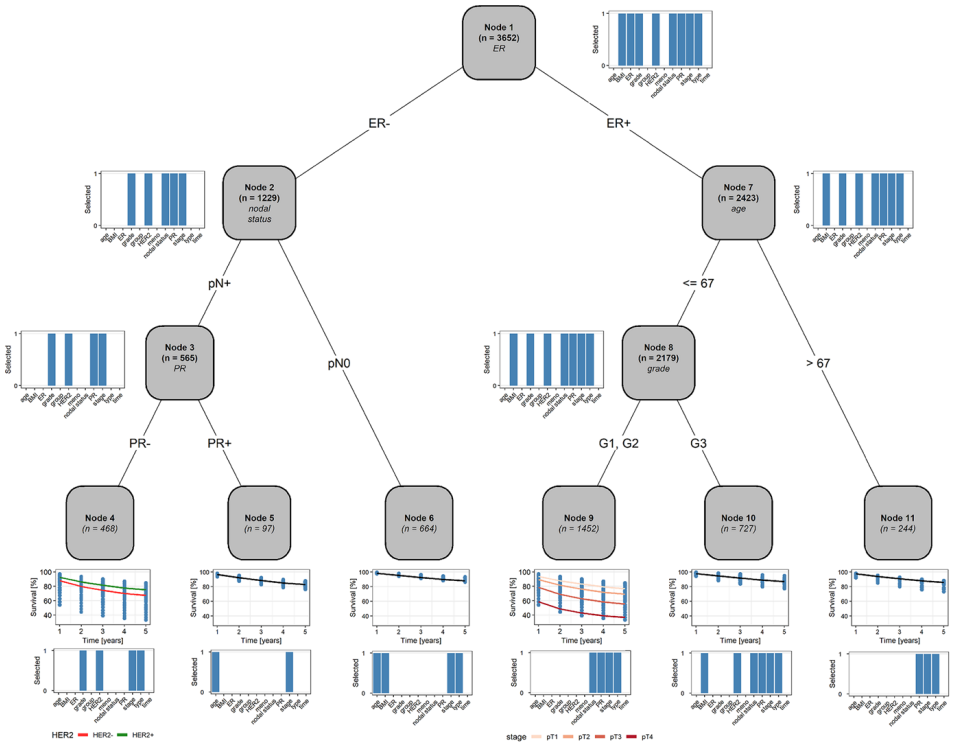


Fig. 7 Analysis of disease-free survival in the SUCCESS-A study data. The figure presents the results obtained from fitting a PRT model with $D = 3$, showing the selected split variables and the sizes of the patient subgroups in the nodes. The blue bars refer to the base-learners selected in the node-wise boosting models. The blue dots and the black lines refer to the fitted values and their averages in the terminal nodes. In Node 4, the mean estimated DFS function of the group of "triple negative" patients (i.e. negative *ER*, *PR* and *HER2*, von Minckwitz et al. 2012) is marked red. The green line refers to mean estimated DFS in the group of HER2 receptor-positive patients. The colored lines in Node 9 refer to the mean estimated DFS functions stratified by tumor stage (light red = pT1, dark red = pT4)

to the methods described in Sect. 4.2. Note that we had to fix the tree depth of the *MOB* method at $D = 1$, as larger values resulted in convergence issues.

The regression tree obtained from the PRT fit is shown in Fig. 7. Overall, the tree structure reflected several established prognostic factors and subgroups, which have been frequently reported in the literature and have also been included in treatment guidelines for breast cancer (Coates et al. 2015; Senkus et al. 2015). Specifically, the first split variable (selected in the root node) was *ER*, indicating the importance of this variable in adjuvant hormonal and chemotherapeutic treatment regimens. The survival advantage of estrogen receptor-positive patients (Goldhirsch et al. 2003) is reflected by the estimated five-year DFS probabilities, which were 80.87% on average in Nodes 4, 5, and 6 and 89.89% on average in Nodes 9, 10, and 11 (corresponding to estrogen receptor-negative patients and estrogen receptor-positive patients, respectively). The split variables in the second level of the tree were *nodal status* and *age* (threshold = 67 years), reflecting the higher risk of lymph node-positive patients (Senkus et al. 2015) and the increased risk of patients aged 67 or older, respectively (Chen et al. 2016). This result is confirmed by the average estimated five-year DFS probabilities, which were smaller in lymph node-positive patients than in lymph node-negative patients (72.46%, Nodes 4 and 5, vs. 88.04%, Node 6), and were higher in patients aged ≤ 67 years than in patients aged > 67 years (90.34%, Nodes 9 and 10, vs. 85.82%, Node 11). Patients with negative estrogen receptor status and positive lymph node status were further split into subgroups defined by *PR*. Of note, progesterone receptor-negative patients were estimated to have the lowest average five-year DFS probabilities (70.31%, Node 4). This group of patients also included the high-risk group of "triple negative" patients (negative *ER*, *PR*, and *HER2*, von Minckwitz et al. 2012), given through the base-learner for *HER2* selected by the boosting model in Node 4. In line with the literature (von Minckwitz et al. 2012), the subgroup of triple negative patients had lower estimated five-year DFS probabilities on average than HER2 receptor-positive patients in Node 4 (red vs. green lines in the lower left panel of Fig. 7). Another prognostic variable is *grade*, which was selected as split variable in the group of patients ≤ 67 years in Node 8. We note that the grouping of *grade* (G1/G2 vs. G3) reflects the grouping in current treatment guidelines for breast cancer (Coates et al. 2015; Senkus et al. 2015). As expected, patients with a low or intermediate grade had a higher estimated five-year DFS probability (G1/G2, 92.03%, Node 9) than patients with a high grade (G3, 86.97%, Node 10). Furthermore, tumor stage (selected by the boosting model) had a strong impact on survival in Node 9 (see colored lines in Fig. 7). Regarding the treatments investigated in the SUCCESS-A trial, we observed that the *group* was not selected in any of the nodes, neither for splitting nor as base-learner in the boosting models. This is in line with the findings in the original study report by de Gregorio et al. (2020), who concluded that the addition of gemcitabine to standard chemotherapy did not improve DFS.

The five-fold cross-validated Brier score and *C*-index values obtained from PRT and the alternative methods are shown in Table 3. It is seen that the Brier score and *C*-index values obtained from PRT were similar for all considered tree depths, suggesting that higher values of *D* did not increase predictive performance but only led to a more difficult interpretation of the models. Overall, PRT were very similar

Table 3 Analysis of the SUCCESS-A study data. The table presents the five-fold cross-validated values of the time-averaged Brier score and the *C*-index at 5 years, as obtained from fitting PRT (with $D \in \{2, 3, 4, 5\}$) and the alternative methods to the study data

Method	Average Brier score	*C*-index at 5 years
PRT ($D = 2$)	0.069	0.660
PRT ($D = 3$)	0.069	0.660
PRT ($D = 4$)	0.069	0.651
PRT ($D = 5$)	0.069	0.662
MOB	0.069	0.666
BoostedTree	0.081	0.618
SRF	0.069	0.668
TreeOnly ($D = 2$)	0.070	0.596
TreeOnly ($D = 3$)	0.070	0.621
TreeOnly ($D = 4$)	0.069	0.639
TreeOnly ($D = 5$)	0.069	0.644
BoostingOnly	0.068	0.670
GEE	0.070	0.667
IPCW-LS	0.291	0.597
Cox	0.068	0.670
Lognormal	0.068	0.669
KaplanMeier	0.072	0.500

to the alternative methods in terms of prediction accuracy (except *TreeOnly* and *KaplanMeier*, which performed worse than PRT as expected, and *BoostedTree* and *IPCW-LS*, which also performed worse than PRT). These results support the plausibility of the above interpretations and the validity of the PRT model in Fig. 7.

6 Discussion

This paper presents a semi-parametric approach for building time-to-event models with a pseudo-value outcome. Our method, entitled pseudo-value regression trees (PRT), results in a piecewise regression model for the survival function, where the "pieces" are obtained by recursively partitioning the covariate space. As described in Sect. 3, developing a model tree algorithm for pseudo-values involved, among other components, a method for multivariate tree building, a loss function for non-normal continuous outcomes, and an appropriately defined time base-learner to ensure monotonicity of the probability estimates. Our numerical experiments in Sects. 4 and 5 demonstrated that the PRT method was able to identify relevant covariates and interactions (Sect. 4.1), showed a favorable estimation accuracy (Sect. 4.2), and yielded highly plausible results in our application on primary invasive breast cancer (Sect. 5). Importantly, by restricting the tree depth to a moderate value ($D \leq 5$), the fitted PRT models had an easily accessible interpretation (see e.g. Fig. 7). This is considered to be a major advantage when the focus is not solely on prediction accuracy, especially when compared to black-box methods like support vector machines,

random forests, or deep neural networks (see e.g. Mogensen and Gerds 2013; van der Ploeg et al. 2014; Zhao and Feng 2020; Rahman et al. 2021).

Conceptually, the PRT method belongs to the class of "direct" modeling approaches, relating the covariates directly to the survival probabilities instead of relating them "indirectly" to $S(t|X_i) = \exp(-\int_0^t \lambda(u|X_i)du)$ via the hazard function $\lambda(t|X_i)$ (as done e.g. by the Cox model and Aalen's additive hazard model). Prominent examples of direct models are the proportional odds model and the Cox-Aalen model, which can be fitted to a set of censored time-to-event data using inverse-probability-of-censoring-(IPC)-weighted binomial regression (Scheike et al. 2008, see also Grøn and Gerds 2014 and the references therein). Analogous to pseudo-value regression, these models provide estimates of $S(t_k|X_i)$ on a pre-defined grid of time points $t_k = t_1, \ldots, t_K$. The same is true for the hierarchical modeling approach by Garcia et al. (2019), which is a mixture of binomial regression and pseudo-value regression; instead of using IPC weights (effectively excluding censored individuals from the estimation equation), the authors replaced the binary values of *censored* individuals by pseudo-values and fitted the (pseudo-)binomial model within the generalized additive modeling framework. Hothorn et al. (2014) proposed the class of *conditional transformation models*, which is a general approach to model the distribution function $F(t|X_i) = 1 - S(t|X_i)$ conditional on a set of covariates (including direct survival models as special cases). Of note, Hothorn et al. (2018) developed a likelihood-based approach for the modeling of $F(t|X_i)$ that does not require pre-specification of a grid of time points (see also Hothorn 2019 and the references therein). Similar to Garcia et al. (2019), Hothorn et al. (2018) proposed to model the baseline risk and the covariate effects using basis functions. Despite the flexibility of the aforementioned approaches, we emphasize that the *building* of survival models (in particular, the specification of the model structure) remains a challenging task. Tree-based methods like PRT are useful in addressing this issue, providing tools for variable selection and the identification of interaction effects. On the other hand, the selection steps performed by PRT (and also by related tree methods) preclude the application of standard hypothesis tests in the nodes (Loh et al. 2019). As a consequence, tree-based methods like PRT should be handled with care if the model structure is fixed and if statistical inference is of major interest.

The PRT approach is also related to other methods for building model trees. In Sect. 4, for instance, we used the *model-based recursive partitioning* approach (MOB) as a comparator to PRT. Conceptually, PRT and MOB are of similar nature; however, they differ with respect to their tree building approaches: While PRT uses the generalized correlation coefficient in (5) for (multivariate) recursive partitioning, MOB applies a test for parameter instability (Zeileis and Hornik 2007) to determine the split variable in each node. In contrast to (5) (which is based on the bivariate relationships between the pseudo-values and the covariates), this instability test requires the node-wise fitting of a regression model including all covariates. As a consequence, MOB is usually more sensitive in detecting interaction effects in the models of interest (controlling for possible confounding instead of considering the "marginal" distributions as in (5)); on the other hand, the validity of the test results might be compromised by multicollinearity, especially when the number of covariates is large relative to the node size (see Sect. 4). To the best of our knowledge,

there exists no regularized version of the MOB algorithm (performing e.g. variable selection like the boosting models in PRT). We further note that the current implementation of MOB in the R package **partykit** does not allow for modeling correlated observations (e.g. via mixed-effects models or a multivariate tree as in PRT). Similar arguments hold for the GUIDE algorithm, which has recently been extended by Loh et al. (2019) to build model trees within the proportional hazards framework.

As stated in Sect. 1, the consistency results by Graw et al. (2009) rely on the assumption that the censoring times C_i are independent of the survival times T_i ("random censoring"). Under this assumption, it can be shown that $\mathrm{E}\,[\hat{\theta}_i(t_k)|X_i] \to \mathrm{E}\,[\mathbb{1}_{\{T_i>t_k\}}|X_i] = S(t_k|X_i)$, so that using the pseudo-values as outcome variable in a statistical model (as done by PRT) is equivalent to substituting the outcome values of interest (here, the unobserved survival probabilities) by consistent estimates of these values. Later, Binder et al. (2014) showed that the random censoring assumption can be relaxed to allow for censoring times C_i that are only *conditionally* independent of T_i given the covariate values X_i. More specifically, the authors considered a scenario where pseudo-values are based on the Aalen-Johansen estimator of the cumulative incidence function $\mathrm{P}\,(T_i \leq t_k)$. By fitting a regression model for the censoring survival function $G(t_k|X_i) := \mathrm{P}\,(C_i > t_k|X_i)$ and incorporating the resulting IPC weights in the Aalen-Johansen estimator, they were able to eliminate the bias occurring from a "naive" calculation of the pseudo-values ignoring dependency on X_i. Analogously, the PRT method could be extended to scenarios with covariate-dependent censoring. This could be done by replacing the Kaplan-Meier estimators in (1) by one minus the respective IPC-weighted Aalen-Johansen estimators. We point out that the results by Binder et al. (2014) rely on the correct specification of the regression model for $G(t_k|X_i)$.

We further note that the pseudo-value methodology is not restricted to the estimation of survival probabilities from right-censored data. For example, Andersen and Pohar (2010) considered a general class of functionals of the form $\mathrm{E}\big[\psi(T_i)|X_i\big]$, suggesting that any of these functionals could be estimated by an appropriately defined pseudo-value regression model. In the same manner, the PRT approach could be adapted to a wider class of functionals, an obvious example being the cumulative incidence function in competing-risks analysis. Following the idea described in Zhao et al. (2020), PRT could also be applied for dynamic risk prediction. Furthermore, PRT can easily be extended to incorporate left-truncated survival times referring to individuals not yet at risk at time $t = 0$. Provided that the truncation times are independent of the survival times T_i (at least conditional on X_i, see above), this could be done by an appropriate definition of the risk sets used in the calculation of the Kaplan-Meier estimators in (1). With regard to the latter, robustness can be increased by computing "stopped" pseudo-values $\hat{\theta}_i(t_k)$ that are based on only those individuals who entered the sample before the respective time points t_k. For details, see Grand et al. (2019).

Finally, we would like to note that PRT is, in general, applicable to high-dimensional scenarios with $p > n$. Naturally, very large numbers of covariates may result

in increased run-times (which are mainly due to the permutation tests conducted for the selection of the split variables). It should also be noted that both conditional inference trees and gradient boosting usually require some sort of pre-filtering in order to work well in "ultra-high"-dimensional scenarios with $p \gg n$. The latter aspects are, of course, not specific to PRT but apply to many methods for modeling high-dimensional data (e.g. to penalized regression and even to random forests).

Software

All computations were carried out using the R Language for Statistical Computing (version 4.1.2, R Core Team 2022). An implementation of the PRT method is available at https://www.imbie.uni-bonn.de/cloud/index.php/s/5oZDBSJjW4pLjtb. Details are given in Section S1 in the supplementary material.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10985-024-09618-x.

Acknowledgements We thank Dr. Lothar Häberle (Department of Gynecology, Obstetrics and Mammology, University Hospital Erlangen, Germany) for supporting us with the analysis of the SUCCESS-A study data.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Andersen PK, Pohar Perme M (2010) Pseudo-observations in survival analysis. Statist Methods Med Res 19:71–99

Andersen PK, Klein JP, Rosthøj S (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 90:15–27

Bacchetti P, Segal MR (1995) Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of AIDS. Lifetime Data Anal 1:35–47

Binder N, Gerds TA, Andersen PK (2014) Pseudo-observations for competing risks with covariate dependent censoring. Lifetime Data Anal 20:303–315

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Taylor & Francis, New York

Bühlmann P, Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting. Statist Sci 22:477–505

Bühlmann P, Yu B (2003) Boosting with the L2 loss: regression and classification. J Am Statist Associat 98:324–339

Chen HL, Zhou MQ, Tian W, Meng KX, He HF (2016) Effect of age on breast cancer patient prognoses: a population-based study using the SEER 18 database. PLoS One 11(10):e0165409

Ciampi A, Negassa A, Lou Z (1995) Tree-structured prediction for censored survival data and the Cox model. J Clin Epidemiol 48:675–689

Coates AS, Winer EP, Goldhirsch A, Gelber RD, Gnant M, Piccart-Gebhart MJ, Thürlimann B, Senn H (2015) Tailoring therapies - improving the management of early breast cancer: St. Gallen international expert consensus on the primary therapy of early breast cancer 2015. Ann Oncol 26:1533–1546

Cox DR (1972) Regression models and life-tables. J Royal Statist Soc Ser B 34:187–220

de Gregorio A, Häberle L, Fasching PA, Müller V, Schrader I, Lorenz R, Forstbauer H, Friedl TWP, Bauer E, de Gregorio N, Deniz M, Fink V, Bekes I, Andergassen U, Schneeweiss A, Tesch H, Mahner S, Brucker SY, Blohmer JU, Fehm TN, Heinrich G, Lato K, Beckmann MW, Rack B, Janni W (2020) Gemcitabine as adjuvant chemotherapy in patients with high-risk early breast cancer - results from the randomized phase III SUCCESS-A trial. Breast Cancer Resh 22(1):111

Demirtas H (2004) Pseudo-random number generation in R for commonly used multivariate distributions. J Modern Appl Statist Methods 3:485–497

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Statist 29:1189–1232

Friedman JH, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. Ann Statist 28:337–407

Garcia TP, Marder K, Wang Y (2019) Time-varying proportional odds model for mega-analysis of clustered event times. Biostatistics 20:129–146

Gerds TA, Kattan MW, Schumacher M, Yu C (2013) Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Statist Med 32:2173–2184

Goldhirsch A, Wood WC, Gelber RD, Coates AS, Thürlimann B, Senn HJ (2003) Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. J Clin Oncol 21:3357–3365

Grand MK, Putter H, Allignol A, Andersen PK (2019) A note on pseudo-observations and left-truncation. Biomet J 61:290–298

Graw F, Gerds TA, Schumacher M (2009) On pseudo-values for regression analysis in competing risks models. Lifetime Data Anal 15:241–255

Greenwell B (2022) Tree-based methods for statistical learning in R. Chapman & Hall/CRC, Boca Raton

Grøn R, Gerds TA (2014) Binomial regression models. In: Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH (eds) Handbook of survival analysis. Chapman and Hall CRC, Boca Raton, pp 221–242

Hofner B, Müller J, Hothorn T (2011) Monotonicity-constrained species distribution models. Ecology 92:1895–1901

Hofner B, Mayr A, Robinzonov N, Schmid M (2014) Model-based boosting in R: a hands-on tutorial using the R package mboost. Computat Statist 29:3–35

Hothorn T (2019) Letter to the Editor response: Garcia et al. Biostatistics 20:546–548

Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Computat Graph Statist 15:651–674

Hothorn T, Kneib T, Bühlmann P (2014) Conditional transformation models. J Royal Statist Soc SerB 76:3–27

Hothorn T, Möst L, Bühlmann P (2018) Most likely transformations. Scandinav J Statist 45:110–134

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Ann Appl Statist 2:841–860

Jia B, Zeng D, Liao JJZ, Liu GF, Tan X, Diao G, Ibrahim JG (2022) Mixture survival trees for cancer risk classification. Lifetime Data Anal 28:356–379

Kalbfleisch JD, Prentice RL (eds) (2002) The statistical analysis of failure time data, 2nd edn. Wiley, New York

Klein JP, Andersen PK (2005) Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics 61:223–229

Kvamme H, Borgan Ø (2023) The Brier score under administrative censoring: problems and a solution. J Mach Learn Res 24:2

Landwehr N, Hall MA, Frank E (2005) Logistic model trees. Mach Learn 59:161–205

LeBlanc M, Crowley J (1992) Relative risk trees for censored survival data. Biometrics 48:411–425

Lee C, Zame W, Yoon J, van der Schaar M (2018) DeepHit: A deep learning approach to survival analysis with competing risks. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI Press, Palo Alto, pp 2314–2321

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Loh WY, Man M, Wang S (2019) Subgroups from regression trees with adjustment for prognostic effects and postselection inference. Statist Med 38:545–557

Mogensen UB, Gerds TA (2013) A random forest approach for competing risks based on pseudo-values. Statist Med 32:3102–3114

Molinaro AM, Dudoit S, van der Laan MJ (2004) Tree-based multivariate regression and density estimation with right-censored data. J Multivar Anal 90:154–177

Overgaard M, Parner ET, Pedersen J (2017) Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. Ann Statist 45:1988–2015

Puth MT, Tutz G, Heim N, Münster E, Schmid M, Berger M (2020) Tree-based modeling of time-varying coefficients in discrete time-to-event models. Lifetime Data Anal 26:545–572

Quinlan JR (1992) Learning with continuous classes. In: proceedings of the 5th Australian joint conference on artificial intelligence, World Scientific, Singapore, pp 343–348

R Core Team (2022) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, https://www.R-project.org/

Rahman MM, Matsuo K, Matsuzaki S, Purushotham S (2021) DeepPseudo: Pseudo value based deep learning models for competing risk analysis. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI Press, Palo Alto, pp 479–487

Scheike TH, Zhang MJ, Gerds TA (2008) Predicting cumulative incidence probability by direct binomial regression. Biometrika 95:205–220

Senkus E, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rutgers E, Zackrisson S, Cardoso F, Guidelines Committee ESMO (2015) Primary breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 26(Suppl. 5):v8–v30

Stensrud MJ, Hernán MA (2020) Why test for proportional hazards? J Am Med Associat 323:1401–1402

Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statist Med 30:1105–1117

van der Laan MJ, Robins JM (eds) (2003) Unified methods for censored longitudinal data and causality. Springer, New York

van der Ploeg T, Datema F, de Jong RB, Steyerberg EW (2014) Prediction of survival with alternative modeling techniques using pseudo values. PLoS One 9(6):e100234

von Minckwitz G, Untch M, Blohmer JU, Costa SD, Eidtmann H, Fasching PA, Gerber B, Eiermann W, Hilfrich J, Huober J, Jackisch C, Kaufmann M, Konecny GE, Denkert C, Nekljudova V, Mehta K, Loibl S (2012) Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin Oncol 30:1796–1804

Vatcheva KP, Lee ML, McCormick JB, Rahbar MH (2015) The effect of ignoring statistical interactions in regression analyses conducted in epidemiologic studies: an example with survival analysis using Cox proportional hazards regression model. Epidemiology (Sunnyvale, Calif) 6(1):216

Zeileis A, Hornik K (2007) Generalized M-fluctuation tests for parameter instability. Statist Neerland 61:488–508

Zeileis A, Hothorn T, Hornik K (2008) Model-based recursive partitioning. J Computat Graph Statist 17:492–514

Zhao L, Feng D (2020) Deep neural networks for survival analysis using pseudo values. IEEE J Biomed Health Inform 24:3308–3314

Zhao L, Murray S, Mariani LH, Ju W (2020) Incorporating longitudinal biomarkers for dynamic risk prediction in the era of big data: a pseudo-observation approach. Statist Med 39:3685–3699

3.3 Publication 3: Modeling the restricted mean survival time using pseudo-value random forests

Schenk A, Basten V, Schmid M. Modeling the restricted mean survival time using pseudo-value random forests. In: Statistics in Medicine 2025; 44 (5): e70031

Link to publication and supplementary information:
https://doi.org/10.1002/sim.70031

Implementations are available at:
https://www.imbie.uni-bonn.de/cloud/index.php/s/6gmJQmayFAMJZHk

Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Modeling the Restricted Mean Survival Time Using Pseudo-Value Random Forests

Alina Schenk[1] | Vanessa Basten[1,2] | Matthias Schmid[1]

[1]Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Bonn, Germany | [2]Department of Mathematics, Informatics and Technology, Koblenz University of Applied Sciences, Rhein-Ahr-Campus, Remagen, Germany

Correspondence: Alina Schenk (alina.schenk@imbie.uni-bonn.de)

Received: 16 August 2024 | **Revised:** 21 January 2025 | **Accepted:** 7 February 2025

Funding: The authors received no specific funding for this work.

Keywords: breast cancer survival | pseudo-values | random forest | restricted mean survival time | survival analysis | treatment contrast

ABSTRACT

The restricted mean survival time (RMST) has become a popular measure to summarize event times in longitudinal studies. Defined as the area under the survival function up to a time horizon $\tau > 0$, the RMST can be interpreted as the life expectancy within the time interval $[0, \tau]$. In addition to its straightforward interpretation, the RMST allows for the definition of valid estimands for the causal analysis of treatment contrasts in medical studies. In this work, we introduce a non-parametric approach to model the RMST conditional on a set of baseline variables (including, e.g., treatment variables and confounders). Our method is based on a direct modeling strategy for the RMST, using leave-one-out jackknife pseudo-values within a random forest regression framework. In this way, it can be employed to obtain precise estimates of both patient-specific RMST values and confounder-adjusted treatment contrasts. Since our method (termed "pseudo-value random forest", PVRF) is model-free, RMST estimates are not affected by restrictive assumptions like the proportional hazards assumption. Particularly, PVRF offers a high flexibility in detecting relevant covariate effects from higher-dimensional data, thereby expanding the range of existing pseudo-value modeling techniques for RMST estimation. We investigate the properties of our method using simulations and illustrate its use by an application to data from the SUCCESS-A breast cancer trial. Our numerical experiments demonstrate that PVRF yields accurate estimates of both patient-specific RMST values and RMST-based treatment contrasts.

1 | Introduction

During the past years, an increasing number of statisticians and applied researchers have advocated the use of the restricted mean survival time (RMST) to summarize event times in longitudinal studies [1, 2]. Defined as the area under the survival function within a time interval $[0, \tau]$, the RMST represents the expected event time between zero and the "time horizon" $\tau > 0$. In medical research, using the RMST as a summary measure offers the following specific advantages: (i) Its interpretation as the life expectancy between 0 and τ is straightforward and easily understood by both clinicians and patients [3], (ii) instead of a single time point (evaluated, e.g., by t-year survival probabilities in cancer research), the entire survival history up to τ is reflected by the RMST, (iii) in contrast to the hazard ratio (HR) derived from Cox regression, the RMST can be used for meaningful

Abbreviations: AFT, accelerated failure time; BMI, body mass index; CART, classification and regression trees; CI, confidence interval; DFS, disease-free survival; ER, estrogen receptor; GEE, generalized estimation equation; HER2, human epidermal growth factor receptor 2; HR, hazard ratio; IPC, inverse-probability-of-censoring; MSE, mean squared error; PFI, permutation feature importance; PH, proportional hazards; PR, progesterone receptor; PVRF, pseudo-value random forest; RCT, randomized controlled trial; RMSE, root mean squared error; RMST, restricted mean survival time; SD, standard deviation; WRSS, weighted residual sum of squares.

treatment comparisons even when the proportional hazards (PH) assumption is violated [1, 4], and (iv) the RMST can be used to define estimands for the causal interpretation of treatment and interventional effects [5]. As a result, the reporting of the RMST in medical studies has become increasingly prevalent [6–8].

In addition to the calculation of absolute RMST values, *differences* between group-wise RMST values have been suggested as a measure of treatment contrasts in longitudinal studies [6]. In medical research, treatment contrasts are often expressed and evaluated by the HR derived from a Cox PH model [1]. However, the interpretation of this type of HR is only valid if the PH assumption holds, assuming the HR to be constant over time. Thus, Stensrud and Hernán [8] recommended to supplement the reporting of HRs by summary measures directly derived from the survival function $S(t) = \mathrm{P}(T > t)$ (with T denoting the survival time). The RMST belongs to this class of measures, as it can be expressed as $\mu(\tau) = \mathrm{E}[\min(T, \tau)] = \int_0^\tau S(t)\, dt$ and therefore directly summarizes the survival function in $[0, \tau]$. Similarly, the RMST difference for two treatment groups A and B with survival functions $S_A(t)$ and $S_B(t)$, defined by $\mu_A(\tau) - \mu_B(\tau) = \int_0^\tau \big(S_A(t) - S_B(t)\big) dt$, can simply be interpreted as the difference in life expectancy or as a gain (or loss) in event-free survival time before τ [3].

This paper is concerned with the estimation of RMSTs conditional on covariates $\mu(\tau|X_i) = \int_0^\tau S(t|X_i)\, dt,\ i = 1, \ldots, n$, from a set of n independent individuals with possibly right-censored event times (in the following referred to as *individual RMSTs*). The covariate values are denoted by $X_i = (X_i^{(1)}, \ldots, X_i^{(p)})^\top \in \mathbb{R}^p$. For ease of notation and without loss of generality, we assume all treatment and interventional variables to be included in X_i. Our method is characterized by a non-parametric approach combining pseudo-value modeling [9] with random forest regression [10, 11]. Using the estimated individual RMSTs, we pursue two goals: (a) To incorporate the effects of a (possibly large and interacting) set of covariates in the estimation of the RMST, and (b) to quantify and assess accuracy of treatment effect estimation through RMST differences in observational longitudinal trials.

Standard approaches to estimate individual RMSTs $\mu(\tau|X_i)$ are the direct integration of group-wise Kaplan–Meier curves (leading to identical RMST estimates for individuals belonging to the same treatment group) and the integration of survival functions estimated through a parametric or semi-parametric time-to-event model with covariates X_i (e.g., a Cox PH model or an accelerated failure time (AFT) model [4, 12]). Using these standard approaches, the estimation of treatment effects through RMST differences is straightforward. Previous research on RMST differences also includes the work by Royston & Parmar [1], Tian et al. [13] and Huang & Kuan [14], who developed hypothesis tests for RMST differences derived by group-wise integration of Kaplan–Meier curves. Clearly, the covariate-free Kaplan–Meier approach is not recommended for use in non-randomized studies, as it ignores the effects of potential confounders on RMST differences. While integrating estimated survival functions derived from Cox PH or AFT models mitigates this problem, the validity of the resulting RMST estimates strongly depends on the correctness of the underlying model and/or distributional assumptions [15].

Instead of estimating individual RMSTs by integrating survival functions derived from time-to-event models, several authors have suggested to *directly* model the RMST [16–18]. In general, the idea of direct modeling approaches is to estimate unconditional individual RMSTs (without using any covariate information) and to subsequently fit a statistical model regressing these values to the covariates. Key advantages of directly modeling the RMST are less restrictive distributional assumptions as well as the straightforward interpretation of the model coefficients [1, 3, 6, 19].

In this paper, we pursue a direct approach for modeling RMST values and their differences. More specifically, the idea of our method is to derive unconditional RMST values from jackknife pseudo-values and to regress these values to the covariates using random forests. Classical pseudo-value regression for the RMST difference [20] is based on parametric models of the form

$$g[\mu(\tau|X_i)] = \alpha + \gamma^T X_i =: \eta_i, \tag{1}$$

with a monotonic link function g, an intercept α and covariate effects γ. Note that we suppress the dependency of α, γ and η_i on τ for ease of notation. Andersen et al. [9] and Andersen & Pohar Perme [20] suggested to estimate unconditional RMST values by leave-one-out jackknife pseudo-values $\hat{\theta}_i(\tau)$ defined as

$$\hat{\theta}_i(\tau) = n \cdot \int_0^\tau \hat{S}_{\mathrm{KM}}(t)\, dt \;-\; (n-1) \cdot \int_0^\tau \hat{S}_{\mathrm{KM}}^{-i}(t)\, dt\,, \quad i = 1, \ldots, n\,, \tag{2}$$

where $\hat{S}_{\mathrm{KM}}(t)$ denotes the Kaplan–Meier estimate evaluated at t calculated on the complete data set and $\hat{S}_{\mathrm{KM}}^{-i}(t)$ denotes the respective Kaplan–Meier estimate calculated on the data set without individual i.

The coefficients in (1) can be estimated by a generalized estimation equation (GEE) approach, with g being the identity or the log link [9]. However, while the GEE approach yields consistent estimates ($n \to \infty$) under the assumption of random censoring [21, 22], its flexibility is limited by the restrictive specification of the main-effects predictor η_i in (2). Although more flexible effect terms (representing, e.g., interaction terms or non-linear main effects) could be included in (1), this approach is not commonly used in practice. Often, this is due to the fact that pre-specifying an extended version of η_i requires detailed knowledge on the, usually unknown, dependency structure between the pseudo-value outcome and the covariates. Further, the GEE approach does, in its basic form, neither incorporate any mechanism for data-driven variable selection nor perform any other sort of regularization to reduce redundant or irrelevant information.

To address these issues, and to achieve the goals stated in (a) and (b), we propose to replace the GEE approach by a random forest regression [10]. This regression model, which uses the pseudo-values $\hat{\theta}_i(\tau)$ as continuous outcome and which will be termed "pseudo-value random forest" (PVRF) in the following, allows for a data-driven selection of covariates and their interaction effects. In this way, the need to pre-specify η_i is eliminated, making PVRF a convenient method for applications involving a large number of covariates compared to the number of

individuals (for instance, in medium-sized observational studies containing many potential confounders). By applying a g-computation formula [23, 24] to the estimated RMST values, the PVRF method further allows for the direct estimation and causal interpretation of RMST differences. To additionally facilitate interpretability of the covariate effects, we propose to use methods for interpretable machine learning, as described in Molnar [25].

The remainder of this paper is organized as follows: In Section 2, we define relevant terms and provide a detailed description of the PVRF method. Sections 3 and 4 contain the results of a simulation study investigating the properties of the PVRF method and comparing the proposed approach to established methods for RMST estimation. Section 5 presents an application of the PVRF method to data from the SUCCESS-A study, a randomized phase III trial investigating the effects of two treatment regimens on the disease-free survival of patients with early breast cancer [26]. Section 6 concludes with the main findings and a brief overview of related approaches.

2 | Methods

We consider a set of n independent individuals subject to right-censoring with covariate values $X_i = (X_i^{(1)}, \ldots, X_i^{(p)})$, $i = 1, \ldots, n$, measured at baseline. The individual survival time and censoring time are denoted by T_i and C_i, respectively. The observed survival time is denoted by $\tilde{T}_i = \min(T_i, C_i)$, and the status variable $\delta_i = \mathbb{1}_{\{C_i > T_i\}}$ indicates whether the i-th individual is censored ($\delta_i = 0$) or whether the event of interest has been observed ($\delta_i = 1$). Following Graw et al. [21], we assume that the censoring times are independent of both the covariates and the event times.

2.1 | Estimation and Modeling of the RMST via Pseudo-Values

When using the RMST, defined as $\mu(\tau) = \mathrm{E}[\min(T, \tau)] = \int_0^\tau S(t)\, dt$, as dependent variable in a regression model, the outcome values are given by $\mu_i(\tau) = \min(T_i, \tau), i = 1, \ldots, n$. By definition, these values depend on the survival times T_i, and, due to censoring, cannot be observed for all individuals. Pseudo-value regression overcomes this problem by replacing the partly incompletely observed outcome values with continuous (real-valued) pseudo-values $\hat{\theta}_i(\tau)$ that can be computed for both censored and uncensored individuals. For the RMST, the i-th pseudo-value at a time horizon τ is given by the right-hand side of Equation (2). The pseudo-values can subsequently be used as a (completely observed) imputation for the outcome variable $\mu_i(\tau)$ in the RMST regression model, facilitating the application of conventional modeling techniques like linear regression or trees [27]. It can be shown that the replacement of $\mu_i(\tau)$ with $\hat{\theta}_i(\tau)$ enables the consistent estimation of covariate effects on the RMST (see Overgaard et al. [22] for details and regularity assumptions). As seen from Figure 1, the characteristics of pseudo-values for the RMST depend on the observed time $\tilde{T}_i$, the censoring proportion in the data set and the time horizon τ. In general, it appears hard to approximate the empirical distribution of the pseudo-values by a parametric distribution.

As outlined in Section 1, the standard approach to pseudo-value regression for the RMST is to use the unconditional pseudo-values $\hat{\theta}_i(\tau)$ as outcome variable in a GEE model (see Equation (1)). The estimated coefficients of this model can be interpreted as direct effects on the RMST if g is the identity link, or on the logarithm of the RMST if g is the log link. For details on GEE estimation, we refer to Graw et al. [21].

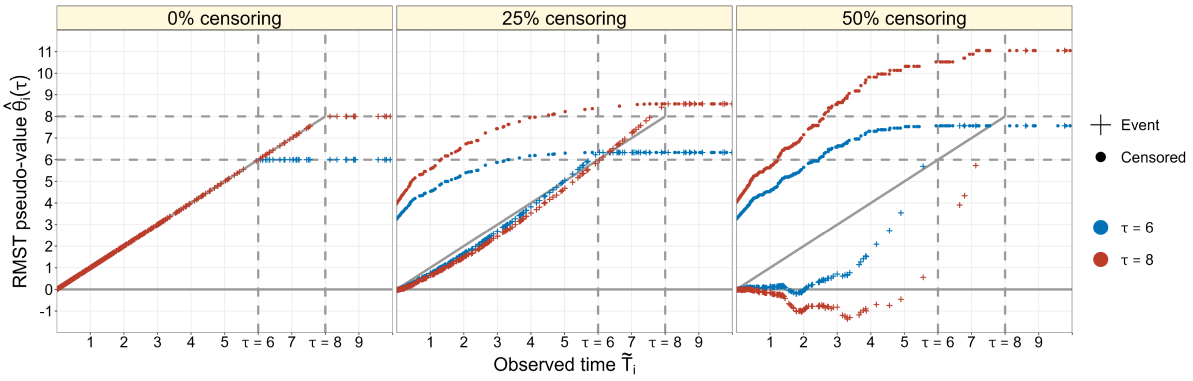


FIGURE 1 | Illustration of pseudo-values, as derived from a synthetic data set with $n = 500$. The dashed vertical lines indicate the time horizons $\tau \in \{6, 8\}$. Pseudo-values of censored and uncensored individuals are represented by dots and plus symbols, respectively (blue: $\tau = 6$, red: $\tau = 8$). In the data with no censoring (left panel), it holds that $\hat{\theta}_i(\tau) = \tilde{T}_i = T_i$ if $\tilde{T}_i < \tau$ and $\hat{\theta}_i(\tau) = \tau$ if $\tilde{T}_i \geq \tau$. For censored individuals (dots in the middle and right panels), it is observed that $\hat{\theta}_i(\tau) > \tilde{T}_i$ for $\tilde{T}_i < \tau$, irrespective of the choice of τ and the censoring proportion. For individuals with an observed event in data sets with censoring, there is no consistent pattern regarding the dependency of pseudo-values on $\tilde{T}_i$: At a lower censoring proportion, pseudo-values of individuals with an observed event closely resemble the observed times (plus symbols in middle panel). In contrast, at a higher censoring proportion, pseudo-values of individuals with an observed event are mostly lower than the observed event time and can even become negative (plus symbols in the right panel). For $\tilde{T}_i \geq \tau$, it holds that $\hat{\theta}_i(\tau) \geq \tau$ for all individuals. Consequently, there is no difference between individuals who were censored after time τ and those who were observed to experience an event after τ. Figure adapted from Andersen & Pohar Perme [20].

Despite the popularity of the GEE approach, it is easily seen from (1) that the estimated RMST values are constrained to a rather restrictive linear combination of the covariate effects. Particularly, (1) does not include any interaction terms. While these terms could be pre-specified in η_i, it is well known that the number of interactions grows exponentially with the number of covariates. This implies numerous coefficients to be estimated and a high variance of the GEE estimator even in cases with a moderate number of covariates. If there is no expert knowledge available to pre-select a suitable (small) number of interaction terms, data-driven variable selection techniques (such as forward or backward selection) could be applied. However, these algorithms usually show a high variability and are not recommended for the selection of interaction terms.

2.2 | Random Forest Regression

To address the issues described in Section 2.1, we propose to model individual RMSTs by using the pseudo-values $\hat{\theta}_i(\tau)$ as continuous outcome variable in a random forest regression model [10, 11]. Random forest regression is characterized by averaging estimates of multiple regression trees trained on different random subsets of the data ("ensemble" of regression trees). In this way, overfitting is avoided, and both interactions and non-linear covariate effects are captured by the model [10].

The general idea of building a regression tree is to derive local estimates of the outcome variable by partitioning the covariate space into a set of mutually exclusive subspaces [28–30]. Beginning with the *root node* containing all individuals $i = 1, \ldots, n$, the idea is to successively evaluate a split criterion and to split individuals into two mutually exclusive sets termed *daughter nodes*. Each daughter node $R_m \subseteq \{1, \ldots, n\}$ is further split into two daughter nodes $R_{m_1} \subset R_m$ and $R_{m_2} \subset R_m$ with $R_{m_1} \cap R_{m_2} = \emptyset$, and splitting is continued until some stopping criterion applies (see Appendix B). In each node R_m, splitting is done by selecting a split variable $X^{(j^*)}$, $j^* \in \{1, \ldots, p\}$, and a corresponding split rule $\mathcal{R}_{mj^*}$ that optimize a pre-defined split criterion (e.g., the mean squared error, see Hothorn et al. [29] and Greenwell [30] for details on split rules). The split criterion is evaluated on the data of the individuals in the respective node R_m. Nodes that are not further split into two daughter nodes because the stopping criterion applies are referred to as *leaf nodes*. For calculating the estimated RMST value of a single individual, the associated leaf node is determined by using the individual's covariate values and by successively applying the split rules from the root node to the leaf node. Afterwards, the RMST value is estimated by averaging the observed pseudo-values $\hat{\theta}_i(\tau)$ in the leaf node.

Random forest regression is characterized by growing large ensembles of regression trees. In this paper, we will use 500 trees unless stated otherwise. Furthermore, we follow the recommendation by de Bin et al. [31] and grow our tree ensemble on subsamples of the complete data without replacement. Thus, each tree in the forest is grown on a different subset of the data, leading to different split rules and different RMST estimates in the leaf nodes. Additionally, only a random subset of the covariates is considered for splitting the nodes of the regression trees. We determine the size of this subset (termed "mtry") using five-fold cross-validation, see Appendix B. The final RMST estimate for individual i is obtained by dropping the covariate values X_i down to the leaf nodes of the 500 trees and by averaging the 500 tree estimates.

In the literature, there exist multiple tree building algorithms that vary in the procedure to select the split variables and the corresponding split rules. In this work, we consider two different tree-building algorithms that will be described briefly in Sections 2.2.1 and 2.2.2: (i) Classification and regression trees (CART) [28] and (ii) conditional inference trees [29]. Correspondingly, the resulting forests will be referred to as *CART random forest* and *conditional random forest*.

2.2.1 | CART Random Forest

In each node R_m, the CART algorithm selects the split variable $X^{(j^*)}$ and the corresponding split rule $\mathcal{R}_{mj^*}$ by minimizing

$$\text{MSE}_{\mathcal{R}_{mj}} = \sum_{i \in R_{m_1}} \left(\hat{\theta}_i(\tau) - \bar{c}_1\right)^2 + \sum_{i \in R_{m_2}} \left(\hat{\theta}_i(\tau) - \bar{c}_2\right)^2 \qquad (3)$$

over $\mathcal{R}_{mj}$, where $\bar{c}_1$ and $\bar{c}_2$ are the averaged pseudo-values in the daughter nodes R_{m_1} and R_{m_2}, respectively. Consequently, the split variable and the split rule minimizing the mean squared errors of the pseudo-values $\hat{\theta}_i(\tau)$ in the daughter nodes are selected jointly in one optimization step. In practice, this leads the CART algorithm to favor split variables with many possible splits, implying that the algorithm is biased towards the selection of covariates with many possible splits (e.g., continuous covariates) [29].

2.2.2 | Conditional Random Forest

Unlike the CART algorithm, conditional inference trees [29] follow a two-step process in each node, selecting the optimal split variable by a set of statistical hypothesis tests *before* determining the corresponding split rule. In this way, a selection bias towards covariates with many possible splits is avoided. More specifically, in the first step, the null hypotheses of independence between the covariates and the outcome values $\hat{\theta}_i(\tau)$ are tested by evaluating a set of generalized correlation coefficients ρ_j, $j = 1, \ldots, p$ (measuring the pairwise associations between the outcome and the covariates), and by computing a permutation-based p-value for each covariate using the conditional distributions of transformed versions of ρ_j under the null. Finally, the covariate with minimum p-value in the permutation tests is selected as a split variable. Since the p-values do not depend on the scales of the covariates, the selection procedure does not show any systematic preference towards covariates with many possible splits. For details on the definition of ρ_j and the test procedure, we refer to Hothorn et al. [29].

The second step is to derive the split rule associated with the selected split variable $X^{(j^*)}$. Analog to the CART algorithm, each possible split rule leads to two possible daughter nodes R_{m_1} and R_{m_2}. To determine the optimal split rule $\mathcal{R}_{mj^*}$, the idea is to maximize a criterion that is constructed in the same way as the generalized correlation coefficients ρ_j, this time measuring the association between the pseudo-values and node membership. Details on the selection procedure are given in Hothorn et al. [29].

2.3 | Evaluating RMST Differences

In medical research, a common aim is to compare subgroups of the population with regard to their survival behavior. Usually, these subgroups are defined by an intervention (e.g., treatment vs. control, see Section 5) or by the presence of a risk factor. Following Royston & Parmar [4], Uno et al. [6] and Dehbi et al. [7], we quantify differences in the survival behavior of population subgroups (in the following termed "treatment contrasts") using differences in RMST values. In randomized controlled trials (RCTs), which usually allow for ignoring all covariates except the intervention due to the randomization procedure, treatment contrasts can simply be estimated by the differences of the average RMST values in the relevant groups. When additional covariates have to be taken into account, particularly in non-randomized studies where the covariates usually take the roles of confounders, we propose to apply g-computation to estimate treatment contrasts [23, 24]. More specifically, denoting the treatment variable of individual i by $X_i^{(\text{trt})}$ and the respective confounders by $X_i^{(-\text{trt})}$, we propose to calculate RMST differences as

$$\hat{\Delta}_i(\tau) = \hat{\mu}(\tau | X_i^{(-\text{trt})}, X_i^{(\text{trt})} = A) - \hat{\mu}(\tau | X_i^{(-\text{trt})}, X_i^{(\text{trt})} = B), \; i = 1, \ldots, n, \tag{4}$$

where $\hat{\mu}$ denotes the RMST estimate obtained from the random forest model (see Hu et al. [32] for alternative ways to define and estimate survival treatment effects). Based on the individual RMST differences, the average treatment effect (= treatment contrast) is estimated by

$$\hat{\Delta}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}_i(\tau) = \frac{1}{n}\sum_{i=1}^{n} \left[\hat{\mu}(\tau | X_i^{(-\text{trt})}, X_i^{(\text{trt})} = A) - \hat{\mu}(\tau | X_i^{(-\text{trt})}, X_i^{(\text{trt})} = B)\right], \tag{5}$$

2.4 | Pseudo-Value Random Forest

Summarizing Sections 2.1 to 2.3, we define our proposed method (termed "pseudo-value random forest", PVRF) by the following steps:

1. Calculate pseudo-values $\hat{\theta}_i(\tau), \; i = 1, \ldots, n$, for the RMST (Equation (2)).
2. Grow a random forest using either the CART algorithm (Section 2.2.1) or conditional inference trees (Section 2.2.2).
3. Estimate RMST values conditional on covariates $\hat{\mu}(\tau | X_i), \; i = 1, \ldots, n$, from the fitted random forest.
4. Depending on the research question,
   a. proceed analyzing estimated RMST values using interpretable machine learning techniques (see Section 5).
   b. estimate treatment contrasts $\hat{\Delta}(\tau)$ from individual RMST differences $\hat{\Delta}_i(\tau), \; i = 1, \ldots, n$ (Equation (5)).

3 | Experiments

To investigate the performance of PVRF, we carried out a comprehensive simulation study in R (version 4.1.2 [33]). The data-generating process was based on a time-to-event model with an additive combination of main and interaction effects. We analyzed the ability of PVRF to estimate RMSTs and RMST differences conditional on covariates between treatment groups in the absence and presence of two-way interactions. To this end, we considered scenarios with time-constant and time-varying treatment effects. The simulation study was based on 100 Monte Carlo replications. In each replication, we generated a data set of size $n = 1000$.

Survival times T_i, $i = 1, \ldots, n$, were generated from a Weibull model with scale parameter $\lambda > 0$, shape parameter $\nu > 0$ and hazard function $h(t|X_i) = \lambda \cdot \exp\big(\eta_i(t)\big) \cdot \nu \cdot t^{\nu-1}$, where $\eta_i(t)$ is the (possibly time-dependent) linear predictor of individual i (depending on X_i, see Equation (6)). The cumulative hazard function was given by $H(t|X_i) = \lambda \cdot \exp\big(\eta_i(t)\big) \cdot t^{\nu}$. The censoring times were generated independently of the survival times, using the same Weibull model with $\eta_i(t) = 0$. The parameters λ and ν were adjusted such that the data-generating process yielded the desired censoring proportions.

Overall, we examined four scenarios, each differing in the calculation of $\eta_i(t)$. Each scenario was characterized by five continuous covariates (denoted by $X_i^{(j)}$, $j = 1, \ldots, 5$) and five dichotomous covariates (denoted by $X_i^{(j)}$, $j = 6, \ldots, 10$). The continuous covariates followed a multivariate normal distribution with zero mean and a covariance matrix as given in Table A1 in Section A. Dichotomous covariates were independent and followed Bernoulli distributions with probability 0.5 each. In addition, we considered a dichotomous treatment variable $X_i^{(\text{trt})}$ (treatment A vs. B, Bernoulli distributed with probability 0.5). The scenarios further differed in the structure of the interactions between the covariates and the strength of the treatment effects. We considered predictors of the form

$$\begin{aligned}\eta_i(t) = &\sum_{j=1}^{10} \delta_j X_i^{(j)} + \sum_{\substack{l \in \{1,\ldots,5\} \\ m \in \{1,\ldots,5\}}} \psi_{lm} X_i^{(l)} X_i^{(m)} \\ &+ \sum_{\substack{r \in \{1,\ldots,5\} \\ s \in \{6,\ldots,10\}}} \varphi_{rs} X_i^{(r)} X_i^{(s)} + \vartheta_{\text{trt}}(t) \mathbb{1}_{\{X_i^{(\text{trt})} = B\}},\end{aligned} \tag{6}$$

where δ_j, $j = 1, \ldots, 10$, denote main effects of the continuous and the dichotomous covariates, ψ_{lm}, $l, m \in \{1, \ldots, 5\}$, represent the interaction effects between the continuous covariates, φ_{rs}, $r \in \{1, \ldots, 5\}$, $s \in \{6, \ldots, 10\}$, represent the interaction effects between the continuous and the dichotomous covariates, and $\vartheta_{\text{trt}}(t)$ denotes the (possibly time-varying) treatment effect. All main and interaction effects were sampled from a continuous uniform distribution on $[-1, 1]$; they were generated independently of each other and were the same in all Monte Carlo replications. Furthermore, we added five independent standard normally distributed noise variables to the covariate set. These were independent of the other covariates and did not affect the predictor $\eta_i(t)$. In Scenarios 1 and 2, the treatment effect was time-constant, whereas in Scenarios 3 and 4, the treatment effect changed at the transition time t_0, resulting in crossing survival curves (Figure 3). Scenarios 1 and 3 included only main effects, whereas Scenarios 2 and 4 additionally included interaction effects. Table 1 provides an overview of the four scenarios, and Figure 3 presents the group-wise Kaplan–Meier curves for each scenario.

TABLE 1 | Overview of the four scenarios used in the simulation study, each characterized by five continuous covariates, five dichotomous covariates, eight interaction effects (Scenarios 2 and 4), and a time-constant (Scenarios 1 and 2) or time-varying (Scenarios 3 and 4) treatment effect. Interaction effects not contained in the fourth column were set to zero.

Scenario	Effects of continuous covariates	Effects of dichotomous covariates	Interaction effects	Treatment effect
1	$\delta_j \sim \mathrm{U}(-1,1)$	$\delta_j \sim \mathrm{U}(-1,1)$	$\psi_{lm} = 0\ \forall l,m$	$\vartheta_{\mathrm{trt}}(t) = -2$
	$j = 1,\ldots,5$	$j = 6,\ldots,10$	$\varphi_{rs} = 0 \forall r,s$	
2	$\delta_j \sim \mathrm{U}(-1,1)$	$\delta_j \sim \mathrm{U}(-1,1)$	$\psi_{13},\psi_{14},\psi_{23},\psi_{25},\psi_{45} \sim \mathrm{U}(-1,1)$	$\vartheta_{\mathrm{trt}}(t) = -2$
	$j = 1,\ldots,5$	$j = 6,\ldots,10$	$\varphi_{17},\varphi_{28},\varphi_{39} \sim \mathrm{U}(-1,1)$	
3	$\delta_j \sim \mathrm{U}(-1,1)$	$\delta_j \sim \mathrm{U}(-1,1)$	$\psi_{lm} = 0\ \forall l,m$	$\vartheta_{\mathrm{trt}}(t) = \begin{cases} -2 & t \le t_0 \\ 2 & t > t_0 \end{cases}$
	$j = 1,\ldots,5$	$j = 6,\ldots,10$	$\varphi_{rs} = 0 \forall r,s$	
4	$\delta_j \sim \mathrm{U}(-1,1)$	$\delta_j \sim \mathrm{U}(-1,1)$	$\psi_{13},\psi_{14},\psi_{23},\psi_{25},\psi_{45} \sim \mathrm{U}(-1,1)$	$\vartheta_{\mathrm{trt}}(t) = \begin{cases} -2 & t \le t_0 \\ 2 & t > t_0 \end{cases}$
	$j = 1,\ldots,5$	$j = 6,\ldots,10$	$\varphi_{17},\varphi_{28},\varphi_{39} \sim \mathrm{U}(-1,1)$	

In each of the four scenarios, we considered three different censoring proportions (25%, 50%, and 75%) and five different values of the time horizon τ. The latter were determined by the 50%, 60%, 70%, 80%, and 90% quantiles of the observed times $\tilde{T}_i$, $i = 1,\ldots,n$, denoted by $q_{50\%}$, $q_{60\%}$, $q_{70\%}$, $q_{80\%}$, and $q_{90\%}$, respectively. The values of τ, which were held fixed across the simulation runs, are given in Table A2 in Section A. The transition time t_0 in Scenarios 3 and 4 was set to $q_{70\%}$. In total, each scenario examined 15 combinations of censoring proportions and time horizons τ. For the values of the coefficients δ_j, ψ_{lm} and φ_{rs}, we refer to the attached R code (see Appendix B).

To evaluate performance of RMST estimates, we considered the mean squared error defined by

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}(\tau|X_i) - \mu(\tau|X_i)\right)^2, \tag{7}$$

where $\hat{\mu}(\tau|X_i)$ and $\mu(\tau|X_i)$ denote the estimated and the theoretical RMSTs, respectively, of individual i at time horizon τ. The root mean squared error (RMSE) is defined as the square root of (7). The theoretical RMST in (7) is derived as

$$\begin{aligned}\mu(\tau|X_i) &= \int_0^\tau S(t|X_i)\,dt = \int_0^\tau \exp(-H(t|X_i))\,dt \\ &= \begin{cases} \int_0^\tau \exp(-H_1(t|X_i))\,dt\,, & \tau \le t_0\,, \\ \int_0^{t_0} \exp(-H_1(t|X_i))\,dt + \int_{t_0}^\tau \exp(-H_1(t_0|X_i) & \\ \quad -H_2(t|X_i) + H_2(t_0|X_i))\,dt\,, & \tau > t_0\,, \end{cases}\end{aligned} \tag{8}$$

where $H_1(t|X_i)$ and $H_2(t|X_i)$ are the cumulative hazard functions before and after the transition point t_0, respectively. Note that in Scenarios 1 and 2, the hazards are constant over time and thus $H_1(t|X_i) = H_2(t|X_i)$, resulting in $\mu(\tau|X_i) = \int_0^\tau \exp(-H_1(t|X_i))dt$ for both $\tau \le t_0$ and $\tau > t_0$. Analogously, we evaluated the accuracy of treatment effect estimates by calculating the mean squared error of the treatment effect, defined as

$$\mathrm{MSE}_\Delta = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Delta}_i(\tau) - \Delta_i(\tau)\right)^2, \tag{9}$$

where $\hat{\Delta}_i(\tau)$ and $\Delta_i(\tau)$ denote the estimated and the theoretical individual treatment effects, respectively (see Section 2.3).

In addition to evaluating the estimation accuracy of the CART and conditional random forest approaches, we compared our method to alternative modeling approaches. These were (i) a GEE pseudo-value model with identity link (*GEE*), (ii) a GEE pseudo-value model with log link (*GEE (log)*), (iii) a Cox PH model (*Cox*), (iv) a parametric AFT model (based on log-transformed survival times and assuming normally distributed errors [34], *Lognormal*), and (v) a correctly specified Cox PH model (*Reference*). For the modeling approaches (i)–(iv), we specified the main effects of all continuous and dichotomous covariates (including the noise variables) but did not consider any interaction terms. The *Reference* model was specified such that it corresponded to the true data-generating process, incorporating the *informative* (= non-zero) main and interaction effects only (see Table 1). The *Reference* model also accounted for the time-dependent treatment effect in Scenarios 3 and 4. This was accomplished by specifying a time-varying stratification variable that enabled the Cox model to estimate a time-dependent treatment effect. Consequently, *Reference* served as a lower benchmark in the RMSE and RMSE_Δ comparisons. For the *Cox*, *Lognormal* and *Reference* models, which do not directly model the RMST, estimates of the RMST were derived through the integration of the estimated survival function. Further details on the implementation of the methods are given in Appendix B.

For the main-effects-only Scenario 1, we expect the *Cox* and *Lognormal* models (both assuming a main effects structure) to show a better performance than the PVRF method. In contrast, we anticipate that the CART random forest and the conditional random forest approaches will outperform the *Cox*, *Lognormal*, *GEE*, and *GEE (log)* models in the scenarios with non-zero interaction terms (Scenarios 2 and 4). Additionally, due to the time-varying treatment effect, we expect the pseudo-value methods to outperform the *Cox* model in Scenarios 3 and 4. Generally, we expect both the RMSE and RMSE_Δ values to increase with τ, since the RMST also rises with τ.

To evaluate the performance of the PVRF method and its comparators in a misspecified scenario, we further conducted a modified version of the previously described simulation study. In this additional study, data were simulated in the same way as before (Table 1), but one informative continuous covariate ($X^{(2)}$) and one informative dichotomous covariate ($X^{(7)}$) were excluded from the set of candidate covariates used for model fitting. As a result, all methods were provided with a reduced set of covariates. In this study, we expect both the RMSE and RMSE_Δ values to increase relative to the study using the full candidate covariate set. Additionally, we anticipate that the *Cox* and *Lognormal* models will be less advantageous in Scenario 1 compared to the PVRF method. Similar to the simulation study above, we expect the PVRF method to be superior to all comparators in Scenarios 2, 3, and 4.

4 | Results

Figure 2 summarizes the simulation results of the four scenarios at a censoring proportion of 50%. In the first scenario (main effects only, time-constant treatment effect), both the average RMSE and the average RMSE_Δ increase with τ, as expected. This is true for all considered models. Notably, there is a clear difference in terms of RMSE between the standard modeling techniques (*Cox* and *Lognormal*) and the pseudo-value methods (*GEE*, *GEE (log)*, CART random forest and conditional random forest), with the best performing model being the *Cox* model followed by the *Lognormal* model. This result can be explained by the fact that the *Cox* model matches the data-generating mechanism in this scenario (except for the noise variables). Among the pseudo-value regression methods, the conditional random forest demonstrates superior performance for $\tau \leq q_{60\%}$ followed by *GEE*, *GEE (log)* and CART random forest. However, this is no longer true when $\tau > q_{60\%}$. In terms of the RMSE for the treatment effect (RMSE_Δ), the *Cox* model demonstrates the best performance, in line with our expectations, followed by the *Lognormal* model. Among the pseudo-value methods, the conditional random forest performs best with regard to treatment effect estimation, followed by the *GEE* model. The application of the log link in the GEE approach (*GEE (log)*) appears to have a negative effect on both performance measures (first column of Figure 2). Notably, the CART random forest shows inferior performance compared to the conditional random forest and to all other comparators.

In Scenario 2 (non-zero interaction effects, time-constant treatment effect), the average RMSE_Δ increases with τ, similar to Scenario 1. The tree-based pseudo-value methods, particularly the conditional random forest, perform best in terms of RMSE, having a slight advantage over the CART random forest. This result demonstrates the ability of tree-based methods to identify and model interactions between the covariates. All other methods perform similarly in this scenario. Regarding treatment effect estimation, the conditional random forest performs best, having slight advantages over the standard *Cox* and *Lognormal* modeling techniques, as well as over the CART random forest. Although the *Cox* and *Lognormal* models do not perform well in terms of RMSE, their performance regarding treatment effect estimation is comparable to the respective performance of the tree-based methods. On the other hand, the GEE with log link shows a poor performance in the estimation of the treatment effect, with estimates getting worse as τ increases (second column of Figure 2).

In Scenario 3 (main effects only, time-dependent treatment effect), the average RMSE values of the standard modeling

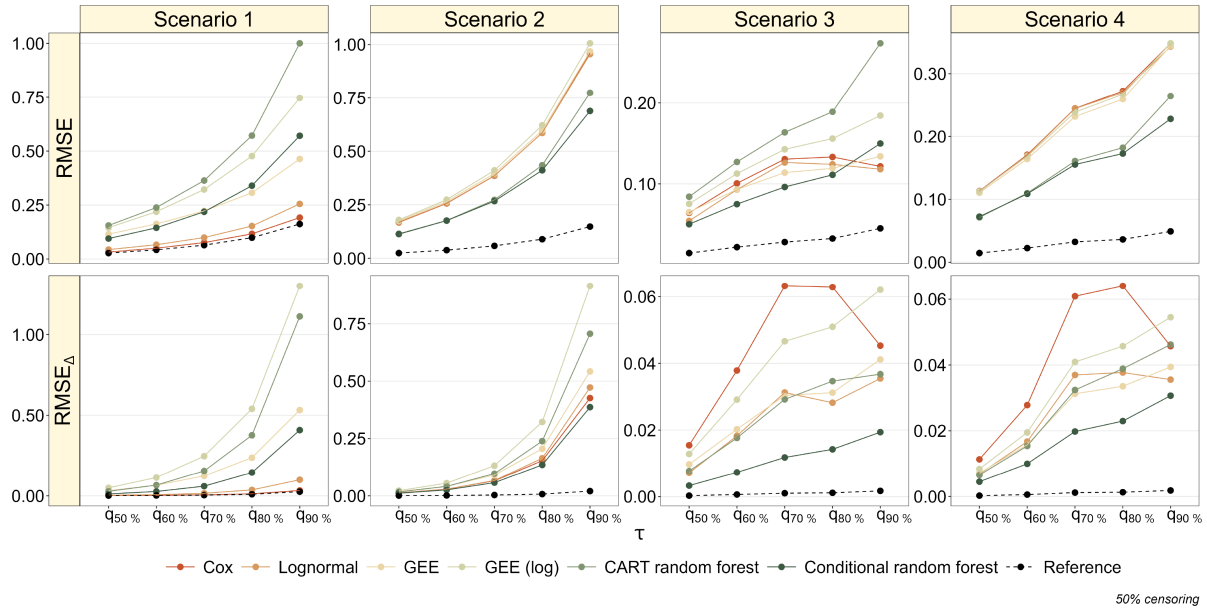


FIGURE 2 | Results of the simulation study (50% censoring). The upper panels present the average RMSE (7), as obtained from the RMST estimates at different values of τ. The lower panels present the average RMSE for the treatment effect (9). The dashed black lines refer to the correctly specified Cox PH model (*Reference*).

techniques *Cox* and *Lognormal* do not monotonically increase with τ, as in Scenarios 1 and 2. Instead, there is a turning point in the RMSE values at $\tau = t_0 = q_{70\%}$. This can be explained as follows: As seen from Figure 3, the *Cox* model assumes a time-constant treatment effect, implying that the effect of treatment B is underestimated when $t \leq t_0 = q_{70\%}$. This in turn leads to strongly biased RMST estimates. On the other hand, the *Cox* model overestimates the effect of treatment B when $t > t_0 = q_{70\%}$. Consequently, as RMST estimates are derived by the area under the survival curve up to τ, the part of the area under the survival curve that is not included in the RMST estimates for $t \leq t_0 = q_{70\%}$ is compensated by the excess area under the estimated survival curve for $t > t_0 = q_{70\%}$. As a result, the *Cox* model yields decreasing RMSE values for $\tau > t_0 = q_{70\%}$. For the *Lognormal* model, analog observations can be made. In contrast, the pseudo-value methods exhibit increasing RMSE values with rising τ, as expected. The conditional random forest demonstrates superior performance compared to the *Cox* and *Lognormal* models with respect to the RMSE for $\tau \leq t_0$. On the other hand, as the RMSE values obtained from the *Cox* and *Lognormal* models decrease for $\tau > t_0$, these methods perform better than the conditional random forest at $\tau = q_{90\%}$. Notably, the CART random forest shows inferior performance compared to all other methods, which is likely due to its selection bias towards (possibly non-informative) continuous covariates. Regarding treatment effect estimation, the RMSE_Δ values obtained from the pseudo-value methods increase with τ. In contrast, the RMSE_Δ values obtained from the *Cox* model increase for $\tau \leq t_0 = q_{70\%}$ but decrease for $\tau > t_0 = q_{70\%}$. Again, this is due to the underestimation (overestimation) of the effect of treatment B for $t \leq t_0 = q_{70\%}$ ($t > t_0 = q_{70\%}$). The conditional random forest consistently performs best, regardless of the value of τ, confirming its ability to capture the time-dependent treatment effect (third column of Figure 2).

In the presence of interaction effects, as in Scenario 4 (non-zero interaction effects, time-dependent treatment effect), there is a clear advantage of the tree-based methods (CART and conditional random forest) with respect to the RMSE. Regarding treatment effect estimation, the conditional random forest performs consistently best with respect to RMSE_Δ across all time horizons τ (fourth column of Figure 2). These results highlight the ability of the conditional random forest to model interaction effects and to capture time-dependent treatment effects simultaneously. As seen from Figure 2, the time-dependent treatment effect influences the trend of the RMSE_Δ values of the *Cox* and *Lognormal* models, similar to Scenario 3, but not the trend of the respective RMSE values. The results obtained for 25% and 75% censoring are similar to the results shown in Figure 2. They are presented in Figures C1 and C2 in Appendix C.

While our primary focus was on evaluating the performance of the PVRF method in estimating RMST values conditional on covariates, we also explored the generalizability of our findings to unseen data, that is, data that were not used for model fitting. The RMSE and RMSE_Δ values derived on unseen data can be found in Figures D1–D3 in Appendix D. In summary, we observed similar results as for the RMSE and RMSE_Δ values obtained from the data used for model fitting, except for the CART random forest, which showed an improved performance.

Figure 4 presents a comparison of the results from Scenario 1 using the full and reduced candidate covariate sets at 50% censoring. As expected, the average RMSE and RMSE_Δ values increase when the reduced candidate covariate set, excluding $X^{(2)}$ and $X^{(7)}$, is used for the modeling approaches. Furthermore, the differences between the *Cox*, *Lognormal*, and PVRF methods decrease, suggesting that the models exhibit a more similar performance (in terms of both RMSE and RMSE_Δ) than when the full candidate covariate set is used. Put differently, the advantages of the *Cox* and *Lognormal* models are way less pronounced when $X^{(2)}$ and $X^{(7)}$ are excluded from the set of candidate covariates, indicating a higher stability of the conditional random forest. In Scenarios 2, 3 and 4, the conditional random forest still outperforms all other methods when $X^{(2)}$ and $X^{(7)}$ are removed from the candidate covariate set (see Figure E1 in Appendix E).

In summary, when considering scenarios with interactions and/or time-varying treatment effects, we find the conditional

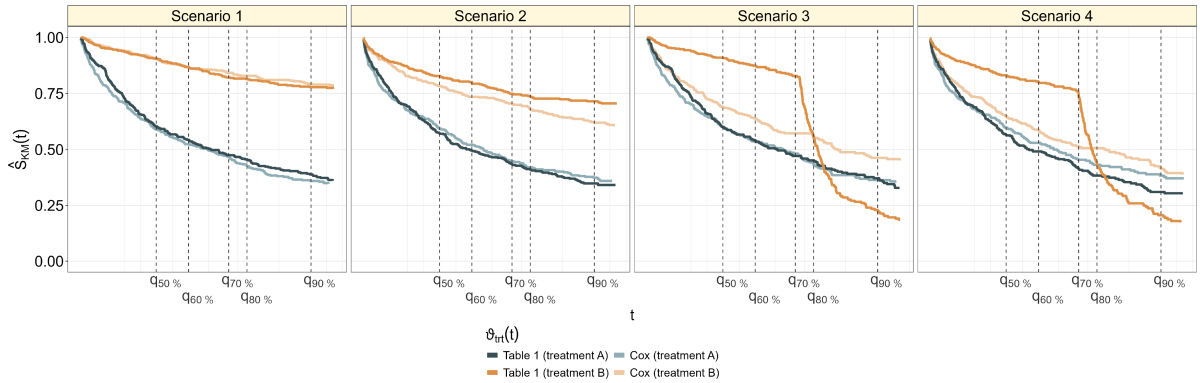


FIGURE 3 | Results of the simulation study (50% censoring). The dark lines depict the Kaplan–Meier curves in the two treatment groups, as obtained from $n = 1000$ individuals with data generated according to Table 1 (including the true treatment effects $\vartheta_{trt}(t)$). The bright lines depict the Kaplan–Meier curves derived from data generated according to Table 1 but including the time-constant average treatment effect estimated by the *Cox* method instead of the true treatment effect (Scenario 1: $\hat{\vartheta}_{trt}^{Cox}(t) = -2.10$, Scenario 2: $\hat{\vartheta}_{trt}^{Cox}(t) = -1.26$, Scenario 3: $\hat{\vartheta}_{trt}^{Cox}(t) = -0.38$, Scenario 4: $\hat{\vartheta}_{trt}^{Cox}(t) = -0.33$).

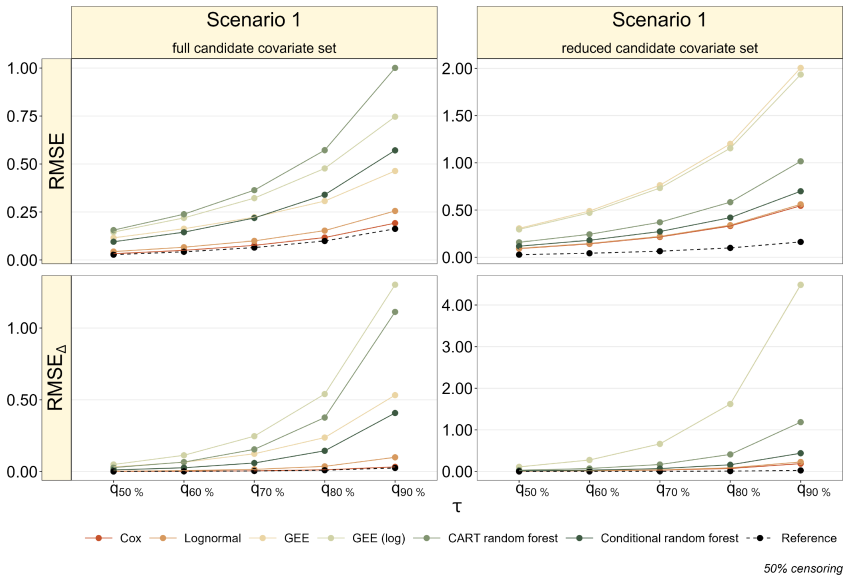


FIGURE 4 | Results of the additional simulation study (50% censoring, main-effects-only Scenario 1, *Cox* and *Lognormal* models expected to perform well). The left panels show the average RMSE and RMSE$_\Delta$ values obtained from the full candidate covariate set, whereas the right panels show the respective values obtained from the reduced candidate covariate set without $X^{(2)}$ and $X^{(7)}$. Note the different scalings of the y-axes in the left and right panels.

random forest to perform better than standard modeling approaches in estimating the RMST and treatment effects. The value of τ relative to the transition time t_0 significantly impacts the performance of the standard modeling techniques *Cox* and *Lognormal*, especially in the scenarios with time-varying treatment effects. In the main-effects-only scenarios, the conditional random forest still performs well when some information is omitted from the set of candidate covariates. Regarding the estimation of treatment effects, the conditional random forest performed considerably better than the CART random forest (both on the data used for model fitting and on unseen data).

5 | Application

To illustrate the PVRF approach, we applied the conditional random forest to data from the multicenter randomized phase III SUCCESS-A trial (NCT02181101). SUCCESS-A enrolled 3 754 patients with a primary invasive breast cancer between September 2005 and March 2007 [26]. Study participants were randomly assigned in a 1:1 ratio to one of two treatment arms, which received either standard chemotherapy (control group) or standard chemotherapy with the addition of gemcitabine (interventional group). For details on the inclusion/exclusion criteria and the design of the study see de Gregorio et al. [26].

The primary aim of the SUCCESS-A trial was to compare the two treatment arms with respect to disease-free survival (DFS), defined as the period from the date of randomization to the earliest date of disease progression (distant metastases, local and contra local recurrence, and secondary primary tumors) or death from any cause [26, 35]. Here, we present the results of a secondary analysis that considered DFS as the outcome of interest. Note that the definition of DFS includes death from any cause. Accordingly, we did not consider death as a competing event.

Patients were censored at the last date at which they were known to be disease-free, resulting in an event proportion of 12.2% (458 events in 3 754 patients). The maximum observation time was 5.5 years (6 months of chemotherapy followed by 5 years of follow-up; median 5.2 years, first quartile 3.7 years, third quartile 5.5 years). Patient characteristics included age at randomization (*age*, in years), body mass index (*BMI*, in kg/m^2) and menopausal status (*meno*, two categories, pre-/post-menopausal) as well as information on the tumor, including stage (*stage*, four categories, pT1/pT2/pT3/pT4), grade (*grade*, three categories, G1/G2/G3), lymph node status (*nodal status*, two categories, pN0/pN+), type (*type*, three categories, ductal/lobular/other) and receptor status of estrogen (*ER*), progesterone (*PR*), and *HER2* (two categories each, negative/positive). A descriptive summary of the variables is given in Table F1 in Section F. Patients with missing values in any of the considered covariates were excluded from our analysis. The analyzed data comprised 3 652 patients.

The main aim of our analysis was to model the RMST for DFS at $\tau = 5$ years, corresponding to the length of the follow-up period. To this end, we applied the conditional random forest, the *Cox* model and the *GEE* model to the SUCCESS-A study data. The covariates were defined by the treatment (control/intervention) and the ten patient/tumor characteristics described above. The accuracy of the models was measured by the weighted residual sum of squares (WRSS), an inverse-probability-of-censoring (IPC) weighted error measure [36] and by a 95% normal bootstrap confidence interval (CI, 1 000 repetitions). The WRSS is defined as

$$\mathrm{WRSS} = \frac{1}{n}\sum_i \left(\min(\tilde{T}_i, \tau) - \hat{\mu}(\tau|X_i)\right)^2 \cdot \hat{w}_i, \tag{10}$$

with $\hat{\mu}(\tau|X_i)$ denoting the estimated RMST for individual i. The IPC weights $\hat{w}_i$ are defined by $\hat{w}_i = \mathbb{1}\{\tilde{T}_i \leq \tau\} \cdot \delta_i \,/\, \hat{G}(\tilde{T}_i|X_i) + \mathbb{1}\{\tilde{T}_i > \tau\} \,/\, \hat{G}(\tau|X_i)$, where $\hat{G}$ is a consistent estimator of the censoring survival function. In this work, we use the Kaplan–Meier method to estimate $\hat{G}$. The average treatment effect (measured in days, control vs. interventional group) was calculated as

$$\hat{\Delta}(\tau) = \frac{1}{n}\sum_i \Big[\hat{\mu}(\tau|X_i^{(-trt)}, X_i^{(trt)} = \text{control}) - \hat{\mu}(\tau|X_i^{(-trt)}, X_i^{(trt)} = \text{interventional})\Big]. \tag{11}$$

To enhance the interpretability of the conditional random forest, we computed the permutation feature importance (PFI_j) along with a 95% normal bootstrap CI and Shapley values for each covariate [25]. PFI_j is defined as the ratio of the WRSS with $\hat{\mu}(\tau|X_i)$ derived from the fitted model but using permuted values of the j-th covariate (numerator), and the WRSS with $\hat{\mu}(\tau|X_i)$ calculated as usual (denominator, see Equation (B1)). Thus, higher PFI_j values indicate a higher importance of the j-th covariate for estimating the RMST. Local Shapley values were derived for 1 000 randomly selected patients. A high absolute local Shapley value indicates a high importance of the respective covariate in the estimation of the RMST.

The results of our analysis are presented in Figure 5. They show that the conditional random forest detected several established prognostic factors and subgroups, which have been consistently

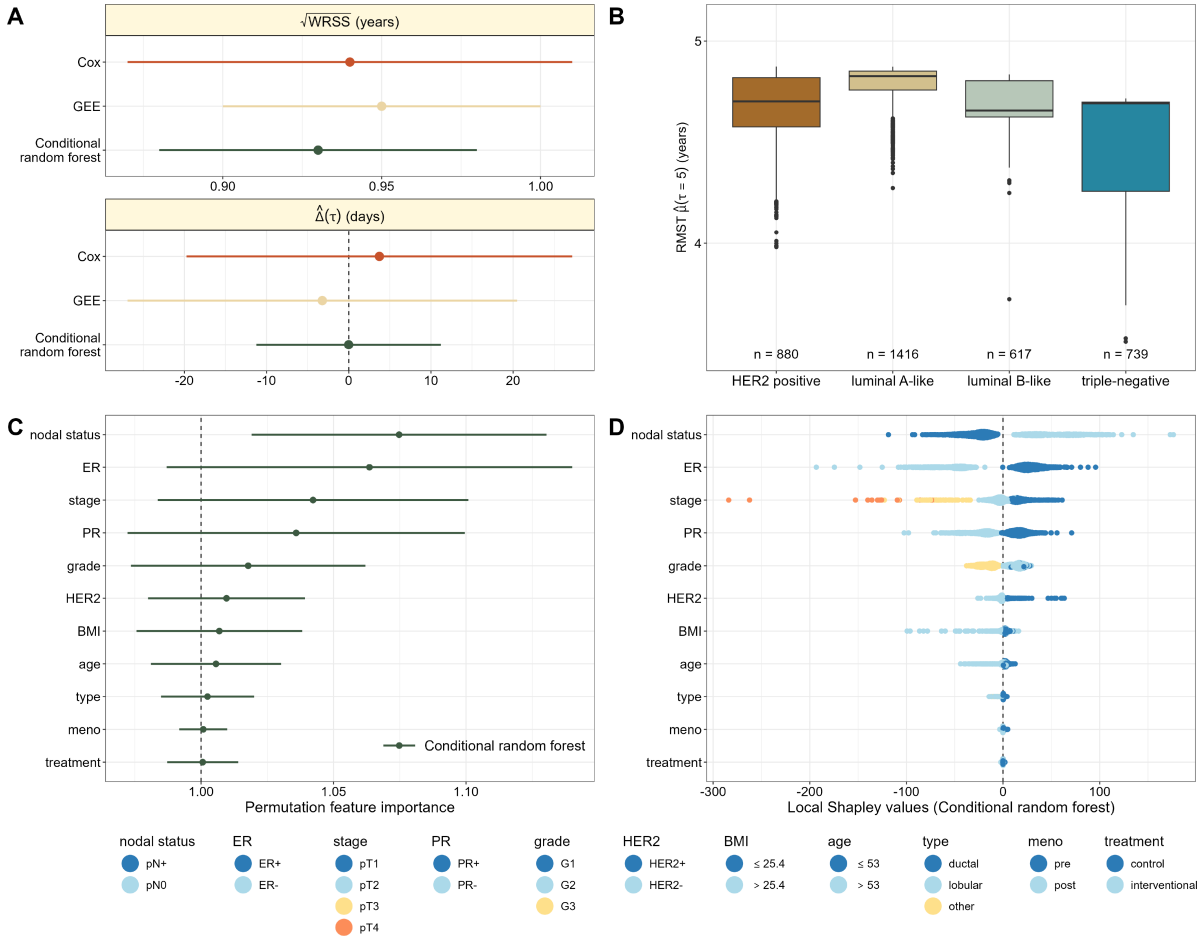


FIGURE 5 | Analysis of DFS in the SUCCESS-A study. The four panels present the results obtained from the *Cox*, *GEE* and conditional random forest methods. Panel (A) shows the square root of the WRSS (measured in years) and the treatment effect $\hat{\Delta}(\tau)$ (measured in days). The dots represent values of $\sqrt{\mathrm{WRSS}}$ and $\hat{\Delta}(\tau)$ derived by the complete cohort, while the bars refer to 95% bootstrap confidence intervals for $\sqrt{\mathrm{WRSS}}$ and $\hat{\Delta}(\tau)$. Panel (B) shows the estimated RMST values in patient groups defined by molecular tumor subtypes, as obtained from the conditional random forest (see Table F2). Panel (C) shows the permutation feature importance of each covariate for the conditional random forest. The dots represent PFI_j values and the bars refer to the respective 95% bootstrap confidence intervals. Panel (D) presents local Shapley values for each covariate, as obtained by evaluating the conditional random forest estimates in 1 000 randomly selected patients. Each dot corresponds to one patient. The color codings used in Panel (D) are presented in the bottom row of the figure.

reported in the literature and have also entered treatment guidelines for primary breast cancer [37, 38]. Regarding the WRSS, the conditional random forest ($\sqrt{\text{WRSS}} = 0.93$ years, 95% CI: 0.88 to 0.98 years) performs best, followed by the *Cox* model ($\sqrt{\text{WRSS}} = 0.94$ years, 95% CI: 0.87 to 1.01 years) and the *GEE* approach ($\sqrt{\text{WRSS}} = 0.95$ years, 95% CI: 0.90 to 1.00 years, Figure 5A). The average treatment effect $\hat{\Delta}(\tau)$ estimated by the conditional random forest is close to zero (−0.01 days, 95% CI: −11.22 to 11.21 days), indicating no advantage of any of the two groups. This supports the results found by de Gregorio et al. [26], who concluded that the interventional treatment does not improve survival in patients with high-risk early breast cancer. In contrast, the *Cox* model indicates a slight advantage of the control group ($\hat{\Delta}(\tau) = 3.73$ days, 95% CI: −19.74 to 27.20 days) while the *GEE* model indicates a slight advantage of the interventional group ($\hat{\Delta}(\tau) = -3.22$ days, 95% CI: −26.92 to 20.49 days). Note, however, that the treatment difference is measured in days, so none of the obtained differences can be considered clinically relevant.

Figure 5B visualizes the RMST values at $\tau = 5$ years in patient groups defined by molecular tumor subtypes [39]. More specifically, *HER2 positive* patients are characterized by *HER2* positive tumors (regardless of *ER* status, *PR* status and *grade*). *HER2* negative tumors are further classified into *luminal A-like* tumors (*HER2* negative, *ER* and/or *PR* status positive, *grade* G1 or G2), *luminal B-like* tumors (*HER2* negative, *ER* and/or *PR* status positive, *grade* G3), and *triple-negative* tumors (*HER2*, *ER* and *PR* status negative and any *grade*, see Table F2). According to Figure 5B, the high-risk *triple-negative* group has the lowest estimated RMST values (mean (SD): 4.46 (0.24) years), which is consistent with findings in the literature [40]. Additionally, when comparing the *luminal A-like* and *luminal B-like* subgroups, there appears to be a slight advantage (corresponding to higher estimated RMST values) of patients with tumor *grade* G1 (*luminal A-like*, 4.78 (0.10) years) compared to those with tumor *grade* G2 or G3 (*luminal B-like*, 4.68 (0.12) years). Again, this result is in line with previous findings in the literature [40]. The comparison of *luminal A-like*, *luminal B-like* and *triple-negative* confirms the ability of the conditional random forest to identify interactions between hormone receptor status and grade.

As illustrated in Figure 5C, the PFI_j values obtained from the conditional random forest identify *nodal status* as the most influential covariate in the estimation of the RMST. The strong influence of *nodal status* on DFS has previously been reported by Senkus et al. [38]. Other important covariates (in terms of PFI_j) are *ER*, *stage*, *PR*, *grade*, *HER2* and *BMI*. Notably, all other covariates appear to have negligible importance in estimating RMST values by the conditional random forest, including *treatment*. This result is in line with the findings of de Gregorio et al. [26].

The local Shapley values in Figure 5D are also in line with previous findings in the literature [38, 40, 41] and with the PFI_j values in Figure 5C. As seen from Figure 5D, lymph node positive patients (*nodal status* = pN+) exhibit higher risk of recurrence or death, reflected by negative local Shapley values of these patients. Furthermore, the high Shapley values for *ER* confirm the importance of this covariate in adjuvant hormonal and chemotherapy. The survival advantage of *ER* positive patients [41] is reflected by positive Shapley values for this group. Conversely, negative Shapley values are observed for *ER* negative, *PR* negative, and *HER2* negative patients, which is consistent with lower estimated RMST values for the *triple-negative* group in Figure 5B [40]. Likewise, the difference in estimated RMST values between *luminal A-like* and *luminal B-like* patients is confirmed by the respective local Shapley values: Patients with *grade* G1 and G2 have a positive contribution to the estimated RMST values, while patients with *grade* G3 contribute negatively. Furthermore, the local Shapley values accurately reflect the hierarchy of tumor stages: Tumor stage pT1 (best prognosis for DFS) has a positive contribution to the estimated RMST, whereas tumor stages pT2 to pT4 have increasingly negative contributions. Shapley values for *treatment* spread around 0 in both groups, suggesting neither a positive nor a negative contribution of the treatment to the RMST. Again, this result is consistent with the findings of de Gregorio et al. [26]. In addition to the bootstrapped estimates, we evaluated cross-validated values of WRSS and PFI_j. As seen from Figure F1 in Section F, these values are very similar to those obtained from the bootstrap procedure.

6 | Discussion

During the past years, the restricted mean survival time has become an increasingly popular measure for summarizing individual event times in medical studies. Compared to other established measures like the hazard ratio, the RMST is derived from survival probabilities measured at the untransformed risk scale, thereby avoiding interpretability and collapsibility issues in the comparison of interventional groups [42, 43]. As a consequence, the RMST is considered a valid survival estimand for the causal interpretation of treatment contrasts in clinical and observational trials [5, 44].

In this work, we proposed the pseudo-value random forest (PVRF) method, which is a non-parametric approach for the quantification of treatment effects by group-specific RMST values. Instead of estimating RMST values from (semi-)parametric models like Cox or AFT regression, the PVRF method combines unconditional pseudo-value RMST estimation with the subsequent fitting of a random forest. Except for the random censoring assumption, both components of our method (pseudo-values and random forests) require minimal assumptions on the data-generating process. While unconditional pseudo-values are based on non-parametric Kaplan–Meier estimates, random forest regression is a model-free algorithm allowing for variable selection and requiring no prior assumptions on the structure of the covariate effects. As a result, the PVRF method is particularly suited for incorporating subgroup characteristics, non-linearities, and higher-order interactions affecting individual RMST values. In non-randomized studies, this approach is particularly useful when treatment effects need to be corrected for higher-dimensional sets of confounders, allowing for the estimation of causal contrasts via g-computation. Furthermore, our method enables model-free comparisons of treatment and control groups in randomized trials. Regarding the latter, we demonstrated that PVRF is able to capture time-dependent treatment effects in a data-driven way (see Section 3, where PVRF performed better than (semi-)parametric approaches in

the scenarios with crossing survival curves). Methods to adjust RMST estimation for covariate-dependent censoring have been studied in Rong et al. [15].

In our numerical studies, we observed that the conditional random forest (correcting for a possible selection bias towards covariates with many possible splits) showed a better performance in terms of RMSE than the traditional CART random forest approach. This finding was particularly evident in the estimation of treatment contrasts, where conditional random forests outperformed CART in all scenarios. We therefore recommend to prefer conditional random forests over CART random forests when the aim is to estimate treatment effects from data with heterogeneous covariate types.

An important topic for future research is the development of hypothesis tests and confidence intervals for PVRF-based RMST differences. Previous research in this field [1, 13, 14] has mainly focused on hypothesis tests for RMST differences derived by group-wise integration of Kaplan–Meier curves (not incorporating additional covariates). Tian et al. [13] compared RMST-based tests to HR-based tests in the context of randomized clinical trials, demonstrating that RMST-based tests outperformed their HR-based counterparts in scenarios where the PH assumption is violated. It would be interesting to conduct analog studies for pseudo-value-based tests of RMST differences, which, to the best of our knowledge, have not yet been explored thus far. In our analysis of the SUCCESS-A study data (Section 5), we constructed confidence intervals for treatment contrasts using bootstrap methods, along the lines of Hernán & Robins [45], Chapter 13.

A general issue in the estimation of RMST values is the choice of a suitable time horizon τ. While choosing a small value of τ may discard a large proportion of the data and will therefore result in a potential loss of information, estimation of the RMST may no longer be possible if τ becomes too large [46]. General recommendations on the choice of τ, have, for instance, been made by Tian et al. [46]: Before data collection (for instance, in the course of planning a clinical trial), it is advisable to pre-select τ based on clinical and feasibility considerations. If pre-selection of τ is not possible (e.g., when the analysis is conceived after data collection), Tian et al., suggest to explore a data-dependent time window for τ and to select the time horizon based on the empirical behavior of the observed times in this window (e.g., by computing quantiles of $\tilde{T}$, as done in our simulations). Alternatively, the RMST could be modeled as a function of τ, as suggested by Zhong & Schaubel [47].

We finally note that pseudo-value-based RMST modeling is not restricted to the use of random forest regression. In this work, we focused on random forests because this method is considered to be "among the best "off-the-shelf" supervised learning methods that are available" [48]. In particular, random forests are known to perform well on medium-sized data (as often encountered in medical applications), with several efficient software implementations being available [49]. However, it is of course possible to extend our approach to other statistical modeling or machine learning techniques, e.g., to gradient boosting [27] or deep neural networks [32, 50, 51].

Author Contributions

A.S and M.S conceived and designed the project. A.S., V.B., and M.S. analyzed and interpreted the results. A.S. and M.S. drafted the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgments

We thank Dr. Lothar Häberle (Department of Gynecology, Obstetrics and Mammology, University Hospital Erlangen, Germany) for supporting us with the analysis of the SUCCESS-A study data. Open Access funding enabled and organized by Projekt DEAL.

Disclosure

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts 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.

References

1. P. Royston and M. K. B. Parmar, "Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials With a Time-To-Event Outcome," *BMC Medical Research Methodology* 13 (2013): 152.

2. F. Ambrogi, S. Iacobelli, and P. K. Andersen, "Analyzing Differences Between Restricted Mean Survival Time Curves Using Pseudo-Values," *BMC Medical Research Methodology* 22 (2022): 71.

3. Z. R. McCaw, A. R. Orkaby, L. J. Wei, D. H. Kim, and M. W. Rich, "Applying Evidence-Based Medicine to Shared Decision Making: Value of Restricted Mean Survival Time," *American Journal of Medicine* 132 (2019): 13–15.

4. P. Royston and M. K. B. Parmar, "The Use of Restricted Mean Survival Time to Estimate the Treatment Effect in Randomized Clinical Trials When the Proportional Hazards Assumption Is in Doubt," *Statistics in Medicine* 30 (2011): 2409–2421.

5. A. Ni, Z. Lin, and B. Lu, "Stratified Restricted Mean Survival Time Model for Marginal Causal Effect in Observational Survival Data," *Annals of Epidemiology* 64 (2021): 149–154.

6. H. Uno, B. Claggett, L. Tian, et al., "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis," *Journal of Clinical Oncology* 32 (2014): 2380–2385.

7. H. M. Dehbi, P. Royston, and A. Hackshaw, "Life Expectancy Difference and Life Expectancy Ratio: Two Measures of Treatment Effects in Randomised Trials With Non-Proportional Hazards," *British Medical Journal* 357 (2017): 2250.

8. M. J. Stensrud and M. A. Hernán, "Why Test for Proportional Hazards?," *Journal of the American Medical Association* 323 (2020): 1401–1402.

9. P. Andersen, M. Hansen, and J. Klein, "Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations," *Lifetime Data Analysis* 10 (2005): 335–350.

10. L. Breiman, "Random Forests," *Machine Learning* 45 (2001): 5–32.

11. U. B. Mogensen and T. A. Gerds, "A Random Forest Approach for Competing Risks Based on Pseudo-Values," *Statistics in Medicine* 32 (2013): 3102–3114.

12. S. E. Leurgans, "Linear Models, Random Censoring and Synthetic Data," *Biometrika* 74 (1987): 301–309.

13. L. Tian, H. Fu, S. J. Ruberg, H. Uno, and L. J. Wei, "Efficiency of Two Sample Tests via the Restricted Mean Survival Time for Analyzing Event Time Observations," *Biometrics* 74 (2018): 694–702.

14. B. Huang and P. F. Kuan, "Comparison of the Restricted Mean Survival Time With the Hazard Ratio in Superiority Trials With a Time-To-Event End Point," *Pharmaceutical Statistics* 17 (2018): 202–213.

15. R. Rong, J. Ning, and H. Zhu, "Regression Modeling of Restricted Mean Survival Time for Left-Truncated Right-Censored Data," *Statistics in Medicine* 41 (2022): 3003–3021.

16. L. Tian, L. Zhao, and L. J. Wei, "Predicting the Restricted Mean Event Time With the Subject's Baseline Covariates in Survival Analysis," *Biostatistics* 15 (2014): 222–233.

17. X. Wang and D. Schaubel, "Modeling Restricted Mean Survival Time Under General Censoring Mechanisms," *Lifetime Data Analysis* 24 (2018): 176–199.

18. T. Hasegawa, S. Misawa, S. Nakagawa, et al., "Restricted Mean Survival Time as a Summary Measure of Time-To-Event Outcome," *Pharmaceutical Statistics* 19 (2020): 436–453.

19. D. H. Kim, H. Uno, and L. J. Wei, "Restricted Mean Survival Time as a Measure to Interpret Clinical Trial Results," *JAMA Cardiology* 2 (2017): 1179–1180.

20. P. K. Andersen and P. M. Pohar, "Pseudo-Observations in Survival Analysis," *Statistical Methods in Medical Research* 19 (2010): 71–99.

21. F. Graw, T. A. Gerds, and M. Schumacher, "On Pseudo-Values for Regression Analysis in Competing Risks Models," *Lifetime Data Analysis* 15 (2009): 241–255.

22. M. Overgaard, E. T. Parner, and J. Pedersen, "Asymptotic Theory of Generalized Estimating Equations Based on Jack-Knife Pseudo-Observations," *Annals of Statistics* 45 (2017): 1988–2015.

23. J. Robins, "A New Approach to Causal Inference in Mortality Studies With a Sustained Exposure Period - Application to Control of the Healthy Worker Survivor Effect," *Mathematical Modelling* 7 (1986): 1393–1512.

24. J. M. Snowden, S. Rose, and K. M. Mortimer, "Implementation of g-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique," *American Journal of Epidemiology* 173 (2011): 731–738.

25. C. Molnar, *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*, 2nd ed. (Independently published, 2022).

26. A. de Gregorio, L. Häberle, P. A. Fasching, et al., "Gemcitabine as Adjuvant Chemotherapy in Patients With High-Risk Early Breast Cancer – Results From the Randomized Phase III SUCCESS-A Trial," *Breast Cancer Research* 22 (2020): 111.

27. A. Schenk, M. Berger, and M. Schmid, "Pseudo-Value Regression Trees," *Lifetime Data Analysis* 30 (2024): 439–471.

28. L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees* (Taylor & Francis, 1984).

29. T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics* 15 (2006): 651–674.

30. B. M. Greenwell, *Tree-Based Methods for Statistical Learning in R* (Chapman & Hall/CRC, 2022).

31. R. de Bin, S. Janitza, W. Sauerbrei, and A. L. Boulesteix, "Subsampling Versus Bootstrapping in Resampling-Based Model Selection for Multivariable Regression," *Biometrics* 72 (2016): 272–280.

32. L. Hu, J. Ji, and F. Li, "Estimating Heterogeneous Survival Treatment Effect in Observational Data Using Machine Learning," *Statistics in Medicine* 40 (2021): 4691–4713.

33. R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, version 4.1.2 (2021).

34. T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model* (Springer, 2000).

35. C. A. Hudis, W. E. Barlow, J. P. Costantino, et al., "Proposal for Standardized Definitions for Efficacy End Points in Adjuvant Breast Cancer Trials: The STEEP System," *Journal of Clinical Oncology* 25 (2007): 2127–2132.

36. A. Cwiling, V. Perduca, and O. Bouaziz, "A Comprehensive Framework for Evaluating Time to Event Predictions Using the Restricted Mean Survival Time," arXiv: arXiv.2306.16075.

37. A. S. Coates, E. P. Winer, A. Goldhirsch, et al., "Tailoring Therapies – Improving the Management of Early Breast Cancer: St.Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015," *Annals of Oncology* 26 (2015): 1533–1546.

38. E. Senkus, S. Kyriakides, S. Ohno, et al., "Primary Breast Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up," *Annals of Oncology* 26, no. Suppl. 5 (2015): v8–v30.

39. C. M. Perou, T. Sørlie, M. B. Eisen, et al., "Molecular Portraits of Human Breast Tumours," *Nature* 406 (2000): 747–752.

40. G. von Minckwitz, M. Untch, J. U. Blohmer, et al., "Definition and Impact of Pathologic Complete Response on Prognosis After Neoadjuvant Chemotherapy in Various Intrinsic Breast Cancer Subtypes," *Journal of Clinical Oncology* 30 (2012): 1796–1804.

41. A. Goldhirsch, W. C. Wood, R. D. Gelber, A. S. Coates, B. Thürlimann, and H. J. Senn, "Meeting Highlights: Updated International Expert Consensus on the Primary Therapy of Early Breast Cancer," *Journal of Clinical Oncology* 21 (2003): 3357–3365.

42. M. A. Hernán, "The Hazards of Hazard Ratios," *Epidemiology* 21 (2010): 13–15.

43. V. Didelez and M. J. Stensrud, "On the Logic of Collapsibility for Causal Effect Measures," *Biometrical Journal* 64 (2022): 235–242.

44. P. Y. Chen and A. A. Tsiatis, "Causal Inference on the Difference of the Restricted Mean Lifetime Between Two Groups," *Biometrics* 57 (2001): 1030–1038.

45. M. A. Hernán and J. M. Robins, *Causal Inference: What if* (CRC Press, 2024).

46. L. Tian, H. Jin, H. Uno, et al., "On the Empirical Choice of the Time Window for Restricted Mean Survival Time," *Biometrics* 76 (2020): 1157–1166.

47. Y. Zhong and D. E. Schaubel, "Restricted Mean Survival Time as a Function of Restriction Time," *Biometrics* 78 (2022): 192–201.

48. T. Coleman, L. Mentch, D. Fink, et al., "Statistical Inference on Tree Swallow Migrations With Random Forests," *Journal of the Royal Statistical Society: Series C: Applied Statistics* 69 (2020): 973–989.

49. M. N. Wright and A. Ziegler, "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R," *Journal of Statistical Software* 77 (2017): 1–17.

50. L. Zhao and D. Feng, "Deep Neural Networks for Survival Analysis Using Pseudo Values," *IEEE Journal of Biomedical and Health Informatics* 24 (2020): 3308–3314.

51. L. Zhao, "Deep Neural Networks for Predicting Restricted Mean Survival Times," *Bioinformatics* 36 (2021): 5672–5677.

52. K. Goldfeld and J. Wujciak-Jens, "simstudy: Illuminating Research Methods Through Data Generation," R package version 0.7.1, (2023).

53. M. Pohar Perme and M. Gerster, *pseudo: Computes Pseudo-Observations for Modeling*, R package version 1.4.3 (2017).

54. T. Hothorn, H. Seibold, and A. Zeileis, "partykit: A toolkit for recursive partytioning," R package version 1.2.20 (2023).

55. T. M. Therneau, "survival: A package for survival analysis in R," R package version 3.5.7 (2023).

56. S. Højsgaard, U. Halekoh, and J. Yan, "The R Package Geepack for Generalized Estimating Equations," *Journal of Statistical Software* 15 (2006): 1–11.

57. C. Molnar, B. Bischl, and G. Casalicchio, "Iml: An R Package for Interpretable Machine Learning," *Journal of Open Source Software* 3 (2018): 786.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix A

Simulation Study

Covariance Matrix of the Continuous Covariates

TABLE A1 | Covariance matrix of the continuous covariates $X^{(j)}$, $j = 1, \ldots, 5$.

	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$
$X^{(1)}$	1.00	−0.08	−0.47	0.73	−0.44
$X^{(2)}$	−0.08	1.00	0.85	−0.05	−0.31
$X^{(3)}$	−0.47	0.85	1.00	−0.38	−0.33
$X^{(4)}$	0.73	−0.05	−0.38	1.00	−0.37
$X^{(5)}$	−0.44	−0.31	−0.33	−0.37	1.00

Values of the Time Horizon τ

TABLE A2 | Time horizons τ used in the simulation study.

Scenario	Censoring proportion	$q_{50\%}$	$q_{60\%}$	$q_{70\%}$	$q_{80\%}$	$q_{90\%}$
1	25%	1.72	2.81	4.64	8.08	16.95
	50%	1.09	1.56	2.23	3.28	5.24
	75%	0.40	0.55	0.74	1.02	1.52
2	25%	1.37	2.38	4.26	8.19	19.64
	50%	0.78	1.15	1.68	2.51	4.07
	75%	0.23	0.32	0.44	0.62	0.92
3	25%	1.41	1.83	2.05	2.23	2.74
	50%	0.64	0.88	1.07	1.22	1.65
	75%	0.31	0.42	0.55	0.64	0.86
4	25%	1.22	1.67	2.03	2.19	2.77
	50%	0.52	0.74	1.01	1.14	1.57
	75%	0.20	0.27	0.36	0.43	0.60

Note: Note that $q_{70\%}$ is approximately equal to the transition time t_0 in Scenarios 3 and 4.

Appendix B

Specification and Implementation of the Methods

The simulation study and the application were carried out in R, version 4.1.2 [33]. Data for the simulation study were generated using the R package `simstudy`, version 0.7.1 [52]. Pseudo-values for the RMST, as defined in Equation (2), were calculated using the `pseudomean` function of the R package `pseudo`, version 1.4.3 [53].

The CART random forest was implemented using the R package `ranger`, version 0.15.1 [49]. The number of trees was set to 500. Data for tree building was sampled without replacement from the complete data using a sampling fraction of 0.632. The number of candidate split variables in each node ("mtry") was tuned using five-fold cross validation. In each tree, the minimum number of observations required to perform an additional split was set to 5. In order to avoid overoptimism in the cross-validation procedure, we computed separate sets of pseudo-values in each of the training and test folds. There were no restrictions on the tree depth and the minimum number of observations in the leaf nodes.

The conditional random forest was implemented using `cforest` function of the R package `partykit`, version 1.2.20 [54]. The number of trees was set to 500. By default, `cforest` implements sampling without replacement from the complete data, using a sampling fraction of 0.632. The number of candidate split variables in each node was tuned using five-fold cross validation. In order to avoid overoptimism in the cross-validation procedure, we computed separate sets of pseudo-values in each of the training and test folds. In each tree, a minimum of 20 observations was required to perform a split, and each leaf node was required to contain a minimum of 7 observations. There was no restriction on the depth of the trees.

The *Lognormal*, *Cox*, and *Reference* methods were implemented using the R package `survival`, version 3.5.7 [55]. *GEE* and *GEE (log)* were implemented using the R package `geepack`, version 1.3.9 [56].

Permutation feature importance values were calculated as

$$\mathrm{PFI}_j = \frac{\mathrm{WRSS}^{\mathrm{perm},j}}{\mathrm{WRSS}^{\mathrm{orig}}}, \tag{B1}$$

where $\mathrm{WRSS}^{\mathrm{orig}}$ denotes the WRSS calculated from the unpermuted data and $\mathrm{WRSS}^{\mathrm{perm},j}$ denotes the respective WRSS calculated from data with randomly permuted values of the j-th covariate. Local Shapley values were calculated using the R package `iml`, version 0.11.2 [57].

The R-code for the simulation study is available at https://www.imbie.uni-bonn.de/cloud/index.php/s/6gmJQmayFAMJZHk.

Appendix C

Simulation Results for Censoring Proportions 25% and 75%

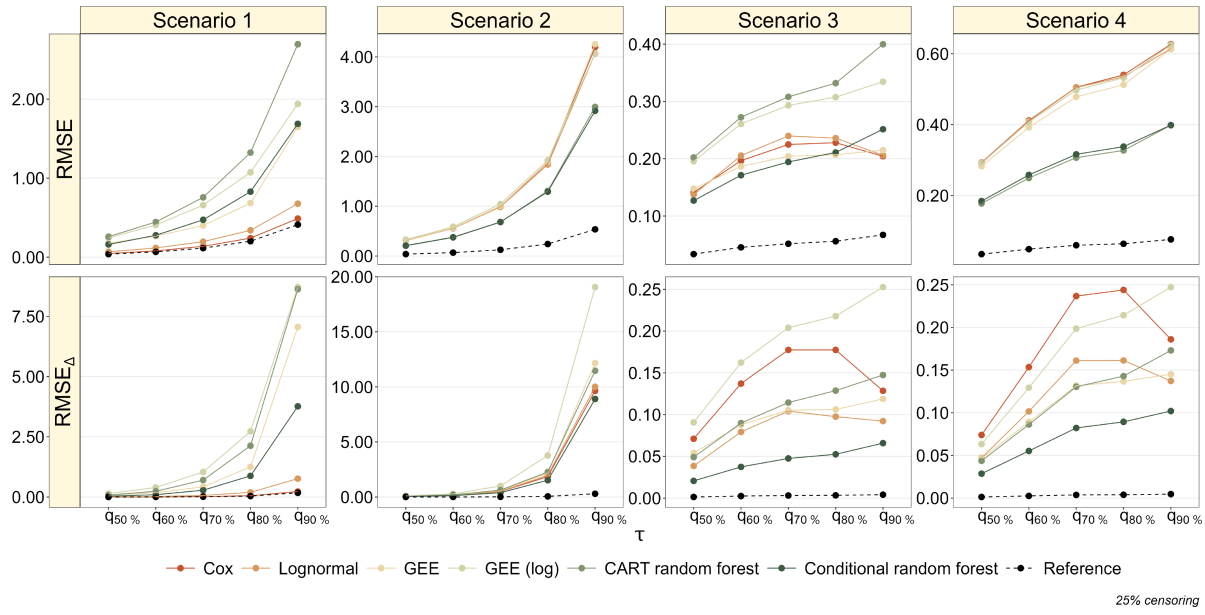


FIGURE C1 | Results of the simulation study (25% censoring). The upper panels present the average RMSE (7), as obtained from the RMST estimates at different values of τ. The lower panels present the average RMSE for the treatment effect (9). The dashed black lines refer to the correctly specified Cox PH model (*Reference*).

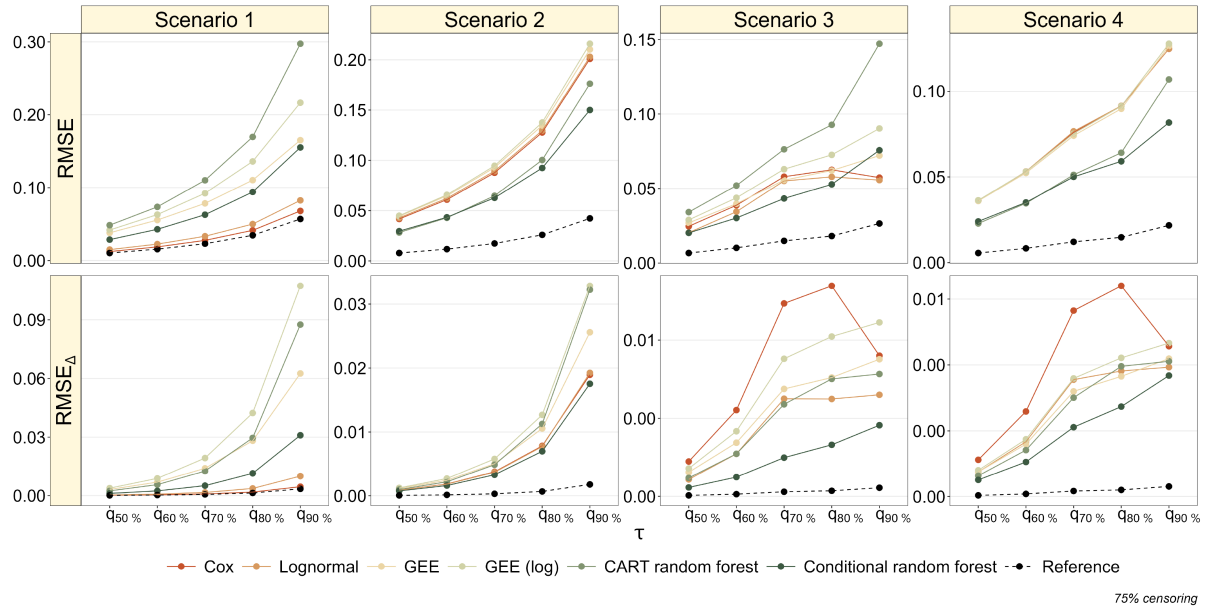


FIGURE C2 | Results of the simulation study (75% censoring). The upper panels present the average RMSE (7), as obtained from the RMST estimates at different values of τ. The lower panels present the average RMSE for the treatment effect (9). The dashed black lines refer to the correctly specified Cox PH model (*Reference*).

Appendix D

Simulation Results on Unseen Data

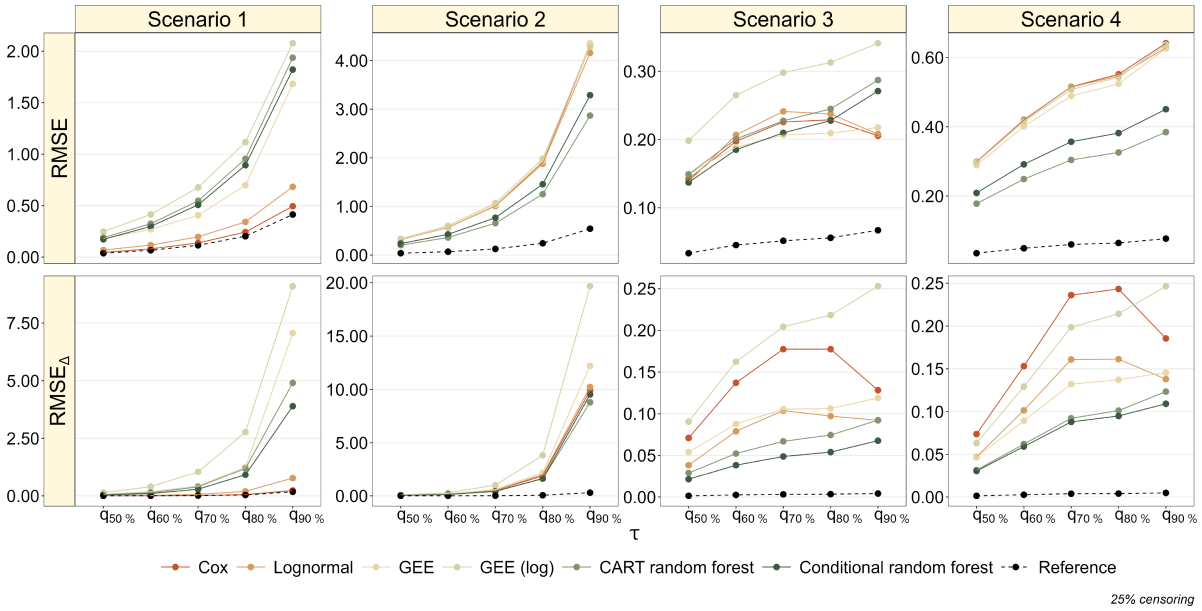


FIGURE D1 | Results of the simulation study (25% censoring). The upper panels present the average RMSE (7), as obtained from the RMST estimates at different values of τ. The lower panels present the average RMSE for the treatment effect (9). All RMSE and RMSE$_\Delta$ values were obtained by applying the model fits from Section 4 to independent data sets of size $n_{\text{test}} = 1000$ each that were generated according to Scenarios 1–4. The dashed black lines refer to the correctly specified Cox PH model (*Reference*).

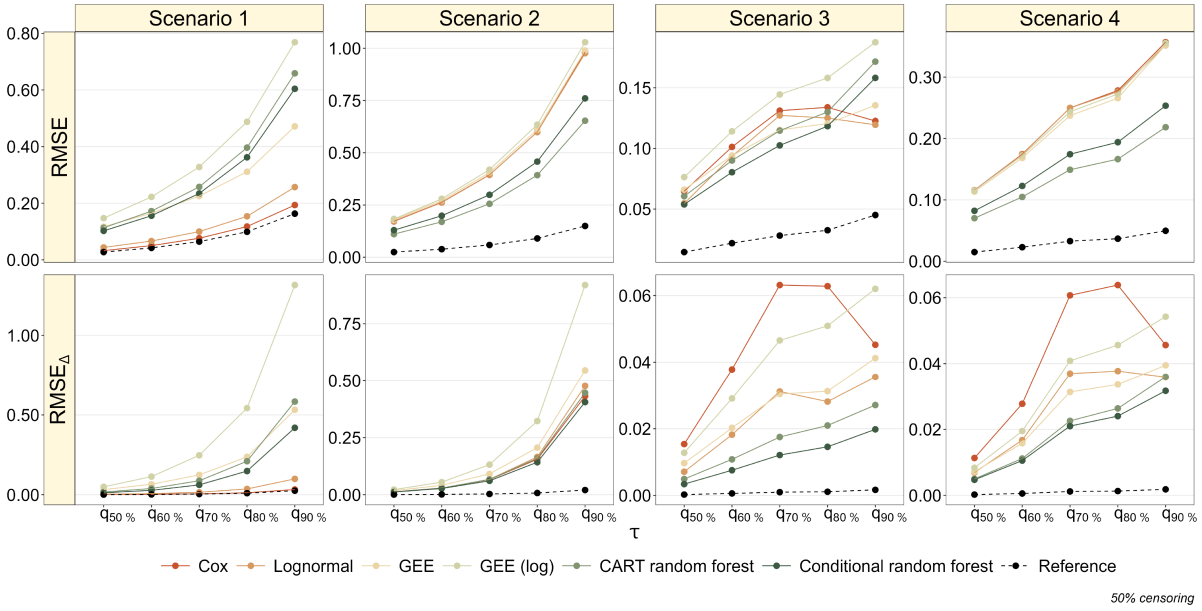


FIGURE D2 | Results of the simulation study (50% censoring). The upper panels present the average RMSE (7), as obtained from the RMST estimates at different values of τ. The lower panels present the average RMSE for the treatment effect (9). All RMSE and RMSE$_\Delta$ values were obtained by applying the model fits from Section 4 to independent data sets of size $n_{\text{test}} = 1000$ each that were generated according to Scenarios 1–4. The dashed black lines refer to the correctly specified Cox PH model (*Reference*).

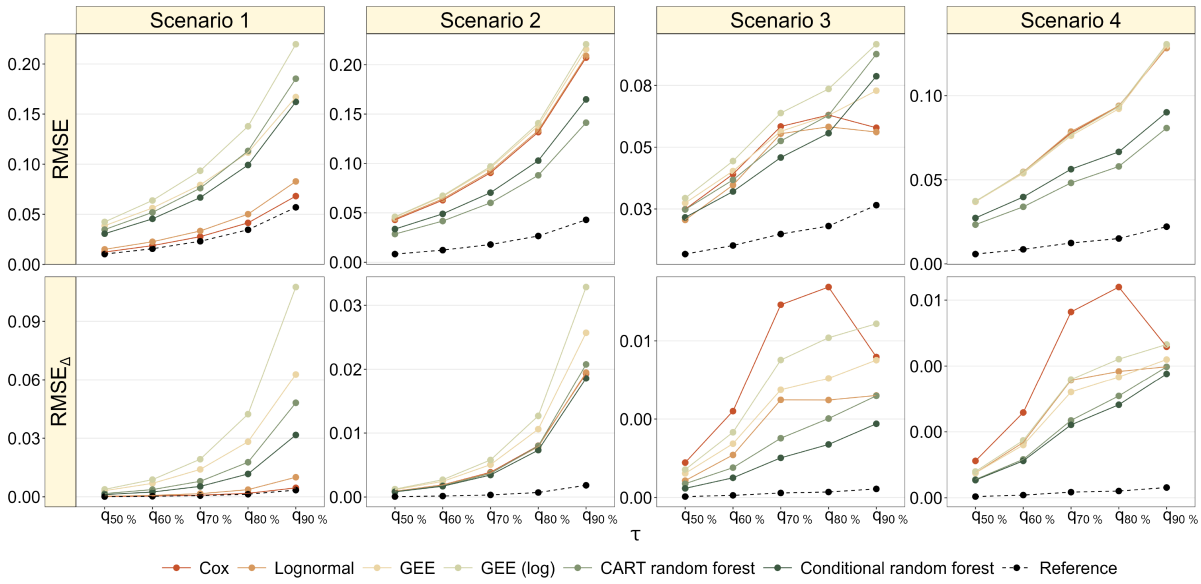


FIGURE D3 | Results of the simulation study (75% censoring). The upper panels present the average RMSE (7), as obtained from the RMST estimates at different values of τ. The lower panels present the average RMSE for the treatment effect (9). All RMSE and RMSE$_\Delta$ values were obtained by applying the model fits from Section 4 to independent data sets of size $n_{\text{test}} = 1000$ each that were generated according to Scenarios 1–4. The dashed black lines refer to the correctly specified Cox PH model (*Reference*).

Appendix E

Simulation Results for the Reduced Candidate Covariate Set (50% Censoring)

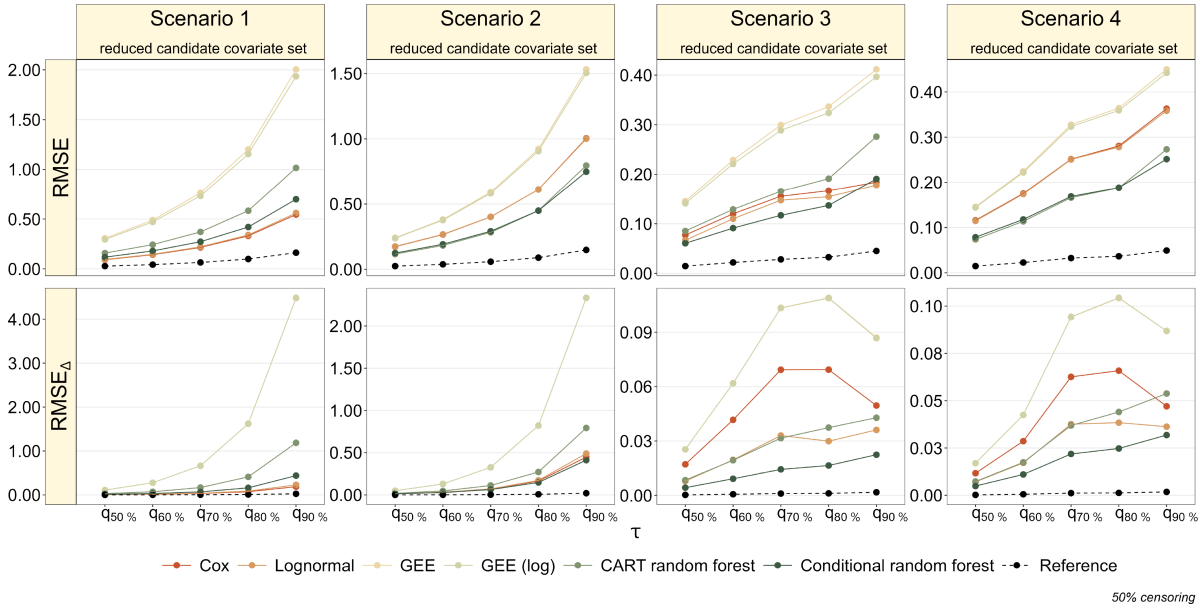


FIGURE E1 | Results of the simulation study (50% censoring). The upper panels present the average RMSE (7), as obtained from the RMST estimates at different values of τ. The lower panels present the average RMSE for the treatment effect (9). All RMSE and RMSE$_\Delta$ values were obtained by applying the methods from Section 3 with the reduced candidate covariate set (excluding $X^{(2)}$ and $X^{(7)}$). The dashed black lines refer to the correctly specified Cox PH model (*Reference*).

Appendix F

Application

Patient and Tumor Characteristics of the SUCCESS-A Study Data

TABLE F1 | Descriptive summary of the SUCCESS-A study data.

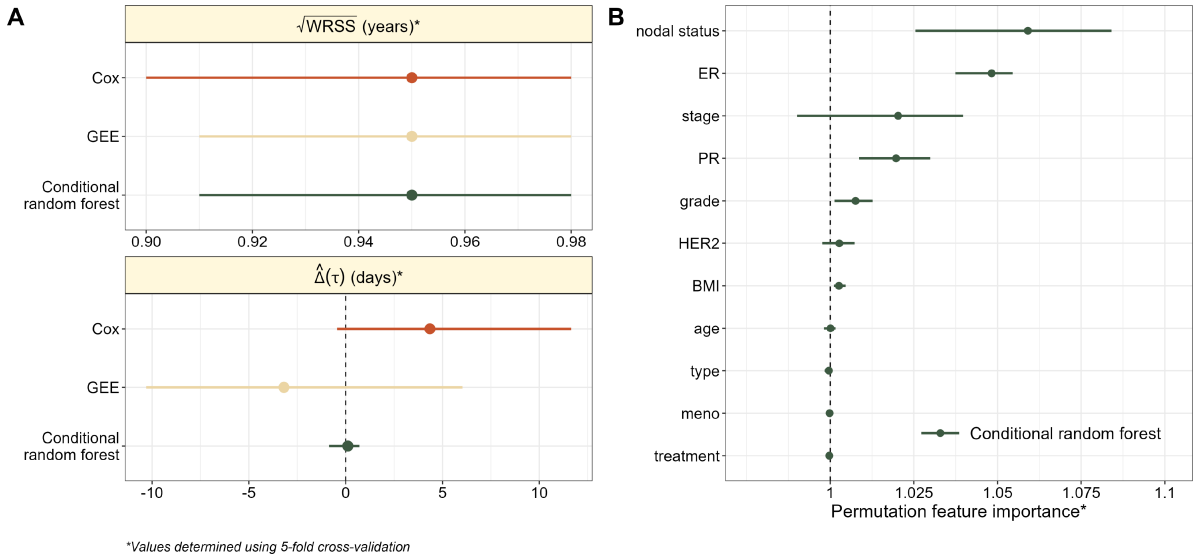
Characteristic		Patients ($n = 3\,754$)
Age (years)	Mean (SD)	53.5 (10.5)
	Median [Min, Max]	53.0 [21.0, 86.0]
BMI (kg/m^2)	Mean (SD)	26.3 (5.03)
	Median [Min, Max]	25.4 [15.4, 53.4]
Tumor stage	pT1	1552 (41.3%)
	pT2	1929 (51.4%)
	pT3	198 (5.3%)
	pT4	52 (1.4%)
	Missing	23 (0.6%)
Tumor grade	G1	176 (4.7%)
	G2	1783 (47.5%)
	G3	1773 (47.2%)
	Missing	22 (0.6%)
Lymph node status	pN+	2452 (65.3%)
	pN0	1273 (33.9%)
	Missing	29 (0.8%)
Tumor type	ductal	3060 (81.5%)
	lobular	419 (11.2%)
	other	253 (6.7%)
	Missing	22 (0.6%)
ER	ER-	1252 (33.4%)
	ER+	2481 (66.1%)
	Missing	21 (0.6%)
PR	PR-	1525 (40.6%)
	PR+	2205 (58.7%)
	Missing	24 (0.6%)
HER2	HER2-	2787 (74.2%)
	HER2+	883 (23.5%)
	Missing	84 (2.2%)
Menopausal status	pre	1565 (41.7%)
	post	2189 (58.3%)
Treatment group	Control	1898 (50.6%)
	Interventional	1856 (49.4%)

Definition of Molecular Tumor Subtypes

TABLE F2 | Definition of molecular tumor subtypes in the SUCCESS-A study data.

Subgroup	HER2	ER/PR	grade
HER2 positive	HER2+	any	any
Luminal A-like	HER2−	ER+ and/or PR+	G1 or G2
Luminal B-like	HER2−	ER+ and/or PR+	G3
Triple-negative	HER2−	ER− and PR−	any

Cross-Validated Results Obtained From the SUCCESS-A Study Data



*Values determined using 5-fold cross-validation

FIGURE F1 | Five-fold-cross-validated analysis of DFS in the SUCCESS-A study. Panel (A) shows the square root of the WRSS (measured in years) and the treatment effect $\hat{\Delta}(\tau)$ (measured in days). The dots represent the five-fold cross-validated values of $\sqrt{\text{WRSS}}$ and $\hat{\Delta}(\tau)$, while the bars refer to the respective ranges of $\sqrt{\text{WRSS}}$ and $\hat{\Delta}(\tau)$ in the five folds. Panel (B) shows the permutation feature importance of each covariate for the conditional random forest. The dots represent the five-fold cross-validated values of PFI_j and the bars refer to the respective ranges in the five folds of the cross-validation procedure.

4 Discussion with references

Accurate and efficient risk quantification by time-to-event models is a crucial challenge in patient care. As the amount of routinely collected clinical data continues to grow, there is a rising demand for advanced modeling techniques in time-to-event analysis that enable automated, data-driven selection of the most informative covariates. Ideally, such algorithms should not only identify relevant covariates from a large set of candidates but also capture interactions and complex relationships among them. This would eliminate the need for pre-specifying covariate effects within model equations, increasing flexibility and reducing manual variable pre-selection. Moreover, advanced modeling approaches for clinical time-to-event data should be applicable across diverse scenarios and data structures, which can be achieved by minimizing restrictive assumptions, further enhancing flexibility. Within this dissertation, well-established and newly developed modeling approaches for survival probabilities and RMSTs are discussed in this context and evaluated in numerical experiments and on different datasets from clinical studies. The publications in this cumulative dissertation encompass the development of a Cox-based scoring system and novel pseudo-value regression techniques based on machine learning approaches.

In clinical practice, simple scoring systems for time-to-event outcomes have long been used for risk stratification. For example, the acute physiology score (APS) was originally developed to assess the risk of death in intensive care unit patients (Knaus et al., 1981). However, since APS is calculated from a large number of physiological measures, simpler versions, such as the simplified acute physiology score (SAPS), SAPS II and SAPS III, were introduced later (Le Gall et al., 1984; Le Gall et al., 1993; Vazquez et al., 2003). This reflects the ongoing need for scoring systems that are easy to apply and provide accurate risk quantification. While these tools often apply to a broad population, covering all age groups or encompassing a wide range of comorbidities, simple scoring systems like PIRATE (Publication 1) are essential for risk assessment in a target population (Schenk et al., 2023). Elderly patients ($\geq$ 80 years) represent a particularly vulnerable group when it comes to mortality risk, especially in the context of interventions with anesthesia. Their age-related physiological decline, comorbidities,

and increased frailty make them more susceptible to complications compared to younger age groups (POSE study group, 2021). Existing risk assessment tools often target a broader age range or focus on elderly patients with undergoing a specific type of intervention (Manach et al., 2016). In contrast, PIRATE is tailored for all elderly patients, regardless of comorbidities, requiring elective or emergency surgery, and encompasses all intervention types, from minor to major. Fast and reliable pre-interventional risk stratification with PIRATE can help clinicians to make multidisciplinary informed shared decisions regarding pre-interventional optimization, peri-interventional monitoring and post-interventional patient care of the elderly population. PIRATE is an easy-to-use tool in clinical practice for patient admission but requires external validation for a routine implementation. One limitation of PIRATE include the reliance on a Cox model assuming the HR to be time-constant. In clinical data, verifying the PH assumption (or distributional assumptions as required for AFT models) is often challenging, and in many cases, the assumptions do not hold reliably. Modeling time-varying hazards to address violated PH assumptions within the Cox framework is challenging due to the lack of information on changes over time. Further, the development of PIRATE is based on a stepwise manual evaluation of risk factors and performance measures, making this development process hardly repeatable on other clinical datasets. All these considerations underscore the need for more flexible but interpretable modeling approaches for time-to-event outcomes with less restrictive assumptions.

The presented PRT and PVRF methods in Publications 2 and 3 combine pseudo-value regression and machine learning techniques, thereby avoiding distributional assumptions on the survival time or on the proportionality of HRs (Schenk et al., 2024; Schenk et al., 2025). The application of regression trees and gradient boosting in PRT ensures data-driven variable selection and the modeling of interactions (Hothorn et al., 2006; Bühlmann and Hothorn, 2007). The monotonic spline base-learner for the time component in the node-wise gradient boosting models additionally allow for capturing time-varying effects. As both the regression tree and the gradient boosting with simple linear base-learner produce interpretable estimates, their combination remains interpretable. Particularly, the terminal nodes of the regression

tree correspond to patient subgroups, with gradient boosting assigning interpretable additive models to each subgroup. PRT showed superior performance in estimating survival probabilities compared to GEE and other (machine-learning) methods in a comprehensive simulation study, especially in the presence of interactions or complex covariate structures (Schenk et al., 2024). Similar to PRT, the PVRF approach provides comparable advantages in modeling the RMST, leveraging the characteristics of pseudo-values and random forests (Breiman, 2001). Moreover, the proposed g-computation formula for RMST differences is particularly useful in non-randomized studies, when treatment effects need to be corrected for higher-dimensional sets of confounders, but can also be applied in RCTs, as demonstrated on the SUCCESS-A study data (Robins, 1986; Snowden et al., 2011; de Gregorio et al., 2020). This allows for the causal interpretation of treatment effects (assuming no unmeasured confounding), even though only the observed outcomes under either treatment or non-treatment, but not the counterfactual ones, are available for each patient. Numerical experiments including scenarios with interactions and time-varying treatment effects (i.e., violated PH assumption) show that the PVRF method provides accurate estimates of RMST values and RMST-based treatment effects and that the PVRF method outperforms modeling approaches like the Cox model, an AFT model or the GEE approach (Schenk et al., 2025). Using data of the SUCCESS-A study, it has been demonstrated that survival probability estimates derived by PRT and RMST estimates derived by PVRF, remain interpretable, either through the algorithm itself (PRT) or by calculating IML measures (PVRF) (de Gregorio et al., 2020). Notably, post-hoc IML measures explain covariate effects without relying on additional model assumptions. For instance, Shapley values leverage methods from game theory without assuming a specific modeling process (Molnar, 2022). The application of both methods on the SUCCESS-A data validated results previously reported in breast cancer literature and confirmed widely recognized treatment guidelines (Coates et al., 2015; Senkus et al., 2015). Deriving estimates for a new patient using PRT and PVRF is more complex than for the easily memorable PIRATE. However, implementing these methods as user-friendly applications, similar to the web-based PIRATE tool, would enable equally simple and accessible use in clinical practice.

The modeling approaches presented in this dissertation contribute to the expanding literature on pseudo-value regression combined with machine learning for time-to-event outcomes, highlighting both the flexibility of pseudo-value regression and the need for new modeling approaches of this type (Zhao and Feng, 2020; Zhao, 2020; Feng and Zhao, 2021; Rahman et al., 2021; Rahman and Purushotham, 2022). While PRT provides interpretable estimates without the need for IML techniques, many of these approaches include deep neural networks, falling into the category of hardly interpretable black-box models. Beyond that, deep neural networks typically require large datasets to achieve good performance and avoid overfitting. In contrast, PRT and PVRF are specifically designed to perform well on medium-sized datasets, which are commonly encountered in clinical studies. While deep neural networks can handle high-dimensional data, they can also process unstructured data like medical images, which is not yet possible with the methods discussed in this thesis. However, PVRF in particular offers new insights into the causal estimation of treatment effects, providing results that are easy to interpret and effectively support risk communication.

While the presented methods offer numerous advantages and perform well on simulated and clinical data, they also have limitations. For right-censored data, pseudo-value regression addresses the challenge of incompletely observed survival times by replacing them with pseudo-values calculated for the outcome of interest. For calculating of pseudo-values in PRT and PVRF, only the independent censoring assumption, required for the Kaplan-Meier estimator, is necessary. However, relating the pseudo-values to covariates, this assumption needs to be extended to the slightly stronger conditional random censoring assumption for consistent estimation of covariate effects (Graw et al., 2009; Overgaard et al., 2017). If this assumption is violated, the pseudo-value technique proposed by Overgaard et al. (2019) could be used to adapt the PRT and PVRF approaches appropriately. Similarly, pseudo-values can straightforwardly be adapted to discrete survival times T, provided a consistent estimator for the outcome of interest is available (Tutz and Schmid, 2016). Both PIRATE and PRT are designed to model survival probabilities at one or K time points, respectively. Future adaptions on PRT comprise the investigation of the optimal grid of time points to maximize performance.

These include investigations on the distance between the time points and the minimum number of time points required for reliable estimates. Similarly, the PVRF approach models the RMST for one time-horizon τ. One extension of PVRF could be to model the RMST on a grid of time horizons simultaneously and to investigate the performance on time horizons not used for model fitting (Zhong and Schaubel, 2022). This could be accomplished by applying an adapted PRT algorithm to pseudo-values for the RMST across a grid of time horizons. PRT is particularly suited for this task, as it is designed to handle multivariate inputs with time-varying effects. Taken together, these extensions would further enhance the flexibility of PRT and PVRF.

In general, time-to-event analysis is not restricted to analyzing one single event of interest, as considered in this dissertation. Instead, it covers a range of different scenarios, often occurring in clinical research. For example, within the competing risks setting, more than one event of interest, such as death from cancer or death from heart disease, are analyzed competitively. Another scenario is the illness-death model, representing a multi-state model estimating the probability of transitioning between three stages (disease-free, diseased and dead). Both the competing events and multi-state settings require special techniques to be analyzed. However, among others, pseudo-value approaches are available for these settings, for example based on the Aalen-Johansen estimator for the cumulative incidence function for the competing risks setting (Andersen and Pohar Perme, 2010). With this, the extension of methods like PRT and PVRF to competing risks or multi-state settings is straightforward (Andersen et al., 2003; Klein and Andersen, 2005; Andersen and Pohar Perme, 2010). All approaches discussed in this dissertation can handle right-censored time-to-event data, the most common type of censoring in clinical research. However, other censoring mechanisms, such as left- and interval-censoring, can also occur (Kalbfleisch and Prentice, 2002). Left-censoring arises when an event is known to have happened before a certain time but the exact timing is unknown. Interval-censoring is characterized by the event known to have happened within a specific time interval with fixed or random limits (Kalbfleisch and Prentice, 2002). The presented modeling approaches do not yet account for left- or interval-censoring.

As a final note, the methods presented in this dissertation are not limited to the application in clinical research. They can be readily applied to other research fields, including economics, social sciences, health services research, and environmental sciences. Here, common survival time examples include the time until a borrower defaults on a loan, the duration of unemployment before securing a job, the time until nursing home admission, or the progression of a wildfire to a specific location. All these examples involve right-censored time-to-event data, for which the presented methods are specifically designed.

4.1 Conclusion

The discussed approaches extend existing time-to-event models, providing interpretable estimates of survival probabilities, RMSTs, and causal treatment effects. PIRATE offers a fast and reliable risk assessment tool for the elderly which can be easily implemented in clinical practice. With less restrictive assumptions, PRT and PVRF offer a high flexibility to be applied to data from observational studies and RCTs. These approaches perform well in medium-sized datasets while automatically selecting main effects, interactions, non-linear, and time-varying effects from a given set of covariates, where standard methods often fail or require manual inclusion of prior knowledge. Thus, PRT and PVRF represent flexible alternatives to existing modeling approaches showing convincing performance and interpretability in different data situations.

4.2 References

Andersen PK, Klein JP, Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. In: Biometrika 2003; 90 (1): 15–27

Andersen PK, Pohar Perme M. Pseudo-observations in survival analysis. In: Statistical Methods in Medical Research 2010; 19 (1): 71–99

Breiman L. Random forests. In: Machine Learning 2001; 45 (1): 5–32

Bühlmann P, Hothorn T. Boosting algorithms: Regularization, prediction and model fitting. In: Statistical Science 2007; 22 (4): 477–505

Coates AS, Winer EP, Goldhirsch A, Gelber RD, Gnant M, Piccart-Gebhart MJ, Thürlimann B, Senn H. Tailoring therapies – improving the management of early breast cancer: St.Gallen international expert consensus on the primary therapy of early breast cancer 2015. In: Annals of Oncology 2015; 26: 1533–1546

de Gregorio A, Häberle L, Fasching PA, Müller V, Schrader I, Lorenz R, Forstbauer H, Friedl TWP, Bauer E, de Gregorio N, Deniz M, Fink V, Bekes I, Andergassen U, Schneeweiss A, Tesch H, Mahner S, Brucker SY, Blohmer JU, Fehm TN, Heinrich G, Lato K, Beckmann MW, Rack B, Janni W. Gemcitabine as adjuvant chemotherapy in patients with high-risk early breast cancer – results from the randomized phase III SUCCESS-A trial. In: Breast Cancer Research 2020; 22 (1): 111

Feng D, Zhao L. BDNNSurv: Bayesian deep neural networks for survival analysis using pseudo values. In: Journal of Data Science 2021; 19 (4): 542–554

Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. In: Lifetime Data Analysis 2009; 15 (2): 241–255

Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. In: Journal of Computational and Graphical Statistics 2006; 15 (3): 651–674

Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. Hoboken: Wiley, 2002

Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. In: Biometrics 2005; 61 (1): 223–229

Knaus WA, Zimmermann JE, Wagner DP, Draper EA, Lawrence DE. APACHE - Acute physiology and chronic health evaluation: A physiologically based classification system. In: Critical Care Medicine 1981; 9 (8): 591–597

Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. In: JAMA: The Journal of the American Medical Association 1993; 270 (24): 2957

Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R, Villers D. A simplified acute physiology score for ICU patients. In: Critical Care Medicine 1984; 12 (11): 975–977

Manach YL, Collins G, Rodseth R, Le Bihan-Benjamin C, Biccard B, Riou B, Devereaux P, Landais P. Preoperative score to predict postoperative mortality (POSPOM): Derivation and validation. In: Anesthesiology 2016; 124 (3): 570–579

Molnar C. Interpretable machine learning (second edition). A guide for making black box models explainable. Independently published, 2022

Overgaard M, Parner ET, Pedersen J. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. In: The Annals of Statistics 2017; 45 (5): 1988–2015

Overgaard M, Parner ET, Pedersen J. Pseudo-observations under covariate-dependent censoring. In: Journal of Statistical Planning and Inference 2019; 202: 112–122

POSE study group. Peri-interventional outcome study in the elderly in Europe: A 30-day prospective cohort study. In: European Journal of Anaesthesiology 2021; 39 (3): 198–209

Rahman MM, Matsuo K, Matsuzaki S, Purushotham S. DeepPseudo: Pseudo value based deep learning models for competing risk analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence 2021; 35 (1): 479–487

Rahman MM, Purushotham S. Pseudo value-based deep neural networks for multi-state survival analysis. In: arXiv 2022; technical report: 2207.05291 [cs.LG]

Robins J. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. In: Mathematical Modelling 1986; 7 (9): 1393–1512

Schenk A, Basten V, Schmid M. Modeling the restricted mean survival time using pseudo-value random forests. In: Statistics in Medicine 2025; 44 (5): e70031

Schenk A, Berger M, Schmid M. Pseudo-value regression trees. In: Lifetime Data Analysis 2024; 30 (2): 439–471

Schenk A, Kowark A, Berger M, Rossaint R, Schmid M, Coburn M, the POSE study group. Pre-Interventional Risk Assessment in The Elderly (PIRATE): Development of a scoring system to predict 30-day mortality using data of the Peri-Interventional Outcome Study in the Elderly. In: PLoS One 2023; 18 (12): e0294431

Senkus E, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rutgers E, Zackrisson S, Cardoso F, ESMO Guidelines Committee. Primary breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. In: Annals of Oncology 2015; 26: Suppl. 5, v8–v30

Snowden JM, Rose S, Mortimer KM. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. In: American Journal of Epidemiology 2011; 173 (7): 731–738

Tutz G, Schmid M. Modeling discrete time-to-event data. Basel: Springer International Publishing, 2016

Vazquez G, Benito S, Rivera R. Simplified acute physiology score III: A project for a new multidimensional tool for evaluating intensive care unit performance. In: Critical Care 2003; 7 (5): 345

Zhao L. Deep neural networks for predicting restricted mean survival times. In: Bioinformatics 2020; 36 (24): 5672–5677

Zhao L, Feng D. Deep neural networks for survival analysis using pseudo values. In: IEEE Journal of Biomedical and Health Informatics 2020; 24 (11): 3308–3314

Zhong Y, Schaubel DE. Restricted mean survival time as a function of restriction time. In: Biometrics 2022; 78 (1): 192–201

5 Acknowledgments

First, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Matthias Schmid, for his invaluable feedback, continuous support, and insightful discussions throughout my PhD journey at the IMBIE. I would like to thank Prof. Dr. Markus Neuhäuser for his support that has already began during my studies in Remagen. I am also thankful to Prof. Dr. Mark Coburn for the opportunity to establish a successful collaboration between the IMBIE and the department of anesthesiology and intensive care medicine. Furthermore, I would like to thank Prof. Dr. Nicole Ernstmann, for her guidance and valuable input beyond statistics.

A special thanks goes to Dr. Lothar Häberle and the SUCCESS-A study group for sharing the SUCCESS-A study data with us.

I sincerely thank my colleagues at the IMBIE and the department of anesthesiology and intensive care medicine, for their support, discussions, and encouragement throughout my PhD.

I am deeply grateful to my family and friends for their unconditional support, patience, and encouragement throughout this journey. Your belief in me has been a constant source of motivation. Special thanks to you, Peter, for always being there, providing love and strength when I needed it most. This achievement would not have been possible without you.

Complete publication list (during PhD)

Methodological research

Schenk A, Basten V, Schmid M. Modeling the restricted mean survival time using pseudo-value random forests. In: Statistics in Medicine 2025; 44 (5): e70031

Schenk A, Berger M, Schmid M. Pseudo-value regression trees. In: Lifetime Data Analysis 2024; 30 (2): 439–471

Schenk A, Kowark A, Berger M, Rossaint R, Schmid M, Coburn M, the POSE study group. Pre-Interventional Risk Assessment in The Elderly (PIRATE): Development of a scoring system to predict 30-day mortality using data of the Peri-Interventional Outcome Study in the Elderly. In: PLoS One 2023; 18 (12): e0294431

Projects anesthesia

Falay D, Schindler E, Mikus M, Boulos A, Schroth S, Schenk A, Baehner T. Ultrasound-guided supraclavicular cannulation of left brachiocephalic versus right internal jugular vein: Comparative analysis of central venous catheter-associated complications. In: Pediatric Anesthesia 2023; 33 (3): 219–228

Neumann C, Breil M, Schild A, Schenk A, Jakobs P, Mikus M, Schindler E. Central venous catheter tip positioning using ultrasound in pediatric patients - A prospective observational study. In: Pediatric Anesthesia 2024; 34 (6): 551–558

Neumann C, Kranenberg E, Schenk A, Kiefer N, Hilbert T, Klaschik S, Keyver-Paik MD, Soehle M. Influence of intraoperative fluid management on postoperative outcome and mortality of cytoreductive surgery for advanced ovarian cancer - A retrospective observational study. In: Healthcare 2024; 12 (12): 1218

Rehm C, Zoller R, Schenk A, Müller N, Strassberger-Nerschbach N, Zenker S, Schindler E. Evaluation of a paper-based checklist versus an electronic handover tool based on the situation background assessment recommendation (SBAR) concept in patients after surgery for congenital heart disease. In: Journal of Clinical Medicine 2021; 10 (24): 5724

Schenk A, Ende J, Hoch J, Güresir E, Grabert J, Coburn M, Schmid M, Velten M. A novel scoring system predicting red blood cell transfusion requirements in patients undergoing invasive spine surgery. In: Journal of Clinical Medicine 2024; 13 (4): 948

Strassberger-Nerschbach N, Magyaros F, Wittmann M, Ehrentraut H, Ghamari S, Schenk A, Neumann C, Schindler E. Quality comparison of remote anesthetic consultation versus on-site consultation in children with sedation for a magnetic resonance imaging examination - A randomized controlled trial. In: Pediatric Anesthesia 2023; 33 (8): 647–656

Other projects

Biener L, Vogelhuber J, Alboany H, Tiyerili V, Weber M, Linhart M, Becher MU, Schenk A, Nickenig G, Skowasch D, Pizarro C. Prevalence of sleep-disordered breathing in patients with mitral regurgitation and the effect of mitral valve repair. In: Sleep and Breathing 2023; 27 (2): 599–610

Goschzik T, Mynarek M, Doerner E, Schenk A, Spier I, Warmuth-Metz M, Bison B, Obrecht D, Struve N, Kortmann RD, Schmid M, Aretz S, Rutkowski S, Pietsch T. Genetic alterations of TP53 and OTX2 indicate increased risk of relapse in WNT medulloblastomas. In: Acta Neuropathologica 2022; 144 (6): 1143–1156

Stösser S, Kleusch L, Schenk A, Schmid M, Petzold GC. Derivation and validation of a screening tool for stroke-associated sepsis. In: Neurological Research and Practice 2023; 5 (1): 32