Advances in tree-based regression for modeling biomedical data

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

Nikolai Spuck
from Erbach (Odenwald)
2025

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

First reviewer: Prof. Dr. Matthias Schmid
Second reviewer: Prof. Dr. Andreas Groll

Day of oral examination: 16.06.2025

From the Institute of Medical Biometry, Informatics and Epidemiology

Table of Contents

List of abbreviations

CART	Classification and regression trees
CI	Confidence interval
LOS	Length of stay
OSCC	Oral squamous cell carcinoma
QoL	Quality of life
SHARE	Survey of Health, Ageing and Retirement in Europe
TSVC	Tree-structured varying coefficient

1 Abstract

Health-related research questions require methods that can deal with the growing complexity and dimensionality of biomedical data sets. A popular alternative to common parametric regression approaches are tree-based models, which recursively partition the data using binary splits to identify subgroups with similar values of an outcome variable of interest. The splitting rules (i.e., the splitting variables and corresponding split points) are selected in a data-driven way. Therefore, the data-driven tree building inherently performs variable selection and is able to detect and include relevant interactions even in high-dimensional data settings. In addition, tree-based models are easily accessible to practitioners due to their intuitive graphical representation.

This cumulative dissertation consists of four projects that aim to extend the class of tree-based models with a focus on application to biomedical research questions. In this vein, novel flexible tree-based approaches for modeling different types of biomedical data and a method for measuring statistical uncertainty and conducting inference on parameters from tree-based models are introduced. The first two projects focus on discrete time-to-event outcomes, which are common in biomedical research, for example, in observational studies, where the possible occurrence of an event of interest is only recorded at certain follow-up times. In the first project, a flexible approach for tree-based modeling of discrete time-to-event outcomes is proposed. In the second project project, a tree-based model for discrete time-to-event analysis is used to identify relevant risk factors for a prolonged length of stay in hospital for patients suffering from oral squamous cell carcinoma. The third project deals with the analysis of clustered data, where observations come in clusters of units, and the heterogeneity between observations from different units needs to be accounted for. A tree-based approach for modeling the effects of the covariates and the heterogeneity between the units with an application to quality of life in older adults is presented. The fourth project addresses the construction of confidence intervals for parameters from tree-based models. In particular, parameters of a tree-structured varying coefficient model are considered. Classical asymptotic normal distribution-based approaches for statistical inference on tree-structured varying

coefficients are invalid as they neglect the uncertainty induced by the data-driven tree building, which constitutes a so-called selective inference problem. To address this selective inference problem, a parametric bootstrap-based method for constructing confidence intervals for tree-structured varying coefficients is introduced. The performance of the methods proposed in the four projects is assessed in simulation studies, and applications to real-world data are considered.

Three research articles have been published in peer-reviewed international journals (*Sections 3.1, 3.2*, and *3.4*). In addition, an unpublished manuscript submitted to Advances in Data Analysis and Classification and available on arXiv is included in this dissertation (*Section 3.3*).

2 Introduction and aims with references

In biomedical research, statistical methods play a crucial role in analyzing various types of data from clinical and epidemiological studies. Therefore, researchers require methods that are able to handle the increasing complexity and dimensionality of biomedical data sets that, for example, arise from the improved imaging technologies in genetics (Chen et al., 2011) or the increasing number of blood parameters measured in modern laboratories in haematology (Gunčar et al., 2018). In particular, an effective identification of relevant variables from (high-dimensional) biomedical data sets is necessary to discover and confirm meaningful biological relationships, improve diagnostic validity, and optimize treatment strategies.

A key aspect of biomedical research is the analysis of time-to-event outcomes, which represent the time to the occurrence of a certain event of interest, such as death, disease progression, or discharge from hospital (Klein and Moeschberger, 2003). In the field of oral and maxillofacial surgery, for example, length of stay (LOS) in hospital serves as a key indicator for clinical severity and required healthcare resources. A prolonged LOS was even shown to be associated with an increased risk of complications and higher mortality (Pirson et al., 2018). Detecting risk factors for and identifying subgroups of patients with a prolonged LOS is therefore of great relevance.

Clustered data, where observations come in clusters of units, are also common in biomedical applications. For example, in multi-centric or cross-national studies, participants are clustered in multiple study centers or countries. In the Survey of Health, Ageing and Retirement in Europe (SHARE), quality of life (QoL) in older adults across 27 European countries and Israel was considered (Börsch-Supan et al., 2013; Bergmann et al., 2024; SHARE-ERIC, 2024). QoL plays an important role in assessing and guiding many health, social, community, and environmental policy decisions (Bowling and Stenner, 2011). Hence, it is of great interest to investigate which individual-level health-related and socioeconomic factors and interactions between them affect a person's QoL.

Tree-based models lend themselves to the two aforementioned research questions as they are able to inherently select relevant covariates, detect interactions, and construct subgroups

of observations that are similar with regard to the outcome (that is, probability of discharge from hospital in the context of oral and maxillofacial surgery patients and QoL in the context of the SHARE data). Tree-based models recursively partition the data into subsets that differ most strongly with regard to the outcome variable using binary splits. More specifically, in each step of the tree building, a parent node (starting with the root node containing all observations in step 1) is split into two child nodes based on some splitting rule, which comprises the variable selected for splitting and the corresponding split point. This process is repeated until a prespecified stopping criterion (for example, a minimal number of observations in a node or a maximal tree depth) is met. The optimal splitting rule in each step is commonly selected based on the Gini impurity or the information gain (Breiman et al., 1984). In each resulting terminal node (also called leaf node), an aggregated measure (for example, the mean) is determined based on the observations in that node. The concept of tree-based modeling originates from the *classification and regression tree* (CART) algorithm by Breiman et al. (1984). Following their idea, various extensions and alternatives have been proposed (for an overview, see Strobl et al., 2009). Due to the data-driven tree building, tree-based models do not require a fixed prespecified model formula that postulates all covariates and interactions to be included before model fitting, unlike classical, parametric regression approaches. Instead, the tree building procedure inherently facilitates variable selection as well as the detection and inclusion of relevant interactions. In addition, tree-based models offer an intuitively interpretable graphical representation and are simulatable (Murdoch et al., 2019). Yet, a major drawback of tree-based approaches is that classical methods for statistical inference on the model parameters are invalid (Neufeld et al., 2022). As the parameters of a tree-based model arise from a data-driven tree building, methods for statistical inference need to take the uncertainty induced by this model selection step into account. Conducting inference on these parameters is a so-called *selective inference* or *post-selection inference* problem as statistical inference after a data-driven model selection is of interest (Berk et al., 2013; Fithian et al., 2014). The concept of statistical inference is essential for most biomedical applications, in particular, for confirming the beneficial effect of a novel treatment method (for

example, the effect of antibody treatment in patients with COVID-19 on their need for oxygen support). To enhance the applicability of tree-based models for the analysis of biomedical data, methods for quantifying uncertainty and assessing statistical significance of the model parameters are required.

2.1 Thesis outline

This cumulative dissertation comprises three publications and one unpublished manuscript that focus on the class of tree-based models and their application to biomedical data. Specifically, the aim is to develop novel tree-based modeling approaches for different types of data and a method for conducting statistical inference on parameters from tree-based models.
The first two articles deal with tree-based models for discrete time-to-event analysis. In particular, a novel tree-based approach for modeling discrete event times is introduced (*Publication 1* in Section 3.1), and a tree-based discrete hazard model is applied to investigate risk factors for a prolonged LOS in hospital in patients suffering from oral squamous cell carcinoma (OSCC; *Publication 2* in Section 3.2). In the *Unpublished Manuscript* in Section 3.3, a tree-based approach for modeling clustered data with an application to QoL in older adults is presented. Finally, the third publication addresses the selective inference problem for parameters from tree-based models using a parametric bootstrap approach to construct confidence intervals (CIs) for *tree-structured varying coefficients* (TSVCs; Berger et al., 2019; *Publication 3* in Section 3.4).
The appendix includes the complete list of publications resulting from projects at the Institute for Medical Biometry, Informatics and Epidemiology during the years of this PhD.

2.1.1 Tree-based modeling of discrete time-to-event data

The analysis of time-to-event outcomes requires methods that are able to handle the challenges specific to this type of data. Event times frequently follow a highly skewed distribution and, even more importantly, are usually subject to censoring. That is, the event times are not fully observed for all individuals. In clinical studies, for example, participants are only ob-

served for a limited amount of time before leaving the study, and therefore the event of interest may not be recorded for all participants.

Most popular models for time-to-event analysis, such as the proportional hazards model by Cox (1972), assume that time is measured on a continuous scale. In many applications, however, event times are measured on a discrete scale, where the exact time of the event is not reported, but only a time interval during which the event occurred. In these cases, one speaks of discrete time-to-event or interval censored data. The *logistic discrete hazard model* is the most widely applied approach for analyzing discrete event times (Tutz and Schmid, 2016). In the classical parametric logistic discrete hazard model, the effects of the covariates are assumed to be linear as well as independent of each other and time.

In *Publication 1*, a novel extension of the logistic discrete hazard model is introduced, where (part of) the parametric predictor function is replaced by a tree. The proposed framework allows modeling and interpretation of the effects of the covariates separate from the effects of time, similarly to the classical parametric model. Furthermore, the *survival tree* by Schmid et al. (2016), which is able to incorporate interactions between covariates and time, can also be fitted within the proposed modeling framework. Predictive performance and variable selection rates of the proposed models are compared with alternative approaches in a simulation study. The models are illustrated based on applications to data of patients suffering from acute odontogenic infection and data of patients suffering from lymphatic filariasis. *Publication 2* presents an application of the survival tree to data from patients suffering from OSCC. Here, the objective is to identify risk factors for a prolonged LOS in hospital after surgery.

2.1.2 Tree-based modeling of clustered data

In clustered data, observations within units are likely to be more similar than observations between units. Therefore, regression approaches for clustered data need to account for this heterogeneity between the units.

The classical *linear mixed effects model* assumes linear effects of the covariates and includes unit-specific random effects, which commonly follow a normal distribution, to account for the

heterogeneity between the units in a parsimonious way (Verbeke and Molenberghs, 2000). Alternative random effects-based approaches that apply tree structures to facilitate inherent variable selection and detection of interactions between the covariates were proposed by Hajjem et al. (2011) and Sela and Simonoff (2012). To avoid the distributional assumption required for the random effects and still enable parsimonious modeling of unit-specific effects, Berger and Tutz (2018) introduced a fixed effects model with linear effects of the covariates that applies tree-structured clustering of units with similar effects on the outcome to account for the heterogeneity.

In the *Unpublished Manuscript*, a novel tree-based approach for modeling clustered data is proposed, combining the ideas of Hajjem et al. (2011), Sela and Simonoff (2012), and Berger and Tutz (2018). Specifically, the proposed model consists of two tree structures, where one tree captures the effects of the covariates, and the other identifies clusters of units with similar effects on the outcome. The proposed model is applied to analyze QoL in SHARE, and goodness of fit as well as variable selection rates are assessed in a simulation study.

2.1.3 Selective inference for tree-based models

Conducting inference on parameters from tree-based models, which arise from a data-driven tree building, is a selective inference problem. Neufeld et al. (2022) proposed a selective inference framework for regression trees based on a truncated normal distribution. For more complex tree-based models, however, alternative approaches for selective inference are required.

Publication 3 addresses this issue in the context of TSVC models proposed by Berger et al. (2019). TSVC models are based on the varying coefficient models originally introduced by Hastie and Tibshirani (1993) and apply recursive partitioning to capture varying effects of the covariates, potentially resulting in several tree structures as part of a single model. More specifically, the coefficient of each covariate may be determined by a tree structure. To address the complex selective inference problem for TSVCs, a parametric bootstrap approach for constructing percentile CIs is introduced. Coverage proportions of the proposed CIs are

considered in a simulation study. For illustration, data of patients suffering from COVID-19 and of patients suffering from acute odontogenic infection are analyzed.

2.2 References

Berger M, Tutz G. Tree-structured clustering in fixed effects models. Journal of Computational and Graphical Statistics 2018; 27: 380–392

Berger M, Tutz G, Schmid M. Tree-structured modelling of varying coefficients. Statistics and Computing 2019; 29: 217–229

Bergmann M, Wagner M, Börsch-Supan A. SHARE Wave 9 Methodology: From the SHARE Corona Survey 2 to the SHARE Main Wave 9 Interview. Munich: SHARE-ERIC, 2024

Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. The Annals of Statistics 2013; 41: 802–837

Börsch-Supan A, Brandt M, Hunkler C, Kneip T, Korbmacher J, Malter F, Schaan B, Stuck S, Zuber S. Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). International Journal of Epidemiology 2013; 42: 992–1001

Bowling A, Stenner P. Which measure of quality of life performs best in older age? A comparison of the OPQOL, CASP-19 and WHOQOL-OLD. Journal of Epidmiology and Community Health 2011; 63: 273–280

Breiman L, Friedman JH, Olshen RA, Stone JC. Classification and Regression Trees. Monteterey, CA Wadsworth: Taylor and Francis, 1984

Chen X, Wang M, Zhang H. The use of classification trees for bioinformatics. Data Mining and Knowledge Discovery 2011; 1: 55–63

Cox DR. Regression Models und Life-Tables. Journal of the Royal Statistical Society Series B: Statistical Methodology 1972; 34: 187–202

Fithian W, Sun D, Taylor J. Optimal Inference after Model Selection. Uploaded to arXiv 2014; 1410.2597

Gunčar G, Kukar M, Notar M, Brvar M, Černelč P, Notar M, Notar M. An application of machine learning to haematological diagnosis. Scientific Reports 2018; 8: 1–12

Hajjem A, Bellavance F, Larocque D. Mixed effects regression trees for clustered data. Statistics & Probability Letters 2011; 81: 451–459

Hastie T, Tibshirani R. Varying-Coefficient Models. Journal of the Royal Statistical Society Series B: Statistical Methodology 1993; 55: 757–779

Klein J, Moeschberger M. The Statistical Analysis of Failure Time Data. New York: Springer, 2003

Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 2019; 116: 22071–22080

Neufeld AC, Gao LL, Witten DM. Tree-Values: Selective Inference for Regression Trees. Journal of Machine Learning and Research 2022; 23: 1–43

Pirson M, Debanne F, Van den Bulcke J, Leclerecq P, Martins D, De Wever A. Evaluation of cost and length of stay, linked to complications associated with major surgical procedures. Acta clinica Belgica 2018; 73: 40–49

Schmid M, Küchenoff H, Hörauf A, Tutz G. A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. Statistics in Medicine 2016; 35: 734–751

Sela RJ, Simonoff JS. RE-EM trees: A data mining approach for longitudinal and clustered data. Machine Learning 2012; 86: 169–207

SHARE-ERIC. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 9. Release version: 9.0.0. Data set, 2024

Strobl C, Malley J, Tutz G. An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychological Methods 2009; 14: 323–348

Tutz G, Schmid M. Modeling Discrete Time-to-Event-Data. New York: Springer, 2016

Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York: Springer, 2000

3 Publications

3.1 Publication 1: Flexible tree-structured regression models for discrete event times

Spuck N, Schmid M, Heim N, Klarmann-Schulz U, Hörauf A, Berger M. Flexible tree-structured regression models for discrete event times. Statistics and Computing 2023; 33: 1-21
https://doi.org/10.1007/s11222-022-10196-x

Supplementary data can be found at:
https://doi.org/10.1007/s11222-022-10196-x

Statistics and Computing (2023) 33:20
https://doi.org/10.1007/s11222-022-10196-x

ORIGINAL PAPER

Flexible tree-structured regression models for discrete event times

Nikolai Spuck[1] · Matthias Schmid[1] · Nils Heim[2] · Ute Klarmann-Schulz[3,4] · Achim Hörauf[3,4] · Moritz Berger[1]

Received: 13 April 2022 / Accepted: 9 December 2022 / Published online: 21 December 2022

Abstract
Discrete hazard models are widely applied for the analysis of time-to-event outcomes that are intrinsically discrete or grouped versions of continuous event times. Commonly, one assumes that the effect of explanatory variables on the hazard can be described by a linear predictor function. This, however, may be not appropriate when non-linear effects or interactions between the explanatory variables occur in the data. To address this issue, we propose a novel class of discrete hazard models that utilizes recursive partitioning techniques and allows to include the effects of explanatory variables in a flexible data-driven way. We introduce a tree-building algorithm that inherently performs variable selection and facilitates the inclusion of non-linear effects and interactions, while the favorable additive form of the predictor function is kept. In a simulation study, the proposed class of models is shown to be competitive with alternative approaches, including a penalized parametric model and Bayesian additive regression trees, in terms of predictive performance and the ability to detect informative variables. The modeling approach is illustrated by two real-world applications analyzing data of patients with odontogenic infection and lymphatic filariasis.

Keywords Discrete time · Hazard models · Non-parametric regression · Recursive partitioning · Time-to-event analysis

1 Introduction

The terms time-to-event data or survival data refer to data sets where the outcome variable corresponds to the time to the occurrence of a certain event of interest. In clinical research, for example, the time to death, the time to disease progression or the duration of hospitalization are widely applied time-to-event outcomes (Klein and Moeschberger 2003).

When building a regression model that relates the event time T to a set of explanatory variables $\boldsymbol{x} = (x_1, \ldots, x_p)$ one must account for certain characteristics that are unique to time-to-event data. In particular, event times are usually subject to censoring, that is, the event of interest is not observed for all individuals under study. An appropriate approach for modeling time-to-event outcomes is to consider the hazard for the occurrence of an event at time t given by $\xi(t) = \lim_{\Delta t \to 0}\{P(t < T \leq t + \Delta t | \, T > t, \boldsymbol{x})/\Delta t\}$. The most popular hazard model is the *Cox proportional hazards model* by Cox (1972), which assumes that the effects of the explanatory variables on the hazard are constant over time. An alternative that does not require the proportional hazards assumption are *accelerated failure time models* (Kalbfleisch and Prentice 2002). This class of models is not dealt with here, but we refer to Kuss and Hoyer (2021) for recent developments in this field.

The classical Cox model assumes that the effects of the explanatory variables can be described by a parametric term (that is, by a linear function of $\boldsymbol{x}$). This assumption, which is too restrictive in many applications, can be relaxed by the specification of smooth, non-linear functions (Sleeper and Harrington 1990; LeBlanc and Crowley 2004) or the use of *recursive partitioning techniques* also called *tree-structured modeling*, see, for example, Segal (1997), Zhang and Singer

✉ Nikolai Spuck
spuck@imbie.uni-bonn.de

1 Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

2 Department of Oral and Cranio-Maxillo and Facial Plastic Surgery, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

3 Institute of Medical Microbiology, Immunology and Parasitology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

4 German Center for Infection Research (DZIF), Partner Site Bonn-Cologne, 53127 Bonn, Germany

(1999), and Bou-Hamad et al. (2011a). Tree-structured models are strong tools that relate the explanatory variables to the outcome in a non-linear way and automatically detect relevant interactions between the explanatory variables if they are present. The visualization as a hierarchical tree makes the model easily accessible for practitioners and simulatable (Murdoch et al. 2019). In addition, tree-structured models inherently perform variable selection, which is particularly useful in high-dimensional settings. Within the scope of hazard models, Gordon and Olshen (1985) formerly extended classification and regression trees (CART, Breiman et al. 1984) and proposed to partition time-to-event data based on different measures of distance between two survival curves. Segal (1995) illustrated the application of tree-structured modeling for the analysis of HIV patient data, and Carmelli et al. (1997) applied tree-structured time-to-event analysis to investigate the relationship between obesity and mortality from coronary heart disease and cancer. More recently, Wallace (2014) applied conditional inference trees (Hothorn et al. 2006) to model time-to-event data, and Rancoita et al. (2016) proposed to use a Bayesian network for survival tree analysis with missing data.

All of the aforementioned methods have in common that they assume time to be measured or approximated by a *continuous* scale. In many applications, however, the event times are intrinsically discrete, or the exact continuous event times were not recorded and it is only known that the events occurred in a certain time interval (this is also referred to as the case of *grouped event times*). Grouped event times are typically observed in clinical and epidemiological studies with fixed prespecified follow-up times. Then time is recorded on a *discrete* time scale $t = 1, \ldots, k$. Two examples, which will be dealt with here, are days of hospitalization after jaw surgery in patients with acute odontogenic infection and time to an acute attack caused by infections through skin lesions in patients with lymphatic filariasis measured in months. In these cases, in which grouping effects are present, the application of statistical models designed for continuous time-to-event data is considered inappropriate by many authors (e.g. Tutz and Schmid (2016) and Berger and Schmid (2018)). Therefore, we consider the class of discrete hazard models that has prevailed for the analysis of discrete time-to-event outcomes (Willet and Singer 1993; Hashimoto et al. 2011; Tutz and Schmid 2016). A comprehensive introduction to parametric discrete hazard models and semi-parametric extensions was recently given by Berger and Schmid (2018). In this article, we propose novel alternatives to the widely used class of parametric discrete hazard models. More specifically, we propose different non-parametric extensions, where (part of) the parametric term is replaced by a tree structure while the common additive form of the predictor function is kept. As will be illustrated, the models are highly flexible and combine the advantages of classical parametric and tree-structured discrete hazard models.

Tree-structured models for discrete time-to-event analysis were so far proposed by Bou-Hamad et al. (2009) and Bou-Hamad et al. (2011b). Their method first grows a tree and then fits a covariate-free discrete hazard model in each terminal node separately. Schmid et al. (2016) suggested a CART approach where both the explanatory variables and the time t are considered as candidates for tree building. Sparapani et al. (2016) expanded the Bayesian additive regression tree (Chipman et al. 2010) for binary outcomes to time-to-event outcomes considering grouped survival times. In addition, Tiendrébéogo et al. (2019) applied a model-based recursive partitioning approach based on the algorithm by Hothorn et al. (2006) to HIV patient data to identify characteristics that are associated with risk of death. For overviews on existing tree-structured methods for discrete-time hazard modeling and extensions for dynamic predictions, see also Kretowska (2019) and Moradian et al. (2021).

Our proposed approach differs from previous methods, as we do not apply a traditional recursive partitioning algorithm, but fitting and tree building is performed within the classical framework of additive discrete hazard models. When growing the trees, in each step the best split is selected among all current non-internal and non-terminal nodes, yielding a sequence of nested subtrees of different size.

The remainder of this article is structured as follows: In Sect. 2 the notation and general methodology for the analysis of discrete time-to-event data are described. In Sect. 3 we propose three different novel tree-structured regression models for time-to-event outcomes. The performance of the different models was assessed in a simulation study, which is presented in Sect. 4. In Sect. 5, we show the results of two applications using the proposed models to analyze data of patients with odontogenic infection and lymphatic filariasis. Finally, our findings and conceptual aspects are discussed in Sect. 6.

2 Notation and methodology

In the following, let T_i denote the event time and C_i denote the censoring time of individual i, $i = 1, \ldots, n$. We assume that T_i and C_i are independent and take discrete values in $\{1, \ldots, k\}$. It is further assumed that the censoring mechanism is non-informative for T_i, in the sense that C_i does not depend on any parameters used to model the event time. Considering right-censored data, the observation time for individual i is given by $\tilde{T}_i = \min(T_i, C_i)$, and $\Delta_i := I(T_i \leq C_i)$ indicates whether the event of individual i was observed ($\Delta_i = 1$) or not ($\Delta_i = 0$). In case continuous time-to-event data has been grouped, the discrete event times $t = 1, \ldots, k$ refer to time intervals

$[0, a_1), [a_1, a_2), \ldots, [a_{k-1}, \infty)$, where $T_i = t$ means that the event occurred in time interval $[a_{t-1}, a_t)$.

An essential tool for modeling discrete time-to-event data is the *discrete hazard function*. For time-constant explanatory variables $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})$ it is given by

$$\lambda(t|\, \boldsymbol{x_i}) = P(T_i = t|\, T_i \geq t, \boldsymbol{x_i}),\ t = 1, \ldots, k, \tag{1}$$

which is the conditional probability for the occurrence of an event at time t given that the event has not occurred until t. From (1) it follows that the *discrete survival function*, which denotes the probability that an event occurs after time t, can be written as

$$S(t|\, \boldsymbol{x_i}) = P(T_i > t|\, \boldsymbol{x_i}) = \prod_{s=1}^{t}(1 - \lambda(s|\, \boldsymbol{x_i})). \tag{2}$$

Based on the definition of the hazard function in (1), for a fixed time t, the discrete hazard represents a binary variable that specifies whether an event occurs at time t or not, given that $T_i \geq t$. Hence, strategies for modeling binary outcome data can be adapted to discrete hazard modeling.

A class of regression models that relates the discrete hazard function to the explanatory variables $\boldsymbol{x_i}$ is defined by

$$\lambda(t|\, \boldsymbol{x_i}) = h(\eta(t, \boldsymbol{x_i})), \quad t = 1, \ldots, k-1, \tag{3}$$

where $h(\cdot)$ is a strictly increasing distribution function and $\eta(\cdot)$ is a real-valued predictor function depending on the explanatory variables and time. A commonly assumed semi-parametric form of the predictor function is given by

$$\eta(t, \boldsymbol{x_i}) = \gamma_0(t) + \boldsymbol{x_i}^\top \boldsymbol{\gamma}, \tag{4}$$

where $\gamma_0(t)$ describes the hazard over time (for any given values of $\boldsymbol{x_i}$, called *baseline hazard*) usually by the use of dummy variables for each time point or a smooth, non-linear function. The effects $\boldsymbol{\gamma} \in \mathbb{R}^p$ of the explanatory variables on the hazard are assumed to be linear and independent of time. Using the logistic distribution function for $h(\cdot)$, Equation (3) yields the *logistic discrete hazard model*

$$\lambda(t|\, \boldsymbol{x_i}) = \frac{\exp(\eta(t, \boldsymbol{x_i}))}{1 + \exp(\eta(t, \boldsymbol{x_i}))}, \tag{5}$$

which is also called the *proportional continuation ratio model* (cf. Tutz and Schmid 2016, Chapter 3). The continuation ratio at time t is given by

$$\Psi(t|\, \boldsymbol{x_i}) = \frac{P(T = t|\, \boldsymbol{x_i})}{P(T > t|\, \boldsymbol{x_i})} = \exp\left(\gamma_0(t) + \boldsymbol{x_i}^\top \gamma\right)$$

and denotes the ratio comparing the probability of an event at time t to the probability of an event after t. Comparing the continuation ratios of two individuals u and v at time t yields

$$\frac{\Psi(t|\, \boldsymbol{x_u})}{\Psi(t|\, \boldsymbol{x_v})} = \exp\left((\boldsymbol{x_u} - \boldsymbol{x_v})^\top \boldsymbol{\gamma}\right).$$

Hence, proportionality is given in the sense that the comparison of two individuals with regard to their continuation ratios is independent of time. This facilitates interpretation of the estimated effects (see our applications in Sect. 5). For the remainder of this article, we will focus on the model using the logistic distribution function and refer to Sect. 6 for a discussion on the characteristics of models with other link functions.

As mentioned above, discrete hazard models can be fitted by means of binary response models. This is because the log-likelihood of Model (3) can be expressed as

$$\ell = \sum_{i=1}^{n}\sum_{t=1}^{\tilde{T}_i}(1 - y_{it}) \log(1 - \lambda(t|\boldsymbol{x_i})) + y_{it} \log(\lambda(t|\boldsymbol{x_i})), \tag{6}$$

which is equivalent to the log-likelihood of a binomial model with independent observations y_{it}. In order to derive coefficient estimates, one has to define the binary outcome values for each individual i as

$$(y_{i1}, \ldots, y_{i\tilde{T}_i}) = \begin{cases} (0, \ldots, 0, 1), & \text{if } \Delta_i = 1, \\ (0, \ldots, 0, 0), & \text{if } \Delta_i = 0. \end{cases}$$

Hence, before fitting the model with software for binary outcomes, the original time-to-event data has to be converted into an *augmented data matrix* that contains the binary outcome values. This results in a design matrix with $\tilde{T}_i$ rows for each individual i, where the vector of explanatory variables is repeated row-wise, and with a total number of $\tilde{n} = \sum_{i=1}^{n} \tilde{T}_i$ rows. Further details on data preparation and estimation of discrete hazard models can be found in Berger and Schmid (2018). Note that, in the following, the term *individual* will be used to refer to one row in the original, non-augmented data matrix, and the term *observation* to refer to one row of the augmented data matrix.

3 Tree-structured discrete hazard models

The logistic discrete hazard model with predictor function (4) assumes that the effect on the hazard can be described by a linear combination of the explanatory variables. This may be too restrictive, in particular when non-linear effects or interactions between the explanatory variables are present. To address this issue, we propose to incorporate tree-based splits into the predictor function of the discrete hazard model (3).

Specifically, either the effects of the explanatory variables $\boldsymbol{x}$ or the effects of $\boldsymbol{x}$ as well as the effect of the time t on the hazard are replaced by a tree structure.

3.1 Smooth baseline coefficients

A tree-structured predictor function that allows for more flexibility, but still preserves the additive structure of the semi-parametric model (4) has the form

$$\eta(t, \boldsymbol{x_i}) = \gamma_0(t) + tr(\boldsymbol{x_i}), \tag{7}$$

where the function $tr(\cdot)$ is determined by a common tree structure. This means that the function $tr(\cdot)$ sequentially partitions the observations into disjoint subsets $N_m, m = 1, .., M$, based on the values of the explanatory variables and assigns a regression coefficient γ_m to each subset N_m. Hence, this function can be written as

$$tr(\boldsymbol{x_i}) = \sum_{m=1}^{M} \gamma_m I\left(\boldsymbol{x_i} \in N_m\right), \tag{8}$$

where $I(\cdot)$ denotes the indicator function. When growing the tree, analogously to CART a binary split partitioning the observations of one parental node into two child nodes is performed in each step (cf. Hastie et al. 2009).

The function $\gamma_0(t)$ in Model (7) is determined by a *P-spline* (Eilers and Marx 1996). More precisely, a number of *B-spline* basis functions (de Boor 1978) are used, and a term to penalize differences between adjacent coefficients is included in the likelihood function. Then the coefficients are fitted by maximizing the penalized log-likelihood

$$\ell_p = \ell - \varepsilon J, \tag{9}$$

where $\varepsilon \in \mathbb{R}^+$ is a penalty parameter and $J \in \mathbb{R}^+$ is the penalty term preventing the estimated function from becoming too rough. When using a P-spline, J is a difference penalty on adjacent B-spline coefficients. For details, see Wood (2011, 2017).

The tree $tr(\cdot)$ is constructed in a stepwise procedure, starting from the model with baseline coefficients, only. Then the first split yields a model with predictor

$$\eta^{[1]}(t, \boldsymbol{x_i}) = \gamma_0(t) + \gamma_1^{[1]} I(x_{ij} \leq c_j), \tag{10}$$

where x_j is the explanatory variable selected for the first split, c_j is the corresponding split point, and $\gamma_1^{[1]}$ is the effect for the first subset. Note that the second node defined by $I(x_{ij} > c_j)$ needs to serve as reference node with $\gamma_2^{[1]} := 0$ to ensure parameter identifiability. Secondly, a different or the same variable and a corresponding split point is selected to further split one of the current nodes. Assuming that the left node is split based on variable x_k with split point c_k yields the predictor

$$\begin{aligned}\eta^{[2]}(t, \boldsymbol{x_i}) =& \gamma_0(t) + \left[\gamma_1^{[2]} I(x_{ij} \leq c_j \wedge x_{ik} \leq c_k)\right. \\ &\left. + \gamma_2^{[2]} I(x_{ij} \leq c_j \wedge x_{ik} > c_k)\right],\end{aligned} \tag{11}$$

where $\gamma_1^{[2]}$ and $\gamma_2^{[2]}$ are the effects for the two new subsets built after the second split. Further splits are performed analogously until a predefined stopping criterion is met (see Sect. 3.4 for details). The design of the augmented data matrices for fitting the model with predictor (10) is illustrated in Online Resource Supplement 1, see Equations (S1) and (S2).

3.2 Piecewise constant baseline coefficients

Model (7) allows for non-linear effects of the explanatory variables as well as (higher-order) interactions between them. Moreover, the effects of the explanatory variables can easily be interpreted from the tree structure. Modeling the baseline coefficients by a smooth (P-spline) function may, however, not be adequate, as abrupt changes of the effect strength appear particularly plausible when considering discrete event times (Puth et al. 2020). Therefore, in the following, we assume that the baseline hazard does not vary over the whole range of t, but is constant within several time intervals.

An alternative to model (7) that also preserves the additive structure of parametric discrete hazard models, but allows to fit *piecewise constant* baseline coefficients is a model with predictor

$$\eta(t, \boldsymbol{x_i}) = tr_0(t) + tr(\boldsymbol{x_i}), \tag{12}$$

where $tr_0(\cdot)$ is a second tree partitioning the observations into subsets $N_{0m_0}, m_0 = 1, \ldots, M_0$, with regard to the time t, and assigning a parameter γ_{0m_0} to each subset N_{0m_0}, and $tr(\cdot)$ is a tree structure defined as in (8). More formally, we have that

$$tr_0(t) = \sum_{m_0=1}^{M_0} \gamma_{0m_0} I\left(t \in N_{0m_0}\right), \tag{13}$$

which represents a piecewise constant function over time. Both trees $tr_0(\cdot)$ and $tr(\cdot)$ are constructed using a similar stepwise procedure as described in Sect. 3.1. Assuming now that a split in t at c_t is selected in the first step yields the predictor

$$\eta^{[1]}(t, \boldsymbol{x_i}) = \left[\gamma_{01}^{[1]} I(t \leq c_t) + \gamma_{02}^{[1]} I(t > c_t)\right]. \tag{14}$$

Then, a split in explanatory variable x_j at split point c_j, in the second step, results in a predictor of the form

$$\eta^{[2]}(t, \boldsymbol{x_i}) = \left[\gamma_{01}^{[2]} I(t \leq c_t) + \gamma_{02}^{[2]} I(t > c_t)\right]$$

$$+ \gamma_1^{[2]} I(x_{ij} \leq c_j), \tag{15}$$

where the node $I(x_{ij} > c_j)$ serves as reference node. In each iteration of the algorithm (see Sect. 3.4 for details), a split in either t or in one of the explanatory variables is performed, expanding $tr_0(\cdot)$ or $tr(\cdot)$, respectively. As a result, the predictor comprises piecewise constant effects over time and tree-structured effects of the explanatory variables. Note that the algorithm treats t as an ordinal variable. Thus, the total number of possible splits in $tr_0(\cdot)$ is $k-2$. When treating t as nominal variable instead, certain forms of the effects of time (e.g. U-shapes) might be detected more easily, particularly in small samples. Yet, this approach would be much more computationally expensive, as the number of possible splits grows exponentially with the number of time points. More specifically, $2^{(k-1)-1} - 1$ possible binary partitions would have to be considered for the first split in $tr_0(\cdot)$. The design of the augmented data matrices for fitting the model with predictor (15) is illustrated in Online Resource Supplement 1, see Equations (S3) and (S4).

3.3 Modeling the effects of t and x by one single tree

The proposed models (7) and (12) are an attractive choice, as the effects can be captured in a very flexible way and the representation as a tree facilitates their interpretation. At the same time the common additive structure of parametric models (separating the effects of $\boldsymbol{x}$ and t) is kept.

If one suspects that interactions between time and explanatory variables are present, the use of a discrete hazard model with predictor

$$\eta(t, \boldsymbol{x_i}) = tr(t, \boldsymbol{x_i}), \tag{16}$$

may be more appropriate. Here, the function $tr(\cdot, \cdot)$ partitions the observations based on the time t and the values of the explanatory variables $\boldsymbol{x_i}$, that is,

$$tr(t, \boldsymbol{x_i}) = \sum_{m=1}^{M} \gamma_j I\left((t, \boldsymbol{x_i}) \in N_m\right). \tag{17}$$

The construction of the tree $tr(\cdot, \cdot)$ is performed in the same manner as described in Sect. 3.1, but now in each step the time t (treated as an ordinal variable) and the explanatory variables are treated together. Essentially, this modeling approach is equivalent to the *survival tree* proposed by Schmid et al. (2016), as in both cases a single tree is built, where the time t and the explanatory variables are simultaneously considered as candidates for splitting. Schmid et al. (2016) proposed to use the Gini impurity measure for the selection of split points, whereas our approach selects the splits based on likelihood ratio (LR) test statistics (see Sect. 3.4), which is equivalent to the entropy criterion (Breiman et al. 1984). If the same splitting criterion is chosen, the model with predictor (16) and the survival tree yield the same results.

Model (16) is highly flexible, as an interaction between $\boldsymbol{x}$ and t implies the presence of time-varying effects on the hazards. This, however, comes at the price that the tree structure is harder to interpret, because each terminal node corresponds to a subset defined by the explanatory variables and to a time interval. An example of the augmented data matrices for fitting Model (16) is given in Equations (S5) and (S6) in Online Resource Supplement 1.

3.4 Fitting procedure

In each step of the tree-building algorithm, the best split among all candidate variables (that is, one component of $\boldsymbol{x}$ or t) and among all possible split points is selected. In order to do so, the two parameters corresponding to the two subsets resulting from the new split, $\gamma_q^{[\ell]}$ and $\gamma_{q+1}^{[\ell]}$ (where q is the current number of terminal nodes and $\ell = q - 1$ is the current number of splits), are tested for equivalence. More specifically, one examines all the null hypotheses H_0: $\gamma_q^{[\ell]} = \gamma_{q+1}^{[\ell]}$ against the alternatives H_1: $\gamma_q^{[\ell]} \neq \gamma_{q+1}^{[\ell]}$ by means of likelihood ratio (LR) tests. For a model with predictor (7), the splitting variable x_q and split point c_q related to the largest LR test statistic are selected. For models with predictor (12) and (16), the split is also selected based on the largest LR test statistic but can either be in t or in one of the explanatory variables (in the second or the same tree structure). Note that in each step of the algorithm all observations (of the augmented data matrix) are used to derive estimates of the model coefficients. That means, all parameters are refitted in each iteration and no previously estimated parameters are kept. Consequently, in case of the models with two components (introduced in Sects. 3.1 and 3.2), the parameter estimates of either of the two components are adjusted for the change through a split in the other. For the model with only one tree component described in Sect. 3.3, an additional split, however, does not change the parameter estimates in the remaining nodes (i.e. there would be no necessity to consider all the observations). Hence, the mechanism is equivalent to the fitting of a common tree, where only the subset of all observations in one node are used to guide the next split.

Three approaches for determining the size of the tree(s) are considered:

(i) The first alternative, which is based on the algorithm proposed by Berger et al. (2019), applies *permutation tests*. Let x_q be the variable selected for splitting. Then, the p-value obtained from the distribution of the maximally selected test statistic $T_q = \max_{c_q} T_{q,c_q}$ provides a measure for the dependence between the time-to-event outcome and variable x_q while accounting for the number of possible split points (Hothorn and Lausen 2003). Therefore, one explic-

itly accounts for the involved multiple testing problem with regard to the number of split points for x_q. The construction of the tree(s) is terminated when the null hypothesis of independence between the time-to-event outcome and the selected explanatory variable (or the time t) cannot be rejected (based on a prespecified error level α). To determine the asymptotic distribution of T_q under the null hypothesis and to derive a test decision, a permutation test is performed. That means one permutes the values of the splitting variable x_q (or t) in the original augmented data matrix and computes the value of the maximally selected LR test statistic based on the permuted data. When this procedure is done repeatedly, an approximation of the distribution of T_q under the null hypothesis can be derived. Each permutation test is performed with local type I error level $\alpha_l = \alpha/s$, where s is the number of candidate variables for splitting in the selected node. Hence, for each tree the probability of falsely identifying at least one variable as splitting variable is bounded by α (i.e. on the tree level the family-wise error rate is controlled by α). To determine the corresponding p-values with sufficient accuracy, the number of permutations should increase with the number of candidate splitting variables.

(ii) A second alternative is a *minimal node size* criterion, which is widely used in tree-building algorithms (Probst et al. 2019). With the minimal node size criterion, the minimal number of observations in the nodes is considered as the main tuning parameter for tree construction. If the number of observations in a current node falls below a prespecified threshold, the node is flagged as terminal node and is no longer available for further splits. The construction of the tree(s) is terminated when all current nodes are flagged as terminal nodes. To find the optimal minimal node size we propose to use the predictive log-likelihood of the model (evaluated on a separate validation sample or by cross-validation). It appears to be a natural choice, because split selection is also based on the likelihood. Note that, for the model with predictor (12), the optimal minimal node size must be determined for both trees, which requires optimization on a two-dimensional grid.

(iii) The third alternative is a post-pruning strategy, where a large tree is grown first and is then pruned to an adequate size to avoid overfitting. More specifically, after building the large tree, the sequence of nested subtrees is evaluated with regard to its predictive performance, for example, by using a validation sample or a k-fold cross validation scheme. Finally, the best-performing subtree is selected. As with the minimal node size criterion, we propose to consider the predictive log-likelihood as evaluation criterion. Note that, also for model (12) with two tree components, this strategy requires a one-dimensional optimization only.

To prevent the algorithm from building extremely small nodes (with only a few observations), an additional *minimal bucket size* constraint may be applied. With the minimal bucket size constraint, the minimum number of observations required in any terminal node is limited downward.

To summarize, the following steps are performed during the fitting procedure, if the permutation test is applied:

1. **Initial model:** Fit the model without any explanatory variables, yielding a single estimate of the intercept $\hat{\gamma}_0$ (or a P-spline function modeling the baseline effects).
2. **Tree building:**
   (a) Fit all candidate models with one additional split regarding one of the explanatory variables (or t in case of a model with predictor (12) or (16)), that fulfill the minimal bucket size constraint, in one of the already built nodes. If none of the additional splits meets the minimal bucket size constraint, terminate the algorithm.
   (b) Select the best model based on the maximal LR test statistic.
   (c) Carry out the permutation test for the variable associated with the selected split with error level α_l. If significant, fit the selected model and continue with step (2a). Otherwise, terminate the algorithm.

If the minimal node size criterion is used, the algorithm can be summarized as follows:

1. **Initial model:** Fit the model without any explanatory variables.
2. **Tree building:**
   (a) Fit all candidate models with one additional split regarding one of the explanatory variables (or t), that fulfill the minimal bucket size constraint, in one of the already built nodes. If none of the additional splits meets the minimal bucket size constraint, terminate the algorithm.
   (b) Select the best model based on the maximal LR test statistic considering the already built nodes that fulfill the minimal node size criterion, only. If none of the already built nodes meets the minimal node size criterion, terminate the algorithm.
   (c) Fit the selected model and continue with step (2a).

Using the post-pruning method, the fitting procedure is as follows:

1. **Initial model:** Fit the model without any explanatory variables.
2. **Tree building:**
   (a) Fit all candidate models with one additional split regarding one of the explanatory variables (or t), that fulfill the minimal bucket size constraint, in one of the

already built nodes. If none of the additional splits meets the minimal bucket size constraint, continue with step (3).

(b) Select the best model based on the maximal LR test statistic.

(c) Fit the selected model and continue with step (2a).

3. **Pruning:** Select the best model with the minimal predictive log-likelihood from the sequence of models built in steps (1) and (2). Terminate the algorithm.

In R, the augmented data matrix for fitting discrete time-to-event models can be obtained by applying the function `dataLong()` of the package **discSurv** (Welchowski et al. 2022). Technically, the proposed algorithm can be embedded into the framework of tree-structured varying coefficients models (TSVC; Berger et al. 2019). The models can therefore be fitted by the eponymous R add-on package **TSVC** (Berger 2021), where the explanatory variables (and the time t) serve as effect modifiers, modifying the effect of a constant auxiliary variable. Online Resource Supplement 2 contains a collection of code examples demonstrating how the proposed models can be fitted using **TSVC**.

4 Simulation study

To assess the performance of the proposed tree-structured models and to illustrate their properties, we considered different simulation scenarios. In all scenarios, the aims were (i) to evaluate how the performance of the approaches is affected by the degree of censoring and the number of noise variables, (ii) to compare the three approaches for determining the size of the tree(s), and (iii) to compare the proposed approaches to alternative methods, in particular to penalized maximum likelihood (ML) estimation and Bayesian additive regression trees. In addition, the computation times of the models were considered. The scenarios were based on a true model with predictor (12) composed of a piecewise constant baseline function and tree-structured effects of the explanatory variables (scenario 1), a true model with predictor (4) composed of a smooth baseline function and linear effects of the explanatory variables (scenario 2), and a true model with predictor (16) including interaction effects between time and explanatory variables (scenario 3). More details on the data-generating models will be given in the following subsections.

In each scenario, the following models were fitted to the simulated data:

(i) the tree-structured model with smooth baseline effects as described in Sect. 3.1, referred to as *SB*,

(ii) the tree-structured model with piecewise constant baseline coefficients as described in Sect. 3.2, referred to as *PCB*,

(iii) the tree-structured model including t and $\boldsymbol{x}$ in a single tree as described in Sect. 3.3, referred to as *ST*,

(iv) a parametric model (4) with a LASSO penalty on each of the regression coefficients and a group LASSO penalty on the baseline coefficients, referred to as *LASSO*,

(v) a nonparametric model using Bayesian additive regression trees with logistic link function, referred to as *BART*,

(vi) a fully specified parametric model (4) including dummy-coded baseline coefficients and all explanatory variables (*Full*),

(vii) a model without any explanatory variables and a constant baseline coefficient only (*Null*), and

(viii) the true data-generating model (*Perfect*).

For the SB, the PCB and the ST model, permutation tests (*PT*), the minimal node size criterion (*MNS*) as well as the post-pruning method (*PR*) were applied to determine the optimal sizes of the trees. The baseline coefficients of the SB model were estimated by a P-spline comprising five cubic B-spline basis functions and using a second order difference penalty. The optimal penalty parameter ε was determined by generalized cross-validation (see Wood 2017).

The LASSO model contains a penalty term for each explanatory variable and for the time t. The group LASSO penalty on the baseline coefficients guarantees that either all or none of the corresponding estimates are equal to zero. The degree of regularization is controlled by a joint penalty parameter ε. See Yuan and Lin (2006) and Meier et al. (2008) for details.

The BART model, as introduced for survival analysis by Sparapani et al. (2016), is built of a sum of trees modeling the effects of time and the explanatory variables jointly (analogously to the ST model) and is fitted within a Bayesian framework. To achieve maximum comparability, we used a logistic link function (instead of the default probit link) and set the number of trees to one. Following Sparapani et al. (2016), the posterior distributions were estimated based on 2000 draws using a Markov chain Monte Carlo (MCMC) algorithm with 100 burn-in draws, a thinning factor of 10, and the default priors (see Sparapani et al. 2021).

In all simulation scenarios, the time-to-event outcome was related to one binary explanatory variable $x_1 \sim B(1, 0.5)$ and two continuous explanatory variables $x_2, x_3 \sim N(0, 1)$. We considered a *low-dimensional* setting with five additional binary noise variables ($\sim B(1, 0.5)$) and two additional standard normally distributed variables, as well as a *high-dimensional* setting with 90 additional binary noise variables and seven additional standard normally distributed noise vari-

ables. In each scenario we performed 100 replications and simulated a training sample, a validation sample and a test sample of size $n = 500$, respectively. The number of discrete time points was set to $k = 11$. The validation sample was used to determine the optimal minimal node sizes and best performing subtrees (post-pruning) for the SB, the PCB and the ST model and the optimal penalty parameter ε for the LASSO model by maximization of the predictive log-likelihood. The permutation tests were based on 1000 permutations with error level $\alpha = 0.05$. The censoring times C_i were sampled independently of the event times T_i using the probability mass function $P(C_i = t) = b^{k+1-t}/\sum_{j=1}^{k} b^j,\ t = 1, \ldots, k$. This resulted in censoring rates of approximately 30% ($b = 0.7$), 50% ($b = 1$) and 70% ($b = 1.3$).

To assess the performance of the different approaches, the predictive log-likelihood was calculated on the test samples. In addition, we considered prediction error (PE) curves as a time-dependent measure of prediction error. In case of BART, we used the mean (predicted) hazards over the MCMC draws to calculate the predictive log-likelihood values and the PE curves. The PE at time point t is given by

$$\widehat{PE}(t) = \frac{1}{n}\sum_{i=1}^{n} w_i(t)\left(\hat{S}_i(t) - \tilde{S}_i(t)\right)^2,$$

where $\hat{S}_i(t) = \prod_{s=1}^{t}(1 - \hat{\lambda}(s|\boldsymbol{x}_i))$ denotes the estimated survival function, $\tilde{S}_i(t) = I(t < \tilde{T}_i)$ denotes the observed survival function of individual i at time t, and $w_i(t)$ are inverse-probability-of-censoring weights, see van der Laan and Robins (2003). We also computed the integrated PE given by

$$\widehat{PE}_{int} = \sum_{t=1}^{k-1} \widehat{PE}(t) \cdot \hat{P}(T = t).$$

The marginal probabilities $P(T = t)$ were estimated using a logistic discrete hazard model with dummy variables for each time point. For more details on PE curves, see Tutz and Schmid (2016).

Furthermore, true positive rates (TPR) and false positive rates (FPR) for the explanatory variables were considered. The true positive rate specifies the proportion of explanatory variables that were correctly identified to have an effect on the hazard (that is, correctly selected for splitting). It is given by

$$TPR_X = \frac{1}{\#\{j : \vartheta_j = 1\}} \sum_{j:\vartheta_j=1} I\left(\hat{\vartheta}_j = 1\right),$$

where $\vartheta_j = 1$ if x_j has an effect on the hazard and $\vartheta_j = 0$ otherwise. The false positive rate describes the proportion of all noise variables that were falsely identified to have an effect on the hazard and is given by

$$FPR_X = \frac{1}{\#\{j : \vartheta_j = 0\}} \sum_{j:\vartheta_j=0} I\left(\hat{\vartheta}_j = 1\right).$$

In case of BART, TPR and FPR were determined by averaging over the MCMC draws.

For the settings with tree-structured effects of time, we also determined a true positive rate for the thresholds in t, which is given by

$$TPR_T = \frac{1}{\#\{k : \delta_t = 1\}} \sum_{t:\delta_t=1} I\left(\hat{\delta}_t = 1\right),$$

where $\delta_t = 1$ if there is a split at t and $\delta_t = 0$ otherwise. Analogously, the false positive rate for the thresholds in t is given by

$$FPR_T = \frac{1}{\#\{k : \delta_t = 0\}} \sum_{t:\delta_t=0} I\left(\hat{\delta}_t = 1\right).$$

4.1 Model with piecewise constant baseline effects and tree-structured effects of the explanatory variables

In the first scenario, the predictor of the true data-generating model had the form

$$\begin{aligned}\eta(t, \boldsymbol{x_i}) =& \gamma_0(t) + \big[\gamma_1 I(x_{i2} \le 0 \wedge x_{i1} = 0) \\ &+ \gamma_2 I(x_{i2} \le 0 \wedge x_{i1} = 1) \\ &+ \gamma_3 I(x_{i2} > 0 \wedge x_{i3} \le 0) \\ &+ \gamma_4 I(x_{i2} > 0 \wedge x_{i3} > 0)\big]\end{aligned}$$

with $\gamma_1 = -2.5$, $\gamma_2 = -1.5$, $\gamma_3 = -0.5$, $\gamma_4 = 0.5$, and baseline function

$$\begin{aligned}\gamma_0(t) =& -1.5 I(t \le 3) - I(3 < t \le 5) \\ &- 0.5 I(5 < t \le 8).\end{aligned}$$

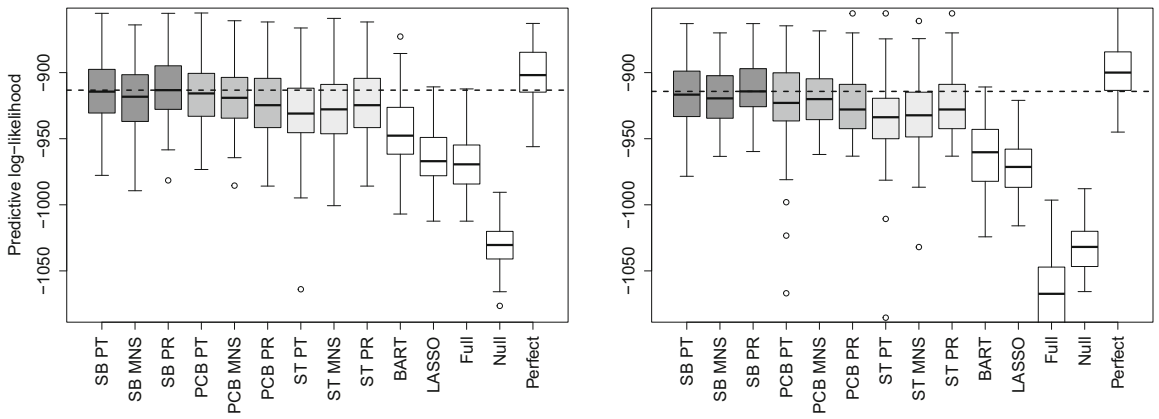
Figure S1 in Online Resource Supplement 3 shows the effects of the explanatory variables represented as a tree structure.

From the results in Table 1 it is seen that all the proposed models were very efficient in detecting the informative variables. The TPR (upper panel) were higher than 0.6 throughout all settings, even for high-dimensional data with strong censoring. Compared to the tree-structured models, LASSO yielded much higher FPR, indicating that the selected models were far too large, while BART yielded much lower TPR. This also resulted in a considerably worse predictive performance of LASSO and BART compared to the proposed models (see Fig. 1, which show the results for the

Table 1 Results of the simulation study: Explanatory variables (scenario 1)

	Model	Stopping criterion	Low-dimensional			High-dimensional		
Censoring rate			0.3	0.5	0.7	0.3	0.5	0.7
TPR_X	SB	PT	0.987	0.940	0.813	0.970	0.863	0.720
		MNS	0.923	0.903	0.833	0.913	0.910	0.730
		PR	0.983	0.926	0.823	0.980	0.917	0.773
	PCB	PT	0.990	0.940	0.810	0.953	0.857	0.660
		MNS	0.930	0.913	0.860	0.910	0.917	0.750
		PR	0.983	0.930	0.837	0.983	0.943	0.777
	ST	PT	0.923	0.883	0.757	0.853	0.750	0.617
		MNS	0.950	0.870	0.760	0.910	0.813	0.670
		PR	0.980	0.907	0.790	0.927	0.853	0.747
	BART	–	0.784	0.761	0.666	0.562	0.477	0.430
	LASSO	–	1.000	0.997	0.980	0.997	0.880	0.740
FPR_X	SB	PT	0.017	0.028	0.010	0.004	0.003	0.004
		MNS	0.081	0.040	0.011	0.007	0.004	0.003
		PR	0.014	0.016	0.010	0.001	0.001	0.001
	PCB	PT	0.011	0.020	0.009	0.004	0.003	0.003
		MNS	0.073	0.047	0.020	0.007	0.005	0.004
		PR	0.023	0.019	0.020	0.001	0.001	0.002
	ST	PT	0.004	0.004	0.004	0.002	0.001	0.001
		MNS	0.041	0.036	0.021	0.005	0.005	0.002
		PR	0.020	0.011	0.009	0.001	0.001	0.001
	BART	–	0.119	0.099	0.119	0.013	0.015	0.011
	LASSO	–	0.637	0.627	0.504	0.201	0.470	0.183

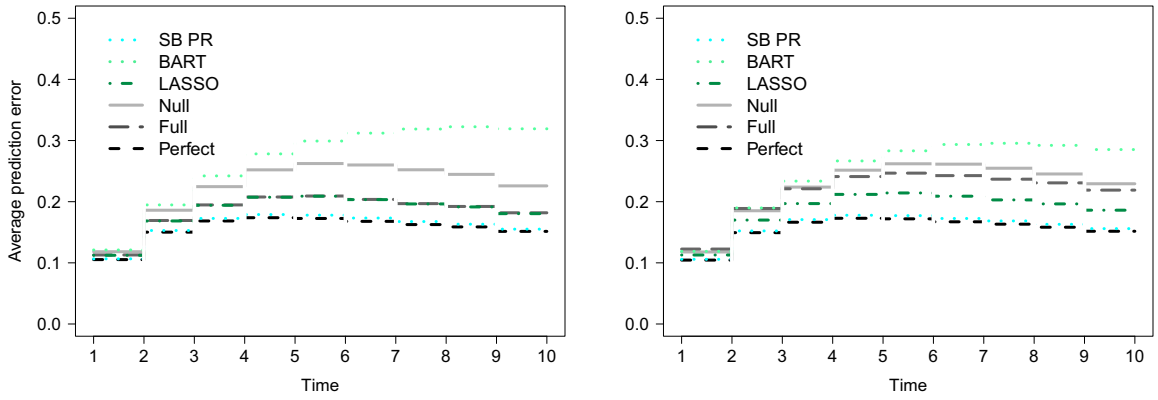
Average true positive rates (TPR_X) and false positive rates (FPR_X) of the explanatory variables for the low-dimensional settings (left) and high-dimensional settings (right) and different degrees of censoring



(a) Low-dimensional setting

(b) High-dimensional setting

Fig. 1 Results of the simulation study: Predictive performance (scenario 1). The figure shows the predicted log-likelihood obtain from fitting the different models for the low-dimensional setting (left panel) and the high-dimensional setting (right panel) with low censoring (30%). For the results with medium and high censoring, see Figure S3 and S4 of Online Resource Supplement 3



(a) Low-dimensional setting (b) High-dimensional setting

Fig. 2 Results of the simulation study: Prediction error curves (scenario 1). The figures show the prediction error curves (averaged over 100 replications) of the proposed tree-structured model with the lowest integrated prediction error (SB PR) and the competing models for the low-dimensional setting (left panel) and the high-dimensional setting (right panel) with low censoring (30%). For the results of the other proposed tree-structured models, see Figures S9 and S10 of Online Resource Supplement 3. Values of the integrated PE for all settings (with low, medium and high censoring) are given in Table S3 of Online Resource Supplement 3

settings with low censoring). The corresponding results for medium and high censoring are given in Online Resource Supplement 3.

It is also seen that the SB and PCB model performed similarly well, whereas the ST model performed worst among the proposed models. This indicates that the true underlying piecewise constant baseline function could be approximated sufficiently by the smooth function of the SB model. Modeling the effects of time and the explanatory variables in one tree, however, did not fit the data very well.

Furthermore, models where permutation tests were applied as stopping criterion mostly showed lower FPR than models using the minimal node size criterion or the post-pruning method (see lower panel of Table 1). This is because the permutation test procedure is intended to control the family-wise error rate α (see Sect. 3.4). Regarding the TPR, the permutation test procedure showed slightly better performance in the settings with low censoring, whereas applying the minimal node size criterion was superior in the settings with high censoring. The post-pruning method performed particularly well in the high-dimensional settings. Of note, in the present setting the true underlying model was determined by a tree with similarly sized terminal nodes. In a more unbalanced setting, the minimal node size criterion tended to be inferior to the permutation test and the post-pruning method (see, for example, Figure S8 of scenario 3 in Online Resource Supplement 3).

Figure 2 shows the PE curves for the settings with low censoring. The figure depicts the results obtained from the proposed tree-structured model with the lowest integrated prediction error (SB PR) and the competing models. It is seen that the differences in PE were larger for later time points, with the proposed SB PR model resulting in the lowest PE values among the competitors across all time points. This result appears to be unaffected by the dimensionality of the data. For the PE curves of the other proposed-tree-structured models, we refer to Figures S9 and S10 of Online Resource Supplement 3.

The selection rates obtained for the thresholds in t given in Table 7 of Online Resource Supplement 3 demonstrate that the PCB model with minimal node size criterion carried out considerably more splits in time yielding higher TPR and FPR. In case of the ST model, most splits were performed when the post-pruning method was applied.

4.2 Model with smooth baseline effects and linear effects of the explanatory variables

The second scenario was based on a predictor of the form

$$\eta(t, \boldsymbol{x_i}) = \gamma_0(t) + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3}$$

with $\gamma_1 = -0.5$, $\gamma_2 = -0.25$, $\gamma_3 = 0.25$ and a sigmoid baseline function

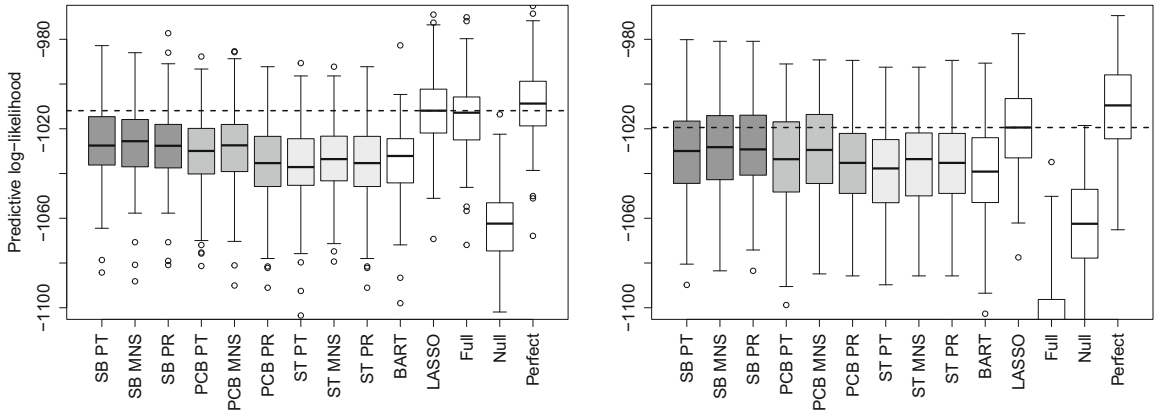
$$\gamma_0(t) = -2.25 + 1.5 \times \exp\left(-(t-6)^{-1}\right).$$

Table 2 and Fig. 3 show that the proposed models yielded lower TPR than in scenario 1 throughout all settings and were outperformed by the LASSO in terms of TPR and predictive ability. The TPR particularly deteriorated in the settings with high censoring. The BART model appeared to be less

Table 2 Results of the simulation study: Explanatory variables (scenario 2)

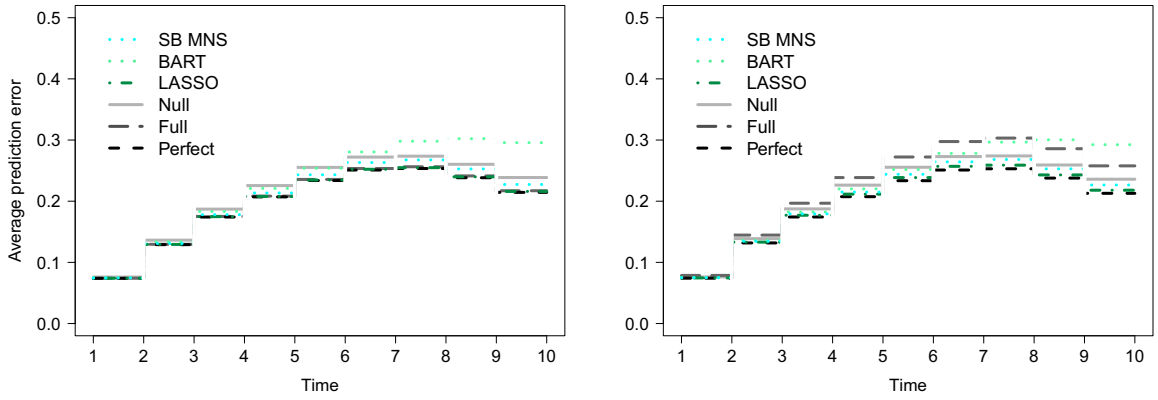
	Model	Stopping criterion	Low-dimensional			High-dimensional		
Censoring rate			0.3	0.5	0.7	0.3	0.5	0.7
TPR_X	SB	PT	0.667	0.507	0.343	0.460	0.320	0.128
		MNS	0.760	0.593	0.350	0.540	0.417	0.256
		PR	0.700	0.580	0.353	0.552	0.363	0.243
	PCB	PT	0.660	0.507	0.307	0.433	0.317	0.128
		MNS	0.737	0.590	0.327	0.547	0.410	0.239
		PR	0.717	0.597	0.397	0.580	0.393	0.230
	ST	PT	0.517	0.380	0.270	0.270	0.190	0.128
		MNS	0.553	0.417	0.287	0.337	0.263	0.182
		PR	0.587	0.397	0.263	0.337	0.243	0.153
	BART	–	0.553	0.491	0.378	0.216	0.209	0.156
	LASSO	–	0.910	0.880	0.740	0.773	0.637	0.527
FPR_X	SB	PT	0.007	0.009	0.012	0.002	0.001	0.001
		MNS	0.030	0.026	0.007	0.002	0.001	0.001
		PR	0.020	0.009	0.003	0.001	0.000	0.001
	PCB	PT	0.006	0.011	0.007	0.001	0.002	0.001
		MNS	0.037	0.020	0.011	0.001	0.001	0.001
		PR	0.026	0.013	0.004	0.001	0.000	0.001
	ST	PT	0.007	0.014	0.006	0.002	0.001	0.001
		MNS	0.021	0.006	0.013	0.002	0.001	0.001
		PR	0.011	0.015	0.009	0.000	0.001	0.001
	BART	–	0.111	0.096	0.077	0.010	0.010	0.011
	LASSO	–	0.620	0.582	0.467	0.204	0.160	0.117

Average true positive rates (TPR_X) and false positive rates (FPR_X) of the explanatory variables for the low-dimensional settings (left) and high-dimensional settings (right) and different degrees of censoring



(a) Low-dimensional setting (b) High-dimensional setting

Fig. 3 Results of the simulation study: Predictive performance (scenario 2). The figure shows the predicted log-likelihood obtained from fitting the different models for the low-dimensional setting (left panel) and the high-dimensional setting (right panel) with low censoring (30%). For the results with medium and high censoring, see Figure S5 and S6 of Online Resource Supplement 3



(a) Low-dimensional setting (b) High-dimensional setting

Fig. 4 Results of the simulation study: Prediction error curves (scenario 2). The figures show the prediction error curves (averaged over 100 replications) of the proposed tree-structured model with the lowest integrated prediction error (SB MNS) and the competing models for the low-dimensional setting (left panel) and the high-dimensional setting (right panel) with low censoring (30%). For the results of the other proposed tree-structured models, see Figures S11 and S12 of Online Resource Supplement 3. Values of the integrated PE for all settings (with low, medium and high censoring) are given in Table S4 of Online Resource Supplement 3

affected by an increasing censoring rate, but showed much lower TPR in the high-dimensional settings.

Yet, the proposed models showed substantially better predictive performance compared to the Null model in the low- and high-dimensional setting as well as compared to the Full model in the high-dimensional setting, which reflects the added value of the variable selection procedure.

Again, the ST models achieved the lowest predictive log-likelihood values among the tree-structured models (seventh, eighth and ninth boxplot of Fig. 3), which demonstrates that fitting one single tree comprising the effects of time and the explanatory variables is not adequate if the underlying model has an additive structure.

The models applying the minimal node size criterion performed slightly better than models using permutation tests and the post-pruning method. The rather weak performance of the permutation test indicates that the test procedure is too conservative to detect numerous informative variables in settings with moderately sized linear effects.

Figure 4 shows that the PE for the settings with low censoring increased until $t = 7$ for all of the models. The LASSO performed best across all time points in both the low- and high-dimensional setting. While the SB MNS model was outperformed by the LASSO, it was superior to the Null and BART model in the low-dimensional setting as well as to the Full model in the high-dimensional setting. For the PE curves of the other proposed-tree-structured models, we refer to Figures S11 and S12 of Online Resource Supplement 3.

4.3 Tree-structured model with interaction effects between time and explanatory variables

The data of the third scenario was generated based on a predictor of the form

$$\begin{aligned}\eta(\boldsymbol{x_i}, t) = \gamma_0 &+ \gamma_1 I(t \leq 4 \wedge x_{i1} = 0 \wedge x_{i3} \leq 0) \\ &+ \gamma_2 I(t \leq 4 \wedge x_{i1} = 0 \wedge x_{i3} > 0) \\ &+ \gamma_3 I(t \leq 4 \wedge x_{i1} = 1) \\ &+ \gamma_4 I(t > 4 \wedge x_{i2} \leq 0) \\ &+ \gamma_5 I(t > 4 \wedge x_{i2} > 0)\end{aligned}$$

with $\gamma_0 = -0.5$, $\gamma_1 = -4$, $\gamma_2 = -3$, $\gamma_3 = -2$, $\gamma_4 = -1$, $\gamma_5 = 0$. Figure S2 in Online Resource Supplement 3 illustrates the predictor function represented as a tree structure.

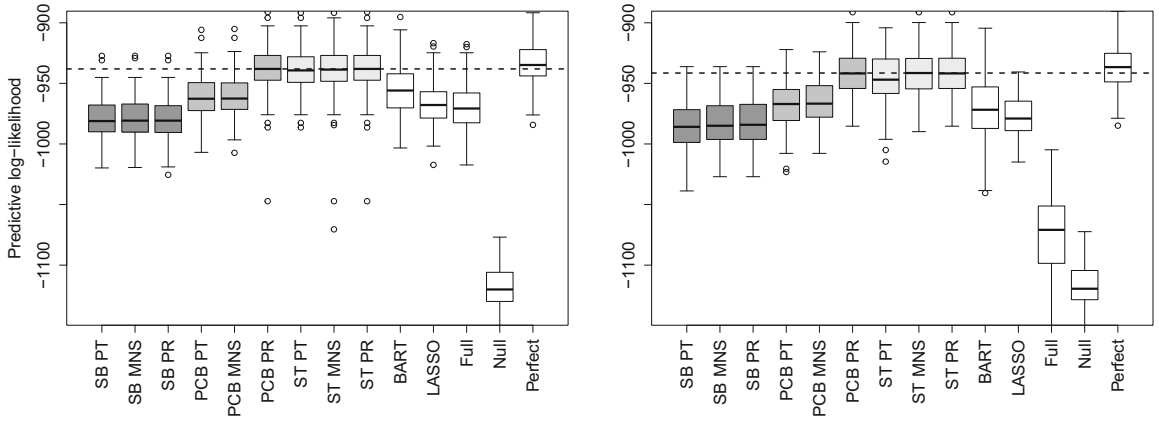
From Table 3 and Fig. 5 it is seen that the performance of the different models deviated from each other more strongly than in the previous scenarios. While the LASSO yielded higher TPR than the tree-structured models, the proposed models resulted in much lower FPR than LASSO and BART and exhibited good predictive performance. In particular, they performed substantially better than the Null model as well as the Full model in the high-dimensional settings).

Among the proposed models, the ST model (seventh, eighth and ninth boxplots in Fig. 5), whose structure corresponds to the form of the true underlying predictor function, was clearly superior to the SB and PCB models. Further, the difference in predictive performance between the SB and PCB models shows that the abrupt change in the hazard at time point $t = 4$ could be captured less adequately by the

Table 3 Results of the simulation study: Explanatory variables (scenario 3)

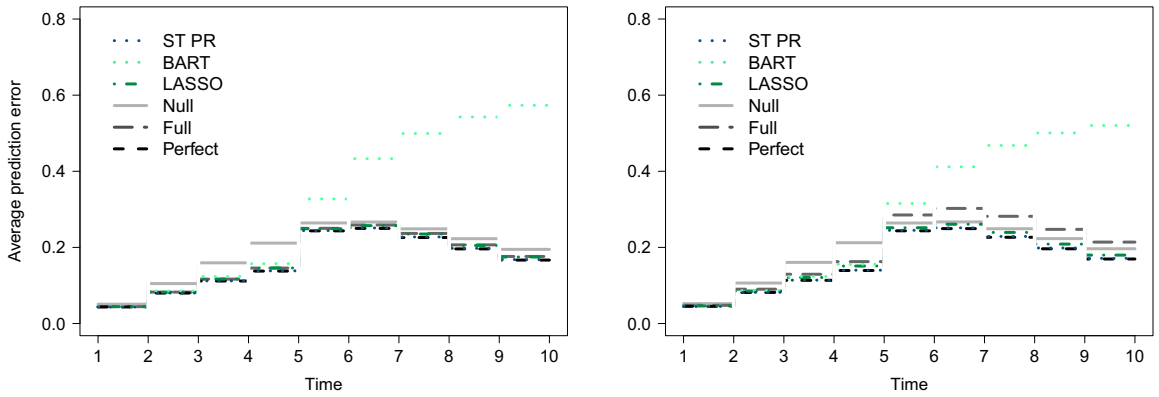
	Model	Stopping criterion	Low-dimensional			High-dimensional		
Censoring rate			0.3	0.5	0.7	0.3	0.5	0.7
TPR_X	SB	PT	0.487	0.463	0.270	0.383	0.310	0.190
		MNS	0.490	0.453	0.363	0.390	0.387	0.247
		PR	0.483	0.463	0.330	0.403	0.367	0.233
	PCB	PT	0.493	0.483	0.300	0.393	0.323	0.193
		MNS	0.483	0.467	0.353	0.400	0.377	0.260
		PR	0.493	0.480	0.347	0.413	0.370	0.247
	ST	PT	0.713	0.693	0.550	0.687	0.650	0.440
		MNS	0.740	0.550	0.360	0.710	0.503	0.330
		PR	0.700	0.710	0.523	0.700	0.677	0.463
	BART	–	0.684	0.624	0.482	0.426	0.380	0.255
	LASSO	–	0.893	0.910	0.830	0.760	0.733	0.590
FPR_X	SB	PT	0.011	0.011	0.011	0.003	0.002	0.002
		MNS	0.033	0.017	0.009	0.002	0.002	0.001
		PR	0.014	0.017	0.004	0.001	0.001	0.000
	PCB	PT	0.011	0.014	0.010	0.003	0.004	0.002
		MNS	0.031	0.029	0.011	0.002	0.002	0.001
		PR	0.018	0.010	0.006	0.001	0.001	0.001
	ST	PT	0.017	0.013	0.016	0.004	0.003	0.002
		MNS	0.007	0.041	0.009	0.002	0.007	0.001
		PR	0.003	0.003	0.004	0.001	0.001	0.001
	BART	–	0.137	0.130	0.125	0.017	0.015	0.020
	LASSO	–	0.696	0.654	0.537	0.239	0.207	0.173

Average true positive rates (TPR_X) and false positive rates (FPR_X) of the explanatory variables for the low-dimensional settings (left) and high-dimensional settings (right) and different degrees of censoring



(a) Low-dimensional setting (b) High-dimensional setting

Fig. 5 Results of the simulation study: Predictive performance (scenario 3). The figure shows the predicted log-likelihood obtained from fitting the different models for the low-dimensional setting (left panel) and the high-dimensional setting (right panel) with low censoring (30%). For the results with medium and high censoring, see Figure S7 and S8 of Online Resource Supplement 3



(a) Low-dimensional setting (b) High-dimensional setting

Fig. 6 Results of the simulation study: Prediction error curves (scenario 3). The figures show the prediction error curves (averaged over 100 replications) of the proposed tree-structured model with the lowest integrated prediction error (ST PR) and the competing models for the low-dimensional setting (left panel) and the high-dimensional setting (right panel) with low censoring (30%). For the results of the other proposed tree-structured models, see Figures S13 and S14 of Online Resource Supplement 3. Values of the integrated PE for all settings (with low, medium and high censoring) are given in Table S5 of Online Resource Supplement 3

smooth function of the SB model. According to the TPR in Table S2 in Online Resource Supplement 3, a split in $t = 4$ was carried out by the PCB and ST models in each replication (as all values were equal to one).

In addition, Table 3 shows that the ST model resulted in much lower TPR when applying the minimal node size criterion compared to the permutation tests or the post-pruning procedure in the settings with medium and high censoring. This may be explained by the unbalanced size of the terminal nodes of the tree according to the true underlying model. This also results in a decreased predictive performance (see Figures S7 and S8 in Supplement 3 of the Online Resource).

The PE curves shown in Fig. 6 were very close at early time points, but for later time points ($t > 4$) in particular the BART model performed worse than the other models, which may be caused by the low number of observations at subsequent time points (leading to a stronger influence of the prior distributions). The proposed ST PR model showed the lowest PE among the competitors across all time points for low- and high-dimensional data. For the PE curves of the other proposed-tree-structured models, we refer to Figures S13 and S14 of Online Resource Supplement 3.

4.4 Run-time

In the last part of the simulation study we investigated the computing times of the proposed tree-structured models and compared them to the alternative approaches LASSO and BART. To do so, we evaluated the run-times in simulation scenario 2 with low censoring. This setting was chosen because increasing the degree of censoring results in a lower number of rows in the augmented data matrix ($\tilde{n}$) and therefore generally reduces run-times. Table 4 shows that the run-times differ strongly depending on the structure of the model (SB, PCB or ST) and the stopping criterion (PT, MNS or PR). Run-times were consistently longer compared to the LASSO model, which on average took only one minute or less in both settings, and also longer than the BART model in some cases. Among the proposed tree-structured approaches the post-pruning method was least expensive, in particular for the ST and PCB model. A higher number of noise variables (high-dimensional setting) increased the run-times when using the minimal node size criterion or the post-pruning method. In contrast, the run-times decreased with a larger number of noise variables if permutation tests were applied. This is because the effort of the permutation test approach directly depends on the actual size of the fitted trees, which was considerably smaller in the high-dimensional setting (resulting in lower TPR, cf. Table 2). Moreover, the long run-times of the SB model demonstrate the computational demand of refitting the smooth baseline function $\gamma_0(t)$ in each iteration, and the low performance of the PCB MNS model resulted from the optimization of the minimal node size on a two-dimensional grid.

5 Application

To illustrate the use of the proposed models, two real-world examples were considered. In both applications, we compared the different approaches also considered in the simulation study and selected the best-performing approach using

Table 4 Results of the simulation study: Run-times

Model	Stopping criterion	Run-time	
		Low-dimensional	High-dimensional
SB	PT	156	95
	MNS	93	231
	PR	33	66
PCB	PT	17	9
	MNS	125	317
	PR	3	6
ST	PT	11	8
	MNS	6	16
	PR	2	5
BART	–	56	55
LASSO	–	>1	1

Run-times in minutes averaged over 100 replications for fitting the different models in simulation scenario 2 with low censoring (30%) in the low- and high-dimensional setting. The calculations were performed on a high performance computing cluster that consists of 35 nodes with a total of 944 cores for general purpose computing. An additional 372 cores are reserved for machine learning computations and aided by 6 A100 or V100S NVIDIA GPUs. The cores are a mixture of older (e.g. Nehalem, Opteron) and modern (AMD Epyc) processors. The total memory of the system is 4.6 terabytes and roughly 400 terabytes of hard drive space are available for storing project data. Connections between nodes are using 40 or 56 gigabits per second Infiniband links

a cross-validation procedure. The corresponding results are presented in the following.

5.1 Patients with acute odontogenic infection

We considered data of a five-year retrospective study investigating hospitalized patients with abscess of odontogenic origin conducted between 2012 and 2017 by the Department of Oral and Cranio-Maxillo and Facial Plastic Surgery at the University Hospital Bonn. Patients with an acute odontogenic infection suffer from pain, swelling, erythema and hyperthermia. If not treated at an early stage, such infections may spread into deep neck spaces and lead to perilous complications by menacing anatomical structures, such as major blood vessels, the upper airway and the mediastinum (Biasotto et al. 2004). The primary objective of the study was to identify risk factors that are associated with a prolonged length of stay (LOS) in the treatment of severe odontogenic infections. As the LOS was recorded in 24-h intervals (that is, time was measured in days $t = 1, \ldots, 18$), the use of a discrete hazard model is appropriate.

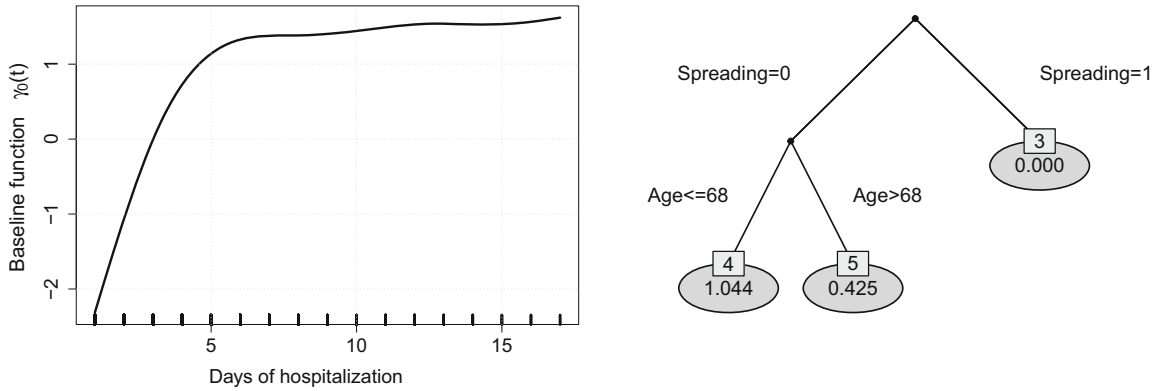
Here data from 303 patients that underwent surgical treatment in terms of incision and drainage of the abscess were considered. Intravenous antibiotics were administered during the operation and for the length of inpatient treatment. Further details on the study can be found in Heim et al. (2019). The characteristics of the patients considered for modeling were: age in years, gender (0: female, 1: male), an indicator of whether the infection spread into other facial spaces (0: no, 1: yes), the location of the infection focus (0: mandible, 1: maxilla), the administered antibiotics (0: ampicillin, 1: clindamycin), the presence of diabetes mellitus type 2 (0: no, 1: yes) and an indicator of whether the infection was already removed at admission (0: no, 1: yes). Basic statistics of the LOS and the patients characteristics are summarized in Table S6 in Online Resource Supplement 4.

Table 5 Analysis of the odontogenic infection data

Model	Stopping criterion	Predictive log-likelihood
SB	PT	−70.31
	MNS	−70.23
	PR	**−69.91**
PCB	PT	−70.43
	MNS	−70.43
	PR	−70.56
ST	PT	−71.12
	MNS	−74.06
	PR	−71.27
BART	–	−71.21
LASSO	–	−70.77
Full	–	−73.36
Null	–	−80.93

Predictive log-likelihood values for the different modeling approaches based on ten-fold cross-validation. The bold value corresponds to the best perfoming model

The logistic discrete hazard model with predictor (4) including linear effects of the explanatory variables that was recently applied for statistical analysis by Heim et al. (2019)



(a) Baseline function (b) Effects of the explanatory variables

Fig. 7 Analysis of the odontogenic infection data. The figure shows the results obtained from fitting the SB model with minimal node size criterion to the odontogenic infection data. On the left, the estimated smooth baseline function is presented. The graph on the right shows the estimated tree obtained from fitting the SB model with minimal node size criterion. The estimated coefficients γ_m are given in each leaf of the tree, where the right node on the first tree level serves as reference node

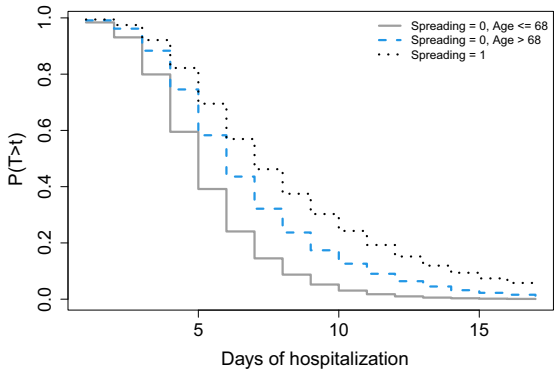


Fig. 8 Analysis of the odontogenic infection data. The figure shows the estimated probability of being of being still at ward for the three different groups of patients determined by the terminal nodes of the fitted tree

indicated that age and spreading of the infection focus into facial spaces significantly affected the LOS (with error level $\alpha = 0.05$), while all the other variables showed no evidence for an effect.

In the first step of our analysis, we compared the tree-structured models (SB, PCB, ST, and BART), the LASSO model, the Full and the Null model with regard to their predictive performance. More specifically, we calculated the predictive log-likelihood values using ten-fold cross-validation. In case of the proposed tree-structured models, permutation tests, the minimal node size criterion, and the post-pruning method were used to limit the size of the trees. Additionally, a minimal bucket size constraint of $mb = \lfloor 0.1\tilde{n} \rfloor$, where $\tilde{n}$ is the number of observations (i.e. rows) in the augmented data matrix, was set for all tree-structured models. The optimal minimal node size, the optimal subtree, and the optimal LASSO penalty parameter were respectively determined by means of leave-one-out cross-validation on each of the ten training samples. Accordingly, predictive log-likelihood values were calculated for each of the ten test samples. Afterwards, the minimal node size, the subtree, and the penalty parameter value with the largest average predictive likelihood were selected.

Table 5 shows that the SB model applying the post-pruning method achieved the best performance, that is, the highest cross-validated log-likelihood value. In the second step of the analysis, we therefore fitted the SB PR model to the complete data. To determine the optimal subtree, we again performed leave-one-out cross-validation and selected the model with the highest predictive log-likelihood. The smooth baseline function of the SB model was fitted by five cubic P-splines with a second order difference penalty.

Figure 7a shows the estimated smooth baseline function $\gamma_0(t)$ obtained from fitting the SB PR model. The figure illustrates that it was highly unlikely for patients to be discharged in the first few days after surgery. The probability of being discharged strongly increased until day five, but subsequently remained constant until day 18. Figure 7b shows the estimated tree-structured effects of the explanatory variables on the LOS. According to the first split, patients with spreading into facial spaces (Spreading=1) had a lower probability of being discharged. In the group of patients without spreading into facial spaces (Spreading=0), age was identified as an additional risk factor affecting the LOS. Patients who were older than 68 years were less likely being discharged than patients who were 68 years of age or younger. For patients

Table 6 Analysis of the lymphantic filariasis data

Model	Stopping criterion	Predictive log-likelihood
PCB	PT	−15.50
	MNS	−14.99
	PR	−18.07
ST	PT	−15.45
	MNS	**−14.92**
	PR	−15.08
BART	–	−15.29
LASSO	–	−15.00
Full	–	−16.45
Null	–	−15.25

Predictive log-likelihood values for the different modeling approaches based on ten-fold cross-validation. The bold value corresponds to the best perfoming model

68 years of age or younger without spreading the continuation ratio was increased by the factor $\exp(1.044) = 2.841$ compared to patients with spreading. Figure 8 depicts the corresponding estimated survival functions for the three groups of patients determined by the tree.

Overall, our results strongly coincide with the findings by Heim et al. (2019). Based on the predictive log-likelihood, however, the tree-structured models SB and PCB surpassed the BART as well as the unrestricted and penalized models with linear effects (Full, Null and LASSO). This indicates that the tree-structured model (detecting an interaction between Spreading and Age) describes the association between the LOS and the explanatory variables best.

5.2 Patients with lymphatic filariasis

As a second example, we considered data from a randomized controlled trial in patients with lymphatic filariasis that was carried out in the western region of Ghana. Lymphatic filariasis is a filarial worm disease transmitted by mosquitoes that affects approximately 120 million persons worldwide and can lead to the development of severe lymphedema (LE) or hydroceles. The main objective of the study was to investigate the effect of antibiotic doxycycline in LE patients. Doxycycline has been shown previously to ameliorate LE severity in a subgroup of the population (Debrah et al. 2006). While the primary outcome of the study was change in LE stages, a key secondary outcome was time to occurrence of acute attacks, which are caused by secondary infections through skin lesions. Patients were examined 3 ($t = 1$), 12 ($t = 2$) and 24 ($t = 3$) months after treatment onset, i.e., time was measured on a discrete scale with unequally spaced time intervals.

A sample of 118 patients was analyzed. The empirical distribution of the event times was 23.73%, 31.36% and 44.92% for $\tilde{T} = 1, 2, 3$, respectively. The censoring rate was 34.75% (patients, who did not suffer from an attack during follow-up). For details on how the trial was conducted, see Mand et al. (2012). The characteristics of the patients that were included in the analysis are: age in years, weight in kilograms, microfiliariae count, gender (0: female, 1: male), infection status (0: negative, 1: positive), administered drug (0: placebo, 1: 6-week course of amoxicillin at 1000 mg/d, 2: 6-week course of doxycycline at 200 mg/d), lymphedema stage before treatment and hygiene status before treatment (0: poor, 1: moderate, 2: good). Based on the inclusion criteria of the study, only patients with LE stages 1–5 according to the scheme by Dreyer et al. (2002) were considered. Two patients were excluded from the analysis because of missing values in the explanatory variables. Summary statistics on the characteristics of the patients are given in Table 12 in Supplement S7 of the Online Resource.

Based on a log-rank test, Mand et al. (2012) found a significant difference between the doxycycline and the placebo group with regard to the occurrence of acute attacks (with error level $\alpha = 0.05$). However, there was no evidence for a difference between the doxycycline and the amoxicillin group. More recently, the study data was analyzed with the survival tree method by Schmid et al. (2016). In their analysis, splits in hygiene status, LE stage and drug were performed, indicating that these characteristics affect the risk for an acute attack.

Our analysis of the data was performed analogously to the analysis of the odontogenic infection data (using ten-fold cross-validation and the minimal bucket size constraint $mb = \lfloor 0.1\tilde{n} \rfloor$). Note that the three patients with LE stage 1 and 4 were excluded from the cross-validation step. Because of the small number of time intervals, fitting a smooth baseline function was not reasonable and the SB model was omitted.

The results of ten-fold cross-validation are shown in Table 6. The highest predictive log-likelihood value was achieved by the ST model applying the minimal node size criterion, which is (apart from the criterion for split selection) equivalent to the survival tree by Schmid et al. (2016). We therefore fitted the ST MNS model to the complete data (after determining the optimal minimal node size using again leave-one-out cross-validation).

Figure 9 shows the estimated tree structure, which concurs with the findings by Schmid et al. (2016), cf. Figure 7 therein. The first split was performed in hygiene showing the importance of good hygiene for lowering the risk for an acute attack. The subsequent splits reveal an interaction between time, lymphedema stage and drug for patients with poor to moderate hygiene. That is, within the first three months of the study, a higher lymphedema stage was shown to be a relevant risk factor, whereas after three months patients treated with doxycycline were shown to be at lower risk for an acute attack than patients in the placebo or amoxicillin

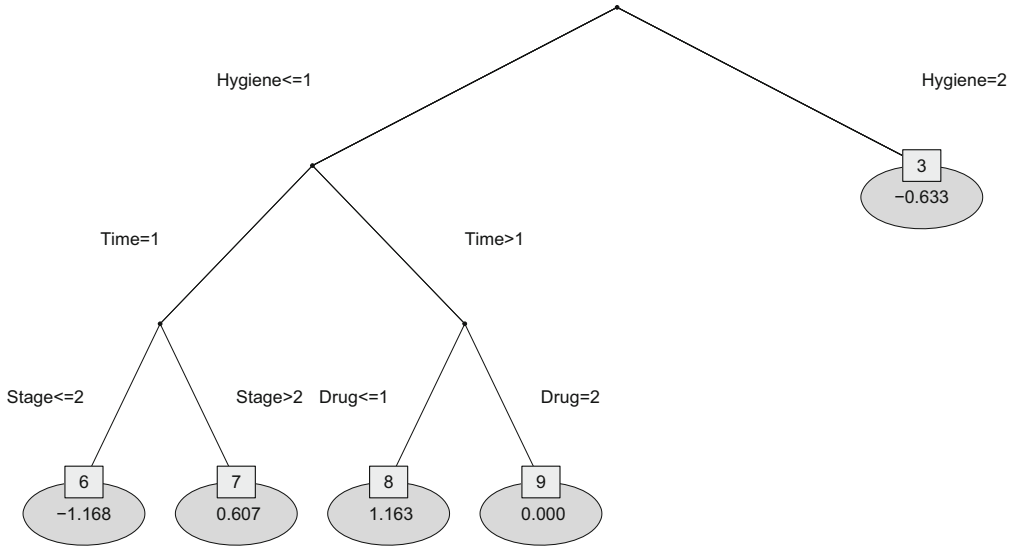


Fig. 9 Analysis of the lymphantic filariasis data. The figure shows the estimated tree obtained from fitting the ST model with minimal node size criterion. The estimated coefficients γ_m are given in each leaf of the tree, where the rightmost node on the third tree level serves as reference node

group. The continuation ratio for patients with poor to moderate hygiene was increased by the factor $\exp(1.163) = 3.199$ after three months for patients treated with placebo or amoxicillin compared to patients treated with doxycycline. This finding confirms the results of the analysis by Mand et al. (2012) who found that the effects of doxycycline for the risk of an acute attack develop over the course of the second and third follow-up interval.

The predictive log-likelihood values for the considered models hardly deviate from each other (apart from the PCB PR model) indicating a lack of evidence for the superiority of one or several models (Burnham and Anderson 2002). However, the estimated ST model coincides with the survival tree established by Schmid et al. (2016), confirming their findings and supporting the validity of our approach.

6 Summary and discussion

The results of the simulation study and the applications to real-world data indicate that the proposed tree-structured models are promising tools for modeling discrete event times. The tree-structured models, unlike common parametric approaches, are able to capture non-linear effects and interactions between the explanatory variables, as illustrated by the results in Figs. 7 and 9. Even though the linear modeling approach surpassed the tree-structured approaches in the simulations, where the true effects were linear (see Sect. 4.2), the proposed models were shown to be highly effective in identifying the informative variables, particularly in high-dimensional settings. The ST model has a more flexible form than the SB and PCB model, as splitting in t allows to account for time-varying effects of the explanatory variables on the hazard. Yet, this may come at the price of interpretability of the tree structure and corresponding effects.

Our approach differs from traditional recursive partitioning algorithms: (i) Fitting of our tree-structured models is done within the framework of parametric discrete hazard models, which allows to apply software for binary regression modeling. Split selection and pruning of the built tree(s) are therefore naturally based on the likelihood. (ii) The SB and PCB model exhibit the common additive form of parametric discrete hazard models, which facilitates the interpretation of effects in terms of continuation ratios. (iii) The predictor function can be specified in a very flexible way including the survival tree by Schmid et al. (2016) as special case. (iv) The framework is easily generalizable, for example, to an additive discrete hazard model of the form

$$\eta(t, \boldsymbol{x_i}) = \gamma_0(t) + tr(\boldsymbol{x_i}) + z_i^\top \boldsymbol{\beta},$$

where $\boldsymbol{z}$ is an additional set of explanatory variables with linear effects on the outcome. An exemplary code how to fit such a model using **TSVC** is given in Online Resource Supplement 2.

One of the most important parameters for tree building is the number of splits that determines the size of the tree. Hence, before fitting one of our proposed models, one needs to consider which of the three criteria (PT, MNS or PR) to apply. In terms of predictive performance, the permutation

test, the minimal node size criterion, and the post-pruning method yielded very similar results. The permutation test, which is the most conservative criterion (resulting in much lower FPR), appears slightly favorable in settings with low censoring, whereas the minimal node size criterion appears advantageous (showing higher TPR) for data with higher censoring rates, and the post-pruning methods appeared beneficial in high-dimensional settings. Both, performing permutation tests as well as determining the optimal minimal node size, may become computationally expensive, depending on the number of permutations or the resampling scheme used. In the latter case, for example, optimization on a two-dimensional grid (within the PCB model) using (leave-one-out) cross-validation requires high computing time (cf. Table 4). In terms of run-time, the post-pruning method proved to be superior, as the optimal subtree is selected from the sequence of nested subtrees that were already built on the training sample during iteration.

Alongside the logistic link function defined by the logistic distribution function (which we focused on in this article), the *probit link* function based on the standard normal distribution, and the complementary log-log (cloglog) link function defining the *Gompertz model* are popular choices. A comparison of parametric discrete hazard models with different link functions (considering information criteria) has been conducted by Hashimoto et al. (2011). All of the corresponding discrete hazard models postulate a proportionality property that affects the interpretation of parameters (Tutz and Schmid 2016, Chapter 3). In case of the logistic link function, the SB and PCB model yield proportionality with respect to the continuation ratios. Assume an SB model (7) with one split in x_j at split point c_j, where γ_1 is the coefficient of the left node and γ_2 is the coefficient of the right node. Then the continuation ratio at time t is given by

$$\begin{aligned}\Psi(t|\, x_j) &= \frac{P(T = t|\, x_j)}{P(T > t|\, x_j)} \\ &= \exp\big(\gamma_0(t) + [\gamma_1 I(x_j \le c_j) \\ &\qquad + \gamma_2 I(x_j > c_j)]\big).\end{aligned}$$

Consequently, the term

$$\exp(\gamma_2 - \gamma_1) = \frac{\Psi(t|\, x_j > c_j)}{\Psi(t|\, x_j \le c_j)}$$

is the factor by which the continuation ratio changes if x_j increases such that it exceeds c_j. Since the effects of the explanatory variable x_j are independent of time t, if $\gamma_2 > \gamma_1$, the continuation ratio of the second subgroup is higher at all time points. For the ST model such proportionality only holds in specific time intervals (see, for example, the tree in Fig. 9).

Regarding the choice of the link function one should note, that the logistic and the probit link function are based on density functions that are symmetric about the y-axis, whereas the Gompertz function, which is the basis of the cloglog model, is asymmetric in the sense that the right-hand asymptote of the function is approached much more gradually. For binary regression, Chen et al. (1999) recommended the use of an asymmetric link function in cases where the number of zeros and the number of ones in the data is highly unequally distributed. This is the case here, because the augmented data matrices used for tree building comprise a disproportionately high number of zero values. In addition, the Gompertz model is equivalent to the Cox proportional hazards model for continuous event times (with regard to the effects of the explanatory variables), if the original data were generated by the Cox model but only grouped event times are recorded (Tutz and Schmid 2016). Hence, the use of the cloglog link function in the scope of the proposed tree-structured hazard models might by worth investigating. For more details on link functions for binary regression models, see also Czado and Santner (1992) and Prasetyo et al. (2019).

Although only time-constant values of the explanatory variables were considered in the simulation study and the presented applications, it is also possible to deal with time-varying information. Instead of repeating the vector of explanatory variable row-wise, the vector of values measured at each time point could be entered in the corresponding rows of the augmented data matrix. The use of time-varying information, however breaks the relation between the survival function and the hazard function (2), because an observed value at t indicates that the individual must have survived up to t and thus $S(t|\, \boldsymbol{x}_{it}) = 1$.

In future research the computation of standard errors for the parameters $\hat{\gamma}$ or for the hazards $\hat{\lambda}(t|\, \boldsymbol{x}_i)$ directly, for example by bootstrap procedures, needs to be investigated. Moreover, the proposed class of models can be extended to an ensemble method, as *survival forests* for continuous (Ishwaran et al. 2008; Moradian et al. 2017; Wang et al. 2018) and discrete-time data (Bou-Hamad et al. 2011b; Schmid et al. 2020; Moradian et al. 2021), and adapted to competing risk data, similar to the approach by Berger et al. (2019).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-022-10196-x.

Acknowledgements Support by the German Research Foundation is gratefully acknowledged.

Author Contributions Conceptualization: Moritz Berger; Methodology: Moritz Berger; Formal analysis and investigation: Nikolai Spuck; Writing—original draft preparation: Nikolai Spuck; Writing—review and editing: Nikolai Spuck, Matthias Schmid and Moritz Berger; Funding acquisition: Moritz Berger; Resources: Nils Heim, Ute Klarmann-Schulz and Achim Hörauf

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the German Research Foundation (DFG; Grant number BE 7543/1-1)

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Berger, M.: TSVC: tree-structured modelling of varying coefficients. R Package Vers. **1**(2), 2 (2021)

Berger, M., Tutz, G., Schmid, M.: Tree-structured modelling of varying coefficients. Stat. Comput. **29**(2), 217–229 (2019). https://doi.org/10.1007/s11222-018-9804-8

Berger, M., Schmid, M.: Semiparametric regression for discrete time-to-event data. Stat. Model. **18**(3–4), 1–24 (2018). https://doi.org/10.1177/1471082X17748084

Berger, M., Welchowski, T., Schmitz-Valckenberg, S., Schmid, M.: A classification tree approach for the modeling of competing risks in discrete time. Adv. Data Anal. Classif. **13**(4), 965–990 (2019). https://doi.org/10.1007/s11634-018-0345-y

Biasotto, M., Pellis, T., Cadenaro, M., Bevilacqua, L., Berlot, G., Lenarda, R.D.: Odontogenic infections and descending necrotising mediastinitis: case report and review of the literature. Int. Dent. J. **54**(2), 97–102 (2004). https://doi.org/10.1111/j.1875-595x.2004.tb00262.x

Bou-Hamad, I., Larocque, D., Ben-Ameur, H.: A review of survival trees. Stat. Surv. **5**, 44–71 (2011). https://doi.org/10.1214/09-SS047

Bou-Hamad, I., Larocque, D., Ben-Ameur, H.: Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. Stat. Model. **11**(5), 429–446 (2011). https://doi.org/10.1177/1471082X1001100503

Bou-Hamad, I., Larocque, D., Ben-Ameur, H., Mâsse, L.C., Vitaro, F., Tremblay, R.E.: Discrete-time survival trees. Can. J. Stat. **37**(1), 17–32 (2009). https://doi.org/10.1002/cjs.10007

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, J.C.: Classification and Regression Trees. Taylor and Francis, Moneterey, CA Wadsworth (1984)

Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference, 2nd edn. Springer, New York, NY (2002)

Carmelli, D., Zhang, H., Swan, G.E.: Obesity and 33-year follow-up for coronary heart disease and cancer mortality. Epidemiology **8**(4), 378–383 (1997). https://doi.org/10.1097/00001648-199707000-00005

Chen, M.H., Dey, D.K., Shao, Q.M.: A new skewed link model for dichotomous qantal response data. J. Am. Stat. Assoc. **94**(448), 1172–1186 (1999). https://doi.org/10.2307/2669933

Chipman, H.A., George, E.I., McCulloch, R.E.: BART: Bayesian additive regression trees. Ann. Appl. Stat. **4**(1), 266–298 (2010). https://doi.org/10.1214/09-AOAS285

Cox, D.R.: Regression models and life tables. J. R. Stat. Soc. Ser. B (Stat. Methodol.) *34*(2), 187–220 (1972). https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Czado, C., Santner, T.J.: The effect of link misspecification on binary regression inference. J. Stat. Plan. Inference **33**(2), 213–231 (1992). https://doi.org/10.1016/0378-3758(92)90069-5

de Boor, C.: A Practical Guide to Splines. Springer, New York, NY (1978)

Debrah, A.Y., Mand, S., Narfo-Debrekyei, Y., Basta, L., Pfarr, K., Labri, J., Lawson, B., Taylor, M., Adjei, O., Hoerauf, A.: Doxycycline reduces plasma VEGF-C/sVEGFR-3 and improves pathology in lymphatic filariasis. PLoS Pathog. **9**(2), e92 (2006). https://doi.org/10.1371/journal.ppat.0020092

Dreyer, G., Addiss, D., Dreyer, P., Noroes, J.: Basic lymphoedema management: treatment and prevention of problems associated with lymphatic filariasis. Hollis Publishing Company, Hollis, NH (2002)

Eilers, P.H.C., Marx, B.D.: Flexible Smoothing with B-splines and Penalties. Stat. Sci. **11**(2), 89–121 (1996). https://doi.org/10.1214/ss/1038425655

Gordon, L., Olshen, R.A.: Tree-structured survival analysis. Cancer Treat. Rep. **69**(10), 1065–1069 (1985)

Hashimoto, E.M., Ortega, E.M.M., Paula, G.A., Barreto, M.L.: Regression models for grouped survival data: estimation and sensitivity analysis. Comp. Stat. Data Anal. **55**(2), 993–1007 (2011). https://doi.org/10.1016/j.csda.2010.08.004

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, New York, NY (2009)

Heim, N., Berger, M., Wiedemeyer, V., Reich, R., Martini, M.: A mathematical approach improves the predictability of length of hospitalization due to acute odontogenic infection. A retrospective invetigation of 303 patients. J. Cranio-Maxillofac. Surg. **47**(2), 334–340 (2019). https://doi.org/10.3844/jmssp.2019.354.365

Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: a conditional inference framework. J. Comp. Graph. Stat. **15**(3), 651–674 (2006). https://doi.org/10.1198/106186006X133933

Hothorn, T., Lausen, B.: On the exact distribution of maximally selected rank statistics. Comp. Stat. Data Anal. **43**(2), 121–137 (2003). https://doi.org/10.1016/S0167-9473(02)00225-6

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. Ann. Appl. Stat. **2**(3), 841–860 (2008). https://doi.org/10.1214/08-AOAS169

Kalbfleisch, J., Prentice, P.: The Statistical Analysis of Failure Time Data, 2nd edn. Wiley Inter-Science, New Jersey, NJ (2002)

Klein, J., Moeschberger, M.: Survival Analysis: Statistical Methods for Censored and Truncated Data. Springer, New York, NY (2003)

Kretowska, M.: Oblique survival trees in discrete event time analysis. IEEE J. Biomed. Health Inform. **24**(1), 247–258 (2019). https://doi.org/10.1109/JBHI.2019.2908773

Kuss, O., Hoyer, A.: A proportional risk model for time-to-event analysis in randomized controlled trials. Stat. Methods Med. Res. **30**(2), 411–424 (2021). https://doi.org/10.1177/0962280220953599

LeBlanc, M., Crowley, J.: Adaptive regression splines in the cox model. Biom. **55**(1), 204–213 (2004). https://doi.org/10.1111/j.0006-341x.1999.00204.x

Mand, S., Debrah, A.Y., Klarmann-Schulz, U., Basta, L., Marfo-Debrekyei, Y., Kwarteng, A., Specht, S., Belda-Domene, A., Fimmers, R., Taylor, M., Adjei, O., Hoerauf, A.: Doxycycline improves filarial lymphedema independent of filarial infection:

a randomized controlled trial. Clin. Infect. Dis. **55**(5), 621–630 (2012). https://doi.org/10.1093/cid/cis486

Meier, L., van de Geer, S., Bühlmann, P.: The Group Lasso for Logistic Regression. J. R. Stat. Soc. **70**(1), 53–71 (2008). https://doi.org/10.1111/j.1467-9868.2007.00627.x

Moradian, H., Larocque, D., Bellavance, F.: L1 splitting rules in survival forests. Lifetime Data Anal. **23**, 671–691 (2017). https://doi.org/10.1007/s10985-016-9372-1

Moradian, H., Yao, W., Larocque, D., Simonoff, J.S., Frydman, H.: Dynamic estimation with random forests for discrete-time survival data. Can. J. Stat. (published online) (2021). https://doi.org/10.1002/cjs.11639

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. **116**(44), 22071–22080 (2019). https://doi.org/10.1073/pnas.1900654116

Prasetyo, R.B., Kuswanto, H., Iriawan, N., Sutijo, B., Ulama, S.: A comparison of some link functions for binomial regression models with application to school drop out rates in east java. AIP Conf. Proc. **2194**, 020083 (2019)

Probst, P., Wright, M.N., Boulesteix, A.L.: Hyperparameters and tuning strategies for random forest. Wiley Interdisciip.: Rev. Data Min. Knowl. Discov. **9**(3), 1301 (2019). https://doi.org/10.48550/arXiv.1804.03515

Puth, M.T., Tutz, G., Heim, N., Münster, E., Schmid, M., Berger, M.: Tree-based modeling of time-varying coefficients in discrete time-to-event models. Lifetime Data Anal. **26**(3), 545–572 (2020). https://doi.org/10.1007/s10985-019-09489-7

Rancoita, P.M.V., Zaffalon, M., Zucca, E., Bertoni, F., De Campos, C.P.: Bayesian network data imputation with application to survival tree analysis. Comput. Stat. Data Anal. **93**, 373–387 (2016). https://doi.org/10.1016/j.csda.2014.12.008

Schmid, M., Küchenhoff, H., Hoerauf, A., Tutz, G.: A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. Stat. Med. **35**(5), 734–1 (2016). https://doi.org/10.1002/sim.6729

Schmid, M., Welchowski, T., Wright, M.N., Berger, M.: Discrete-time survival forests with Hellinger distance. Data Min. Knowl. Discov. **34**, 812–832 (2020). https://doi.org/10.1007/s10618-020-00682-z

Segal, M.R.: Extending the elements of tree-structured regression. Stat. Methods Med. Res. **4**(3), 219–236 (1995). https://doi.org/10.1177/096228029500400304

Segal, M.R.: Features of tree-structured survival analysis. Epidemiology **8**(4), 344–446 (1997)

Sleeper, L.A., Harrington, D.P.: Regression splines in the cox model with application to covariate effects in liver disease. J. Am. Stat. Soc. (1990). https://doi.org/10.1080/01621459.1990.10474965

Sparapani, R.A., Logan, B.R., McCulloch, R.E., Laud, P.W.: Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). Stat. Med. **35**(16), 2741–2753 (2016). https://doi.org/10.1002/sim.6893

Sparapani, R.A., Spanbauer, C., McCulloch, R.: Nonparametric machine learning and efficient computation with Bayesian additive regression trees: the BART R package. J. Stat. Software **97**(1), 1–66 (2021). https://doi.org/10.18637/jss.v097.i01

Tiendrébéogo, S., Somé, B., Kouanda, S., Gbété, S.D.: Survival analysis of data in HIV infected persons receiving antiretroviral therapy using a model-based binary tree. J. Math. Stat. **15**, 354–365 (2019)

Tutz, G., Schmid, M.: Modeling Discrete Time-to-Event-Data. Springer, New York, NY (2016)

van der Laan, M.J., Robins, J.M.: Unified Methods for Censored Longitudinal Data and Causality. Springer, New York (2003)

Wallace, M.L.: Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. Stat. Med. **33**(27), 4790–4804 (2014). https://doi.org/10.1002/sim.6261

Wang, H., Chen, X., Li, G.: Survival forests with R-squared splitting rules. J. Comp. Biol. **25**(4), 388–395 (2018). https://doi.org/10.1089/cmb.2017.0107

Welchowski, T., Berger, M., Koehler, D., Schmid, M.: discSurv: Discrete Time Survival Analysis. R package version 2.0.0 (2022)

Willet, J.B., Singer, J.D.: Investigating onset, cessation, relapse, and recovery. J. Consult. Clin. Psychol. **61**(6), 952–65 (1993). https://doi.org/10.1037/0022-006X.61.6.952

Wood, S.N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **73**, 3–36 (2011). https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S.N.: Generalized Additve Models: An Introduction with R, 2nd edn. Chapman & Hall, Boca Raton, FL (2017)

Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) *68*(1), 49–67 (2006). https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zhang, H., Singer, B.H.: Recursive Partitioning in the Health Sciences. Springer, New York, NY (1999)

3.2 Publication 2: Mathematical approach improves predictability of length of hospitalisation due to oral squamous cell carcinoma: a retrospective investigation of 153 patients

Elahi F, Spuck N, Berger M, Kramer FJ, Heim N. Mathematical approach improves predictability of length of hospitalisation due to oral squamous cell carcinoma: a retrospective investigation of 153 patients. British Journal of Oral and Maxillofacial Surgery 2023; 61: 605-611
https://doi.org/10.1016/j.bjoms.2023.09.004

Available online at www.sciencedirect.com

ScienceDirect

British Journal of Oral and Maxillofacial Surgery 61 (2023) 605–611

BRITISH Journal of Oral and Maxillofacial Surgery

www.bjoms.com

Mathematical approach improves predictability of length of hospitalisation due to oral squamous cell carcinoma: a retrospective investigation of 153 patients

Franziska Elahi [a,*], Nikolai Spuck [b], Moritz Berger [b], Franz-Josef Kramer [a], Nils Heim [a]

[a] *Department of Oral and Cranio-Maxillo and Facial Plastic Surgery, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany*
[b] *Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany*

Received 13 April 2023; revised 6 July 2023; accepted in revised form 18 September 2023
Available online 21 September 2023

Abstract

Oral squamous cell carcinoma (OSCC), a common cancer of the head and neck, is a major public health problem. The length of stay in hospital (LOS) of patients with OSCC, which can range from a few days to several months, has implications for the patient's recovery. The aim of the study was to identify and evaluate risk factors that have an impact on the prolongation of inpatient hospital stay. A four-year retrospective study reviewed hospital records of 153 inpatients with OSCC. A statistical model for discrete time-to-event data, with the LOS in hospital measured in days for which the event of interest was discharge from hospital, was applied. The model utilises a tree-building algorithm to identify relevant risk factors for a prolonged LOS. Age, type of flap, and occurrence of complications turned out to be relevant variables. Before, and on day 12, the LOS was mainly dependent on flap type and age, whereas after day 12 it was influenced by the presence of early complications. Predicting the likelihood of discharge can improve the management and resource utilisation of the healthcare system among inpatients.

Keywords: **O**ral squamous cell carcinoma; Length of hospitalization; LOS; Reconstruction

Introduction

The most common neoplasm of the oral cavity is squamous cell carcinoma (OSCC), which accounts for over 90% of all cases of head and neck cancer.[1,2] With 377,000 incident cases and 177,000 deaths related to lip and oral cavity cancer worldwide in 2020,[3] OSCC is a major public health problem. Despite diagnostic and therapeutic advances there has been no significant improvement in prognosis over the past decade, and the five-year survival rate remains low, at around 50%.[2,4,5]

Length of stay (LOS) in hospital is an important indicator of clinical severity and resource consumption. The LOS of patients with OSCC can range from a few days to several months, and these variations have implications for a patient's recovery. Prolonged LOS is associated with an increased incidence of complications, higher mortality, and delay in adjuvant therapy. It also has a negative impact on healthcare resources and hospital costs.[6–8]

The aim of the study was to identify and evaluate risk factors that have an impact on the prolongation of inpatient hospital stay. We present a statistical model that enables the prediction of LOS by revealing the most important clinical determinants.

Material and methods

Patients

A four-year retrospective study reviewed the hospital records of 153 patients who were admitted with histologi-

* Corresponding author.
E-mail addresses: ffritz1@uni-bonn.de (F. Elahi), spuck@imbie.uni-bonn.de (N. Spuck), moritz.berger@imbie.uni-bonn.de (M. Berger), franz-josef.kramer@ukbonn.de (F. -J. Kramer), Nils.Heim@ukbonn.de (N. Heim).

https://doi.org/10.1016/j.bjoms.2023.09.004

cally confirmed OSCC and treated as inpatients at the Department of Oral and Maxillofacial Plastic Surgery of the University of Bonn, Germany from January 2014 to December 2017.

The variables considered included age at diagnosis, gender, localisation of the tumour, size, spread to nearby lymph nodes and metastasis of the tumour (TNM), tumour grade, and R-classification. The flaps used for reconstruction were divided into local flap reconstruction/primary closure (rotation flap, dorsal tongue flap, transposition flap), free flap (radial forearm free flap, fibular free flap, scapular free flap, anterolateral thigh flap), and pedicle flap reconstruction (pectoralis major flap, nasolabial flap, supraclavicular flap).

Furthermore, comorbidities, HPV status, habits of smoking and/or alcohol consumption, malignancies in medical history, haemoglobin level, blood transfusion requirement, occurrence of complications during hospital stay, and necessity for and length of stay in the intensive care unit, were also evaluated.

All patients had a pathological diagnosis of OSCC and an oncological resection with subsequent inpatient hospitalisation. Exclusion criteria were outpatient care or palliative treatment only.

The length of stay was defined as the time between the day of admission until the day of discharge. Patients were usually admitted to the hospital one day before surgery.

Management of missing values

Of the 153 patients, 33 had missing values in at least one of the variables used for statistical modelling. To facilitate use of all the data in the statistical analysis, missing values were imputed using multiple imputation based on random forests.

Statistical model

The survival tree proposed by Schmid et al[9] was fitted to the imputed data to identify relevant risk factors associated with a prolonged LOS. Their model estimates the hazard for a discrete time-to-event outcome, which equals the conditional probability that an event occurs at time point t given that the event has not occurred before t, and given the values of the considered variables. Basically, the discrete hazard drives a binary variable that indicates whether the event occurred at time t or not, so parametric and tree-based modelling strategies for binary response data can be applied (see Tutz and Schmid[10] and Berger and Schmid[11] for more details).

Patients with a LOS of more than 50 days were treated as censored. Time since admission and all the aforementioned variables (except HPV status and requirement for blood transfusion) were considered as candidates for splitting during tree building. Following the approach by Schmid et al,[9] splits were selected by minimising the Gini impurity (which is widely applied in tree-based modelling), and the optimal minimal node size was determined based on the Bayesian information criterion.[12,13]

Sensitivity analysis

The imputation strategy involves a stochastic element. To assess the dependence of the results of the statistical analysis on the imputation mechanism, the model was fitted on 100 imputed versions of the data. For all repetitions, the integrated average difference between the survival functions of the corresponding model and the main model was calculated.[14]

Results

Patients

A total of 153 patients (65.4 % male, 34.6% female), mean (range) age 64.8 (32 – 94) years, were included in the study. The median (range) LOS was 15 (4-75) days. The most affected subsites were the tongue (35.8%), lower jaw (17.6%), and floor of the mouth (16.6%) (Table 1). Due to the fact that in some cases the tumour extended into several areas simultaneously, the total number of tumour locations was higher than the number of patients.

The TNM classification is shown in Table 2. In one case each, no information on nodal metastasis and R-status was provided in the records. In 15 cases no grading information could be found. In total, 49.7% of the patients underwent local flap reconstruction, 39.9% received a free flap, and 10.5% a pedicle flap.

Postoperative complications during the first 21 days occurred in 30.1% of cases. The most frequent were postoperative bleeding and flap necrosis, followed by wound dehiscence, general wound healing disorder, and formation of a fistula (Table 3). Multiple complications occurred in 16 patients (10.5%) which led to a higher number of cases.

Systemic comorbidities were found in 58.2%, of whom 41.6% had multiple comorbidities. Cardiovascular disease occurred in 58.7% of the cases, internal diseases in 12.0%, pulmonary diseases in 12.8%, and metabolic diseases in 16.5%.

Twenty-four per cent had malignancies in their medical history and 79.7% required a stay in the intensive care unit, the ICU-LOS ranging from 1-21 days (median 1 day). The perioperative dynamics of haemoglobin was Hb pre/Hb post

Table 1
Tumour localisation and frequency.

Tumour localisation	No. of occurences
Tongue	69
Lower jaw	34
Floor of the mouth	32
Cheek intraoral	17
Upper jaw	15
Lip	11
Soft palate	7
Hard palate	7
Pharynx	1
Patients in total	153

Table 2
Patient and tumour characteristics. At T0, the primary tumour was already completely removed by excisional biopsy. Surgery was performed to widen the safety margin according to the guidelines.

	Total (n=153)
Gender:	
Male	100
Female	53
Tumour size:	
T0	10
T1	56
T2	49
T3	14
T4	24
Nodal metastasis:	
N0	97
N1	20
N2	33
N3	2
Extracapsular growth	31
Metastasis:	
None	152
Distant	1
Tumour grade:	
Well differentiated	32
Moderately differentiated	78
Poorly differentiated	27
Undifferentiated	1
R status:	
R0	141
R1	11

Table 3
Frequency of complications during the 21 days after surgery.

Complications	No. of occurences
Postoperative bleeding	11
Flap necrosis	11
Ischaemia of the flap	1
Wound healing disorder	7
Abscess	3
Dehiscence	7
Fistula	6
Major haematoma	2
Postoperative acute coronary syndrome	1
Hypertensive crisis	2
Myocardial infarction	1
Cardiac decompensation	1
Haemorrhagic shock	1
Soft tissue emphysema	2
Pulmonary embolism	2
Pneumonia	2
Pneumothorax	2
Respiratory insufficiency	4
Pleural effusion	4
Hyperactive delirium	1

13.3/10.2 g/dl. Sixteen per cent of the patients needed a blood transfusion, and in 24 cases no information could be found. Thirty-two per cent of patients had therapeutic anticoagulation before surgery. HPV infection was detected in two, 22 patients had a negative HPV test, and in 129 there was no information on HPV status. A high percentage had a history of smoking (66.7%) and/or drinking (44.4%).

The results obtained from fitting the survival tree in Figure 1 show that age at diagnosis, type of flap, and occurrence of complications ($\leq$21 days after surgery) affect LOS. Specifically, the conditional probability of discharge within the first 12 days after admission in patients with free or pedicle flaps was estimated to be 0.004 if the patient was 64 years of age or younger (left most panel), and 0.016 if the patient was older than 64 years (second panel from the left). For patients with local flaps the conditional probability of discharge was 0.017 before day seven (third panel) and 0.109 from days 7 to 12 (fourth panel). After day 12 patients with early complications were less likely to be discharged (conditional probability: 0.075; fifth panel) than patients without early complications (conditional probability: 0.183; right most panel). Please note that the actual day of an early complication was not recorded and that the model fit could be refined if this information was available.

Survival probability

The panels in Figure 2 show that patients with free or pedicle flaps who were aged 64 years or younger had a median LOS of 21 days if an early complication occurred (black line, figure on the left) and a median LOS of 16 days if no early complication occurred (black line, figure on the right). The probabilities of still being on the ward after 15 days $P(T_i > 15)$ were estimated as 0.756 and 0.521 with and without the occurrence of an early complication, respectively. For patients older than 64 years with free or pedicle flaps the median LOS was 19 days with the occurrence of an early complication ($P(T_i > 15) = 0.650$; blue dashed line, figure on the left) and 15 days without the occurrence of an early complication ($P(T_i > 15) = 0.458$; blue dashed line, figure on the right). Patients with local flaps had a median LOS of 12 days independent of whether an early complication occurred ($P(T_i > 15) = 0.357$; grey dotted line, figure on the left) or not ($P(T_i > 15) = 0.246$; grey dotted line, figure on the right).

Sensitivity analysis

Of the 100 fitted survival trees, 86 were identical to the main model presented, indicating that the results were consistent across the 100 imputed data sets, and that the influence of the imputations on the model fit were minor.

The means (and standard deviations) of the integrated average differences across the 100 repetitions for the identified risk groups were -0.011 (0.003; top right panel), -0.001 (0.002; top right panel), 0.003 (0.008; centre left panel), -0.005 (0.014; centre right panel), 0.002 (0.005; bottom left panel), and 0.003 (0.007; bottom right panel) (Fig. 3).

Discussion

The type of flap used for reconstruction depends on multiple factors, including size, thickness and location of the anatomical defect, but patient-related factors such as patients' preop-

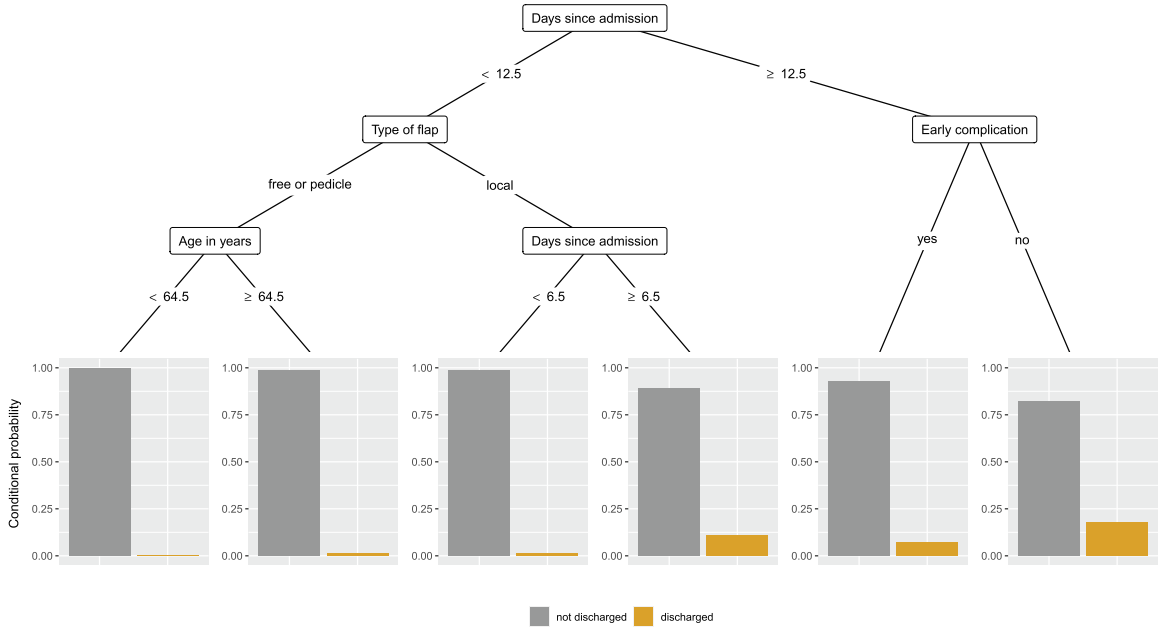


Fig. 1. Survival tree model. The panels in the leaf nodes of the tree show the conditional probabilities of discharge on day t given that the patient is still on the ward (orange bars), and conditional probabilities of not being discharged on day t (grey bars) for the identified risk groups.

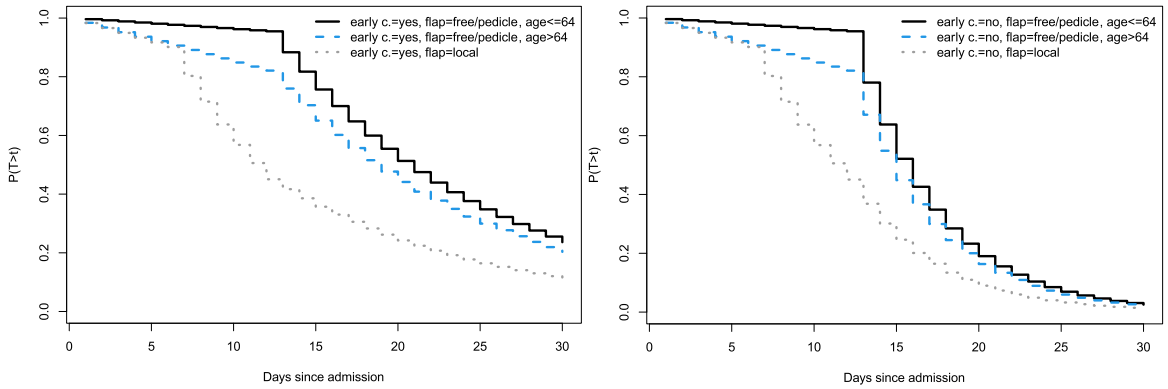


Fig. 2. Estimated survival probabilities. The survival functions $S(t|x_i) = P(T_i > t)$ for the different risk profiles were determined based on the hazards estimated from the model. The panels show the estimated survival probabilities (the probabilities of still being on the ward at day t) for the six identified risk profiles.

erative conditions, their concomitant diseases and constitutional type, also play a role.[15] Patients with a microvascular pedicle or free flap are highly unlikely to be discharged before day 12; for patients with a local flap, the probability of discharge is substantially higher, between 6 and 13 days.

Local flap techniques are suitable for the reconstruction of small surgical defects due to local anatomical limitations. Larger defects are usually covered by pedicle or free flaps.[16] Since resection and reconstruction are more extensive, the operating time is statistically longer.[17] A prolonged operating time has been shown to significantly increase the likelihood of complications, with the risk approximately doubling when the time exceeds two hours. The risk of complications also increases progressively as operating time increases.[18] In addition, the setting of several wound surfaces and the associated longer healing and mobilisation time prolong the LOS.

The average age at diagnosis in this study was 64.8 years. Advanced age is not an exclusion criterion for any surgery, but comorbidities increase the risk of complications during and after surgery, and have an impact on morbidity and mortality.[19,20] Several studies on head and neck cancer have

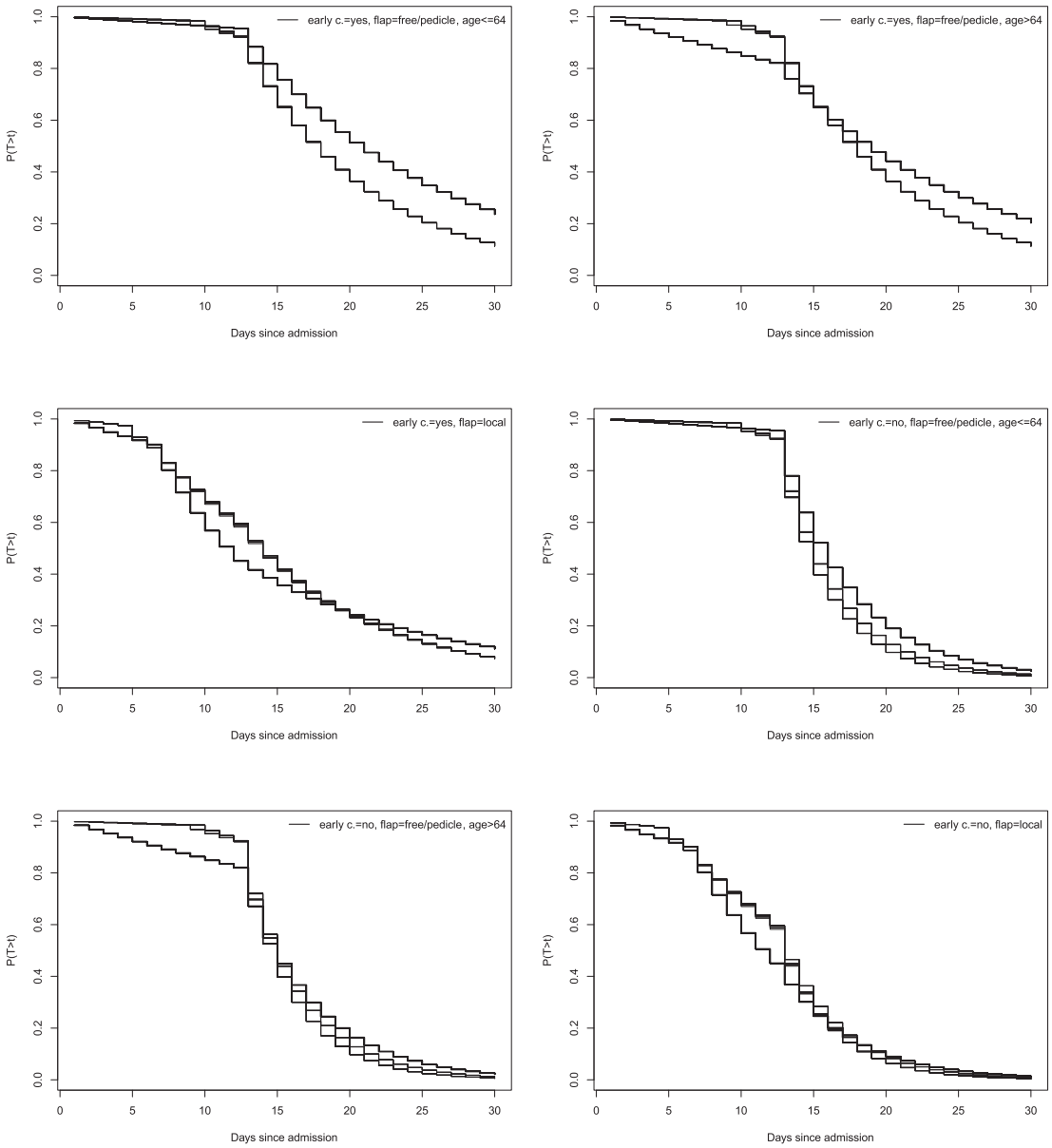


Fig. 3. Sensitivity analysis. The figures show the estimated survival probabilities for the risk groups identified by the main model determined by the 100 models fitted for sensitivity analysis. To calculate the survival probabilities for the models, where variables different from the main model were selected for splitting, categorical variables (except early complication and flap type) were set to their mode value, and the mean was used for continuous variables (except age).

reported that the American Society of Anesthesiologists status (ASA), not age, is the determining factor for the occurrence of postoperative complications.[19,21]

A correlation between duration of surgery and the presence of complications, especially in older patients, has been mentioned in several studies.[22,23] This may be attributed to the fact that older patients are often more sensitive to the effects of prolonged anaesthesia and are less able to compensate for fluid shifts and major blood loss.[24]

In our study the overall complication rate was 30.1%, with multiple complications occurring in 21 of these cases (13.7 %), which is similar to other studies.[25] As mentioned above, increases in operating time and comorbidities have an impact on the development of complications.

During data collection, complications were divided into early complications (≤21 days after surgery) and late complications (>21 days). This information must be taken into account because the actual day on which the complication occurred was not recorded. Availability of this information would have refined the estimates obtained in our study.

The median LOS in this study was 15 days, which is in the middle range of comparable studies reporting a median LOS of 10-24 days.[26–29]

Similar to our study, the occurrence of complications and type of flap were among the factors responsible for a prolonged hospital stay. Additionally, initial high T status (T3-4) and lymph node stage (N2-3), tracheotomy, ASA score of 3-4, prognostic inflammatory and nutritional index (PINI) of more than 2, transfusion, and prior radiation therapy, were highlighted as factors for a prolonged hospital stay.

The applied survival tree model identified different risk groups that each comprised patients with a certain combination of factors. A classic parametric modelling approach for the discrete hazard yields separate effect estimates for each of the included variables, which may be advantageous in many situations. Here, however, a tree-based approach was deemed favourable due to the large number of considered variables, and its ability intrinsically to select relevant variables and detect interactions.

To avoid loss of statistical power due to incomplete data, missing values were imputed, which may affect the results of the fitted model. The survival tree was fitted on 100 different versions of imputed data to investigate the dependence of the results on the imputations. This sensitivity analysis showed that the effect of the imputations on the results of the statistical analysis was strongly limited.

Because of the possible absence of other confounding variables that cannot be captured in the patient record or in this study, the study is limited by its retrospective design. In future studies, it would be worthwhile examining additional variables such as duration of surgery, necessity for tracheostomy, and classification of complication according to severity. These factors could provide further insights into their potential impact on patient outcomes and LOS.

The sample size of this study was relatively small and included patients from one institution only. A future goal should be to test the mathematical approach in cohorts with a wider, more diverse study population, which may help to confirm these findings and examine its usefulness in a broader context.

Research on LOS is critical for effective resource management and patient care. By predicting LOS based on determining factors, hospitals can more accurately predict the number of inpatient days, optimise resource allocation, and develop effective clinical pathways. For example, this knowledge can help develop personalised methods to stratify risk in patients, develop improved recovery programmes, meet patient expectations, and improve overall healthcare outcomes.

Conclusion

The LOS among patients with OSCC continues to pose a major challenge to patients' health and public healthcare resources. Using our statistical model, the probability of discharge based on the decisive criteria can be predicted, which may improve the management and resource utilisation of the healthcare system among inpatients.

Conflict of interest

We have no conflicts of interest.

Ethics statement/confirmation of patient permission

Not applicable.

References

1. De Paz D, Kao HK, Huang Y, et al. Prognostic stratification of patients with advanced oral cavity squamous cell carcinoma. *Curr Oncol Rep* 2017;**19**:65.
2. Ghantous Y, Yaffi V, Abu-Elnaaj I. Oral cavity cancer: epidemiology and early diagnosis. *Refuat Hapeh Vehashinayim (1993)* 2015;**32**:55–63, 71. In Hebrew.
3. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021;**71**:209–249.
4. Marsh D, Suchak K, Moutasim KA, et al. Stromal features are predictive of disease mortality in oral cancer patients. *J Pathol* 2011;**223**:470–481.
5. Mehrotra R, Yadav S. Oral squamous cell carcinoma: etiology, pathogenesis and prognostic value of genomic alterations. *Indian J Cancer* 2006;**43**:60–66.
6. Pirson M, Dehanne F, Van den Bulcke J, et al. Evaluation of cost and length of stay, linked to complications associated with major surgical procedures. *Acta Clin Belg* 2018;**73**:40–49.
7. Wachter RM, Goldman L. The hospitalist movement 5 years later. *JAMA* 2002;**287**:487–494.
8. Lee MK, Nalliah RP, Kim MK, et al. Prevalence and impact of complications on outcomes in patients hospitalized for oral and oropharyngeal cancer treatment. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 2011;**112**:581–591.
9. Schmid M, Küchenhoff H, Hoerauf A, et al. A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Stat Med* 2016;**35**:734–751.
10. Tutz G, Schmid M. *Modeling discrete time-to-event data. (Springer series in statistics)*. Springer; 2016.
11. Berger M, Schmid M. Semiparametric regression for discrete time-to-event data. *Statistical Modelling* 2018;**18**:322–345.
12. Breiman L, Friedman J, Olshen RA, et al. *Classification and regression trees*. Routledge; 1984.
13. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978;**6**:461–464.
14. Zhao L, Tian L, Uno H, et al. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clin Trials* 2012;**9**:570–577.
15. Vishnoi JR, Misra S. Loco regional flaps a boon for surgeons in head and neck reconstruction even in the era of microvascular flaps. *Indian J Surg Oncol* 2019;**10**:575–576.
16. Shah JP, Gil Z. Current concepts in management of oral cancer–surgery. *Oral Oncol* 2009;**45**:394–401.

17. McCrory AL, Magnuson JS. Free tissue transfer versus pedicled flap in head and neck reconstruction. *Laryngoscope* 2002;**112**:2161–2165.
18. Cheng H, Clymer JW, Po-Han Chen B, et al. Prolonged operative duration is associated with complications: a systematic review and meta-analysis. *J Surg Res* 2018;**229**:134–144.
19. Ferrari S, Copelli C, Bianchi B, et al. Free flaps in elderly patients: outcomes and complications in head and neck reconstruction after oncological resection. *J Craniomaxillofac Surg* 2013;**41**:167–171.
20. Jin F, Chung F. Minimizing perioperative adverse events in the elderly. *Br J Anaesth* 2001;**87**:608–624.
21. Coskunfirat OK, Chen HC, Spanio S, et al. The safety of microvascular free tissue transfer in the elderly population. *Plast Reconstr Surg* 2005;**115**:771–775.
22. Chick LR, Walton RL, Reus W, et al. Free flaps in the elderly. *Plast Reconstr Surg* 1992;**90**:87–94.
23. Serletti JM, Higgins JP, Moran S, et al. Factors affecting outcome in free-tissue transfer in the elderly. *Plast Reconstr Surg* 2000;**106**:66–70.
24. Mahieu R, Colletti G, Bonomo P, et al. Head and neck reconstruction with pedicled flaps in the free flap era. *Acta Otorhinolaryngol Ital* 2016;**36**:459–468.
25. Singh B, Cordeiro PG, Santamaria E, et al. Factors associated with complications in microvascular reconstruction of head and neck defects. *Plast Reconstr Surg* 1999;**103**:403–411.
26. Yang J, Wan SQ, Huang L, et al. Analysis of hospitalization costs and length of stay for oral cancer patients undergoing surgery: evidence from Hunan. *China. Oral Oncol* 2021;**119** 105363.
27. Lindeborg MM, Sethi RK, Puram SV, et al. Predicting length of stay in head and neck patients who undergo free flap reconstruction. *Laryngoscope Investig Otolaryngol* 2020;**5**:461–467.
28. Choi JE, Kim H, Choi SY, et al. Clinical outcomes of a 14-day in-hospital stay program in patients undergoing head and neck cancer surgery with free flap reconstruction under the National Health Insurance system. *Clin Exp Otorhinolaryngol* 2019;**12**:308–316.
29. Girod A, Brancati A, Mosseri V, et al. Study of the length of hospital stay for free flap reconstruction of oral and pharyngeal cancer in the context of the new French casemix-based funding. *Oral Oncol* 2010;**46**:190–194.

3.3 Unpublished Manuscript submitted to Advances in Data Analysis and Classification: Flexible tree-structured regression for clustered data with an application to quality of life in older adults

Spuck N, Schmid M, Berger M. Flexible tree-structured regression for clustered data with an application to quality of life in older adults. Unpublished manuscript uploaded to arXiv and submitted to Advances in Data Analysis and Classification 2025; 2501.12787.

https://arxiv.org/abs/2501.12787

Flexible tree-structured regression for clustered data with an application to quality of life in older adults

Nikolai Spuck[1*], Matthias Schmid[1] and Moritz Berger[1]

[1]Institute of Medical Biometry, Informatics, and Epidemiology, Medical Faculty, University of Bonn, Venusberg-Campus 1, Bonn, 53127, Germany.

*Corresponding author(s). E-mail(s): spuck@imbie.uni-bonn.de;

Abstract

Tree-structured models are a powerful alternative to parametric regression models if non-linear effects and interactions are present in the data. Yet, classical tree-structured models might not be appropriate if data comes in clusters of units, which requires taking the dependence of observations into account. This is, for example, the case in cross-national studies, as presented here, where country-specific effects should not be neglected. To address this issue, we present a flexible tree-structured approach that achieves a sparse modeling of unit-specific effects and identifies subgroups (based on individual-level covariates) that differ with regard to the outcome. The methodological advances were motivated by the analysis of quality of life in older adults using data from the survey of Health, Ageing and Retirement in Europe. Application of the proposed model yields promising results and illustrated the accessibility of the approach. A comparison to alternative methods with regard to variable selection and goodness-of-fit was performed in several simulation experiments.

Keywords: CASP score, clustered data, tree-based models, tree-structured clustering
MSC Classification: 62J02 , 62P25

Funding: Support by the German Research Foundation (DFG) , grant BE 7543/1-1 , is gratefully acknowledged.

1 Introduction

People's quality of life (QoL) is essential in evaluating and guiding many health, social, community and environmental policy actions (Bowling and Stenner, 2011). Often, QoL is of particular interest in the group of older adults since they tend to make up a larger proportion of the population in most industrialized countries each year and are most likely to experience events that negatively affect their autonomy and everyday life (Borrat-Besson et al., 2015). According to Eurostat (European Commission and Eurostat, 2024) the median age of the population in the EU increased from 39.0 years in 2003 to 44.5 years in 2023. In order to provide an explicit and well-defined measure for QoL in older adults, Hyde et al. (2003) developed the so-called Control, Autonomy, Self-Realization and Pleasure (CASP) scale comprising 19 Likert-type items on these four domains. The CASP-19 scale has become a widely applied and well-established tool in studies investigating QoL in older adults, see, among others, Sim et al. (2011), Howel (2012), Kim et al. (2015), and Frias-Goytia et al. (2024).

Here, we analyze data from the survey of Health, Ageing and Retirement in Europe, in short SHARE (see Börsch-Supan et al., 2013, for methodological details). The main objective of SHARE is to collect panel data that enables researchers to investigate the impact of socio-economic and health-related factors on the ageing process. Moreover, SHARE constitutes a cross-national survey that is aimed to explore the differences between European countries in

dealing with the consequences of population ageing. SHARE provides information on individuals aged 50 years and older gathered in 27 European countries and Israel. QoL was measured on a SHARE-specific CASP scale, which utilizes an adapted 12-item version of the CASP-19 questionnaire (Borrat-Besson et al., 2015).

When analyzing QoL in SHARE one has to deal with the issue that the data is clustered by country and therefore observations can not be treated as independent. It appears sensible to assume that measurements within units (here countries) tend to be more similar than measurements between units. This heterogeneity needs to be taken into account using an appropriate regression approach. In our application we consider a sample of $n = 45,038$ observations from the ninth wave of SHARE collected from October 2021 to October 2022 (Bergmann et al., 2024; SHARE-ERIC, 2024), which contains between 391 (Israel) and 3,116 (Belgium) observations per country. In this paper, we propose a novel approach for modeling the CASP score using a tree methodology that (i) accounts for heterogeneity between the 28 different countries by sparse modeling of country-specific effects, and (ii) is able to identify distinct subgroups of individuals which differ with regard to their CASP score based on socio-economic and health-related factors as well as their interactions.

Regression approaches for modeling heterogeneity among units are manifold. The most popular tool is *mixed effects regression*, for example, in SHARE a model with country-specific random intercepts. Mixed effects regression models postulate that the random effects follow a common predefined distribution (typically a normal distribution), which results in a parsimonious model specification (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). This strong assumption, however, comes at the price that statistical inference may be sensitive to a misspecification of the random effects distribution (Heagerty and Kurland, 2001; Litière et al., 2007). In addition, Grilli and Rampichini (2011) showed that a correlation between random effects and explanatory variables may lead to biased effect estimates. An alternative to mixed models are *fixed effects models*, in which each country has its own parameter. In the literature fixed effects models are also referred to as "no-pooling" models (Gelman and Hill, 2007) and are based on the assumption that the country-specific effects are unrelated and exist completely independently from each other (Bell et al., 2018).

To overcome both the limitations of mixed and fixed effects models, it can alternatively be assumed that the unit-specific effects follow a more flexible discrete distribution. This implies that there are groups of units sharing the same effect. In our application, the identification of groups of countries that are similar with regard to their QoL and the interpretation of relevant differences are of great interest. Clustering of units can be achieved by using finite mixtures of regression models (Grün and Leisch, 2007), by Bayesian mixed models with Dirichlet process prior (Heinzl and Tutz, 2013) and within fixed effects models applying penalized maximum likelihood estimation (Tutz and Oelker, 2017) or tree-based splits (Berger and Tutz, 2018). The latter, which we are focusing on here, is based on a fixed effects model containing tree-structured unit-specific intercepts and a linear function of a set of explanatory variables. Berger and Tutz (2018) demonstrate that their approach is very flexible in capturing heterogeneity among units particularly in scenarios where the distribution of random effects is skewed and in scenarios with correlation between random effects and covariates. Yet, the approach by Berger and Tutz (2018) is still limited as it only uses a linear combination of the explanatory variables in the predictor function. When modeling associations in SHARE the assumption of linearity may be too restrictive as it does not account for possible non-linear effects and interactions between socio-economic and health-related factors of interest (for example, level of income and chronic diseases). To address this issue, we propose a regression model extending the approach by Berger and Tutz (2018) that comprises two tree structures: One tree determining unit-specific (country-specific) effects, and one tree modeling the effects of covariates (individual-level health-related and socio-economic factors).

The underlying concept of *recursive partitioning* or *tree-based modeling* originates from the framework of classification and regression trees (CART) proposed by Breiman et al. (1984). When growing a classical tree the predictor space is partitioned into a set of disjoint subsets by sequentially applying binary splits. In each subset a simple model (for example, a constant) is fitted. Overviews and comparisons of recursive partitioning methods have been given by Strobl et al. (2009), Doove et al. (2014) and Kern et al. (2019). The tree methodology applied here (see Section 3 for a detailed description of the algorithm) slightly differs from theses approaches, as we do not apply a traditional recursive partitioning algorithm, but fitting and tree building is performed within the framework of fixed effects models. The key advantages of our proposed

model are (i) the flexibility in capturing the effects of individual-level factors (including non-linear effects and interactions), (ii) its built-in mechanism to select the relevant factors, and (iii) sparse modeling of unit-specific effects assuming a discrete distribution.

The remainder of this article is structured as follows: In Section 2 we introduce the notation, describe the proposed tree-structured model and discuss alternatives based on random effects. Details of the fitting procedure are outlined in Section 3. In Section 4 we apply the proposed model for analyzing the CASP score in the SHARE data. In Section 5 the proposed model is compared to alternative methods based on several simulation experiments. The article concludes with a summary and discussion on the different methods for modeling heterogeneity (Section 6).

2 Regression for clustered data

Consider clustered data with n units given by $(y_{ij}, \boldsymbol{x}_{ij})$, $i = 1, \ldots, n$, $j = 1 \ldots, n_i$, where y_{ij} denotes the value of the outcome variable of observation j from unit i and $\boldsymbol{x}_{ij}^\top = (x_{ij1}, \ldots x_{ijp})$ denotes the vector of a set of covariates. In general, it is assumed that the values of the covariates vary within units and that the number of observations per unit n_i may differ across units. In the following, alternative parametric and non-parametric approaches for modeling clustered data are considered. The focus is mainly on models with unit-specific intercepts.

2.1 Models with random effects

In classical *generalized linear mixed effects models* (GLMMs; Verbeke and Molenberghs, 2000) with random intercepts, the expectation of the outcome variable $\mu_{ij} = \mathbb{E}(y_{ij}|\boldsymbol{x}_{ij}, b_i)$ is linked to the covariates in the form

$$g\left(\mu_{ij}\right) = \eta(\boldsymbol{x}_{ij}, b_i) = \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b_i\,, \tag{1}$$

where $g(\cdot)$ denotes a suitable link function, $\boldsymbol{\beta}$ is the vector of regression coefficients (that is, the vector of fixed effects of the covariates) and b_i denotes the random intercept of unit i. It is commonly assumed that the random intercepts follow a normal distribution, i.e. $b_i \sim N(0, \sigma_b^2)$. This distributional assumption on the random intercepts makes the GLMM very efficient, as only the variance parameter has to be estimated in the random effects part of the model.

The simple form of the GLMM in Equation (1) comes with the drawback that only linear main effects of the covariates on the outcome are assumed. This, however, may be too restrictive in real-world data (for example, in our application to SHARE), as it does not account for possible non-linear effects and interactions between covariates. To address this issue Hajjem et al. (2011) and Sela and Simonoff (2012) simultaneously proposed a flexible non-parametric approach using recursive partitioning. Their approaches, referred to as mixed effects regression trees (MERT) and RE-EM trees, respectively, combine a simple random intercept model with a standard regression tree. The predictor function of a RE-EM tree can be written as

$$\eta(\boldsymbol{x}_{ij}, b_i) = \beta_0 + tr(\boldsymbol{x}_{ij}) + b_i\,, \tag{2}$$

where the function $tr(\cdot)$ is determined by a tree structure. This means that $tr(\cdot)$ sequentially partitions the observations into disjoint subsets $N_m, m = 1, \ldots, M$, based on the values of the covariates and assigns a constant γ_m to each subset N_m (by averaging the respective outcome values). The constant γ_m can also be interpreted as the regression coefficient in N_m. Hence, the function $tr(\cdot)$ is given by

$$tr(\boldsymbol{x}_{ij}) = \sum_{m=1}^{M} \gamma_m\, I(\boldsymbol{x}_{ij} \in N_m)\,, \tag{3}$$

where $I(\cdot)$ denotes the indicator function. Importantly, higher-order interactions can be captured by the tree in a very flexible way. RE-EM trees are fitted iteratively by alternating between two steps: (i) Fitting the tree $tr(\boldsymbol{x}_{ij})$ by applying the CART algorithm, while keeping the random effects fixed, and (ii) estimating the random intercepts, while keeping the tree structure fixed. More recently, Fu and Simonoff (2015) introduced an adapted version of the RE-EM tree that applies conditional inference trees instead of CART. In a similar vein, an flexible tree-based approach building on the framework of conditional inference trees has been proposed by Fokkema et al. (2018).

Both, GLMMs and RE-EM trees specify normally-distributed random intercepts to describe the heterogeneity between units. This is useful if the focus mainly is on the effects of the covariates (particularly, in scenarios, where $n \ggg n_i$). Yet, in our analysis of the CASP score in SHARE, we are explicitly interested in cross-national differences, that is, in the country-specific effects. We therefore propose to use a fixed effects model instead, which is outlined in the next section.

2.2 Models with tree-structured fixed effects

An alternative to the mixed effects models introduced in the previous section, are *fixed effects models* (FEMs) with predictor function

$$\eta(\boldsymbol{x}_{ij}, \beta_{0i}) = \beta_{0i} + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}\,, \tag{4}$$

where each unit has its own parameter β_{0i}. The specification of one parameter per unit can easily turn into problems, because it results in a very large number of coefficients, which affects estimation accuracy and complicates the interpretation of effects. For example, in wave 9 of SHARE 28 country-specific intercepts need to be estimated when using the FEM in (4). To deal with this issue, Berger and Tutz (2018) proposed a tree-structured FEM, which assumes that there are groups of units that share the same effect on the outcome variable. Building clusters of units highly reduces the number of parameters and increases stability of the estimates. The tree-structured FEM by Berger and Tutz (2018) has the form

$$\eta(\boldsymbol{x}_{ij}) = tr_0(i) + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}\,, \tag{5}$$

where $tr_0(\cdot)$ describes the unit-specific intercepts represented by a tree structure. The tree forms clusters of units with the same effect on the outcome and is given by

$$tr_0(i) = \sum_{c=1}^{C} \beta_{0c}\, I(i \in N_{0c})\,, \tag{6}$$

where C denotes the number of identified clusters N_{0c} and β_{0c} is the corresponding cluster-specific intercept. To obtain $tr_0(i)$, the observations are sequentially partitioned into disjoint subsets using the unit number as the only covariate, while the other parameters (effects of the covariates) are fitted simultaneously to all observations. Berger and Tutz (2018) proposed to treat the unit number as ordinal variable by ordering the units with respect to their means of the outcome variable before tree building.

Just like the GLMM, the tree-structured FEM in Equation (5) is restricted to linear main effects of the covariates, only. To overcome this limitation and inspired by the works of Hajjem et al. (2011) and Sela and Simonoff (2012) on mixed effects models, we propose a tree-structured FEM, where the effects of the covariates are also determined by a tree structure. Specifically, the predictor function of our proposed model contains two trees and can be written as

$$\eta(\boldsymbol{x}_{ij}) = tr_0(i) + tr(\boldsymbol{x}_{ij})\,, \tag{7}$$

where $tr(\cdot)$ and $tr_0(\cdot)$ are defined as in Equations (3) and (6), respectively. The model in (7) is constructed in a stepwise procedure, where in each step either a split in the tree of the covariates $tr(\cdot)$ or in the tree that determines the clustering of units $tr_0(\cdot)$ is performed. The starting point is a simple model with a global intercept, only. Assuming that a split in x_k at split point c_k is selected in the first step, results in a model with predictor function

$$\eta^{[1]}(\boldsymbol{x}_{ij}) = \beta_0^{[1]} + \gamma_1^{[1]} I(x_{ijk} \leq c_k)\,, \tag{8}$$

where $\beta_0^{[1]}$ is a global intercept and $\gamma_1^{[1]}$ is the effect on the outcome in the left node. Note that the right node $\{x_{ijk} > c_k\}$ in $tr(\cdot)$ serves as a reference node to ensure parameter identifiability. In the second step, either one of the current nodes is split further or a split with regard to the one of the units in the intercept tree is performed. Let us assume splitting the units into the clusters N_{01} and N_{02} is the second step. This yields the predictor function

$$\eta^{[2]}(\boldsymbol{x}_{ij}) = \beta_{01}^{[2]} I(i \in N_{01}) + \beta_{02}^{[2]} I(i \in N_{02}) + \gamma_1^{[2]} I(x_{ijk} \leq c_k)\,, \tag{9}$$

where $\beta_{01}^{[2]}$ and $\beta_{02}^{[2]}$ are the unit-specific intercepts in the two selected nodes and $\gamma_1^{[2]}$ is an update of the parameter from the previous iteration. To determine the split in $tr_0(\cdot)$ the unit number is treated as an ordinal variable (see Section 3 for details on the fitting procedure). A third split in $tr(\cdot)$ with regard to x_ℓ at split point c_ℓ in the left node then results in a predictor of the form

$$\begin{aligned}\eta^{[3]}(\boldsymbol{x}_{ij}) =& \beta_{01}^{[3]} I(i \in N_{01}) + \beta_{02}^{[3]} I(i \in N_{02}) \\ &+ \gamma_1^{[3]} I(x_{ijk} \leq c_k) I(x_{ij\ell} \leq c_\ell) + \gamma_2^{[3]} I(x_{ijk} \leq c_k) I(x_{ij\ell} > c_\ell)\,, \end{aligned} \tag{10}$$

with the new effects $\gamma_1^{[3]}$ and $\gamma_2^{[3]}$.

It is important to note that the coefficients of the tree-structured model in (7) can only interpreted with regard to a reference node. For example, if the outcome variable y_i is metrically scaled and $g(\cdot)$ is the identity link, the coefficient $\beta_{01}^{[3]}$ denotes the expected values of the outcome variable in cluster N_{01} given that $x_{ijk} > c_k$ (that is, for the subgroup in the reference node). Analogously, the coefficients $\gamma_1^{[3]}$ and $\gamma_2^{[3]}$ determine the effects on the outcome variable compared to the subgroup in the reference node. To allow for a more intuitive interpretation of the model coefficients, we propose to apply the adjustment

$$\begin{aligned}\tilde{\beta}_{0c} &= \beta_{0c} + \bar{\gamma} \quad \text{and} \\ \tilde{\gamma}_m &= \gamma_m - \bar{\gamma}\,, \end{aligned} \tag{11}$$

where $\bar{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} tr(\boldsymbol{x}_{ij})$ denotes the mean of the coefficients in the tree of the covariates across all individuals. The coefficients $\tilde{\beta}_{0c}$ can then be interpreted as the *average cluster-specific intercepts* and the coefficients $\tilde{\gamma}_m$ represent subgroup-specific effects compared to these averages. In case of a metrically scaled outcome variable (see also our application to SHARE in Section 4) this translates into the expected values for each cluster ($\tilde{\beta}_{0c}$) and subgroup-specific deviations from these expectations ($\tilde{\gamma}_m$). More details on the fitting procedure are given in the next section.

3 Fitting procedure

In each step of the tree-building algorithm, the best split among all candidate variables (that is, one component of $\boldsymbol{x}$ or the unit number i) and among all possible split points is selected, starting from a predictor function with global intercept, only. For this, all possible models with one additional split in either the tree of the covariates $tr(\cdot)$ or the tree that determines the clustering of units $tr_0(\cdot)$ are evaluated and the best performing one yielding the smallest deviance is selected. In FEMs the deviance is a quite natural measure of the model fit. This criterion is also equivalent to minimizing the entropy, which has been used as a splitting criterion already in the early days of tree construction (Breiman et al., 1984). Note that, in contrast to common trees, in each step of the algorithm all the observations are used to derive estimates of the model parameters. Hence, all parameters are refitted in each iteration and no previously estimated parameters are kept. This ensures that one obtains valid estimates of the two tree components (the coefficient estimates of either of the two components are adjusted for the change through a split in the other) together with the splitting rule.

When selecting the first split in $tr_0(\cdot)$ with regard to the unit number, one has to consider 2^{n-1} possible partitions, which may be a very large number. To avoid this exponential computational cost, we instead order the units with respect to their means of the outcome variable $\bar{y}_i$ beforehand and treat the unit number as an ordinal variable during tree building. Therefore, only $n-1$ possible splits have to be considered. This approach, which has also been used by Berger and Tutz (2018), has been shown to work well in earlier research, see Breiman et al. (1984) and Ripley (1996) for binary outcomes and Fisher (1958) for continuous outcomes.

To determine the optimal number of splits and hence the size of the trees, we use a post-pruning strategy, where a large number of splits $S_{\max}$ is carried out first and afterwards the trees are pruned to an adequate size to prevent overfitting. Running the stepwise algorithm (with a sufficiently large number of splits) results in a sequence of nested models. These models can be evaluated with regard to their goodness of fit applying a likelihood-based criterion. Specifically, we suggest to select the optimal number of splits by maximizing the cross-validated predictive

log-likelihood. In the simulation study and our application to SHARE, we use 10-fold cross-validation and additionally apply the one standard error (1SE) rule. That is, one selects the model yielding a cross-validated log-likelihood value within one standard error of the model with the maximal value. This is in line with the algorithm by Sela and Simonoff (2012). Subsequently, the final model with the selected number of splits is fitted to the entire data.

To prevent the algorithm from building extremely small nodes (with only a few observations), an additional *minimal bucket size* constraint n_{mb} may be applied. With the minimal bucket size constraint, the minimum number of observations required in any terminal node is limited downward.

To summarize, the following steps are performed during the fitting procedure:

1. **Ordering of units:** Order the units $i \in \{1, \ldots, n\}$ according to the average values of the outcome variable in each unit $\bar{y}_i$ and initialize the corresponding ordinal variable.
2. **Initial model:** Fit the model without any covariates, yielding a single estimate of the intercept $\hat{\beta}_0$.
3. **Tree building:** Set $s = 1$.
   (a) Fit all candidate models with one additional split regarding one of the covariates or the unit i, that fulfill the minimal bucket size constraint, in one of the already built nodes. If none of the additional splits meets the minimal bucket size constraint, terminate the algorithm.
   (b) Select the best performing model based on the minimal deviance.
   (c) Fit the selected model and set $s = s + 1$. If $s < S_{\text{max}}$, continue with step (3a).
4. **Post-pruning:** Select the optimal model from the sequence of nested models generated in steps (2) and (3) based on the predictive log-likelihood applying k-fold cross-validation with the 1SE rule. Then fit the model with the corresponding number of splits to the complete data set.

Technically, the proposed algorithm can be embedded into the framework of tree-structured varying coefficients models (TSVC; Berger et al., 2019). The models can therefore be fitted by the eponymous R add-on package **TSVC** (Berger, 2023), where the covariates $\boldsymbol{x}$ and the unit number i serve as effect modifiers, modifying the effect of constant auxiliary variables.

4 Application: Quality of life in SHARE

SHARE is a longitudinal, cross-national survey that collects data from individuals aged 50 years and older living in the European Union and Israel (Börsch-Supan et al., 2013). Data collection for the first wave of SHARE started in 2004 in 19 different countries. Since then a total of nine waves have been conducted. The survey was mainly designed to provide information on how socio-economic and health-related factors influence the aging process. Here, we analyze data from the ninth wave collected from October 2021 to October 2022 across 28 countries (Bergmann et al., 2024; SHARE-ERIC, 2024). The objective of our analysis was the flexible modeling of QoL in terms of the CASP score by (i) accounting for country-specific effects in a sparse way, and (ii) identifying subgroups of individuals which differ in their CASP score based on socio-economic and health-related factors.

In a preliminary step, for households with more than one individual participating in the survey one representative was selected at random. This resulted in an analysis data set of $n = 45,038$ individuals from 28 countries. Figure 1 shows the distribution of individuals included in the analysis by country. The country with the largest number of participants was Belgium with $n = 3,116$, whereas only $n = 391$ participants from Israel were eligible for our analysis (which constitutes the lowest number of participants). The individual-level factors considered for modeling were: sex, age (in years), number of people living in the household, number of children, number of chronic diseases, educational level, employment status, and the level of income (the income decile which the household falls in by country). Summary statistics of these factors are given in Table 1. For more details on the ninth wave of SHARE, see also Bergmann et al. (2024) and SHARE-ERIC (2024).

We fitted the proposed tree-structured FEM (7) to the analysis data set, where the socio-economic and health-related factors presented in Table 1 and level of income were considered as covariates in $tr(\cdot)$ and the countries were treated as the units in $tr_0(\cdot)$. The maximal number of splits considered was $S_{\text{max}} = 20$, and the optimal number of splits was selected based on the

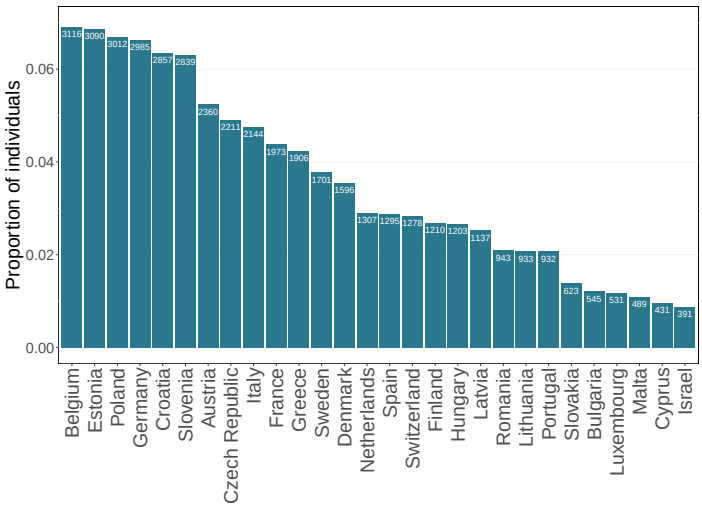


Fig. 1: Analysis of the SHARE data: Distribution of individuals by country. Absolute and relative frequencies of individuals per country included in the analysis data set

Table 1: Analysis of the SHARE data. Summary statistics of the individual-level factors included in the analysis

Variable		**Summary statistics**					
		x_{min}	$x_{0.25}$	x_{med}	$\overline{x}$	$x_{0.75}$	x_{max}
Age		50	62	69	69.3	76	105
Household size		1	1	2	1.9	2	11
Number of children		0	1	2	2.0	3	17
Number of chronic diseases		0	1	2	1.9	3	14
Sex	Male (0)						18 166 (40.3%)
	Female (1)						26 872 (59.7%)
Education	Pre-education (0)						1 123 (2.5%)
	Primary education (1)						5 242 (11.6%)
	Secondary education first stage (2)						7 214 (16.0%)
	Secondary education second stage (3)						17 919 (39.8%)
	Post-secondary education (4)						2 295 (5.1%)
	Tertiary education first stage (5)						10 864 (24.1%)
	Tertiary education second stage (6)						391 (0.9%)
Employment	Unemployed or retired (0)						35 197 (78.1%)
	Employed or self employed (1)						9 841 (21.9%)

10-fold cross-validation with the 1SE rule. The minimal bucket size was set to $n_{\text{mb}} = 100$ and the maximal depth of the tree to $d_{\max} = 4$.

Figure 2 visualizes the results with regard to $tr_0(\cdot)$. Five clusters of countries were identified when fitting the model: The first cluster with the lowest expected CASP score is the smallest comprising only two countries (Bulgaria and Greece). The cluster with the second lowest expected QoL contains Eastern and Southern European countries (Cyprus, Italy, Latvia, Lithuania, and Romania). Central to Eastern European as well as the countries from the Iberian peninsula comprise the third cluster (Croatia, Czech Republic, Estonia, Hungary, Israel, Slovakia, Spain, Poland, and Portugal). The cluster with the second highest expected QoL is composed mostly of Central European as well as Scandinavian countries (Belgium, Finland, France, Germany, Slovenia, and Sweden). Finally, the cluster with the highest expected CASP score contains the five countries Austria, Denmark, Luxembourg, the Netherlands, and Switzerland. These results indicate

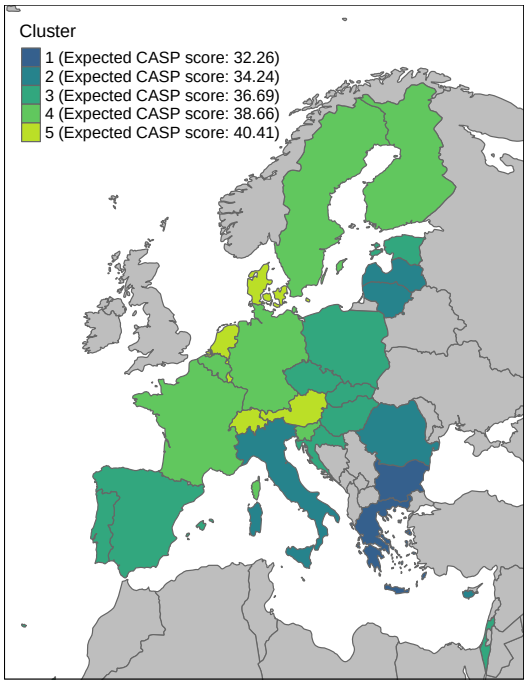


Fig. 2: Analysis of the SHARE data. Identified clusters of countries in $tr_0(\cdot)$ when fitting the TTSC model

that the populations of wealthier countries tend to experience a higher QoL, which was shown previously in a study by Diener and Diener (1995) based on data from 101 nations. Specifically, Greece and Bulgaria, which constitute the cluster with lowest expected CASP score, exhibited the lowest gross domestic product (GDP) per capita of all countries in the EU in 2021, whereas Luxembourg, Denmark, the Netherlands, and Austria (i.e. EU countries in the fifth cluster) exhibited the highest, third, fourth and seventh highest GDP, respectively (European Commission and Eurostat, 2024). In addition, Niedzwiedz et al. (2014) analyzed data from wave 2 and 3 of SHARE and found that older adults from countries with more generous welfare regimes experienced higher QoL, which is confirmed by our findings: Scandinavian countries and countries with Bismarckian welfare regimes (e.g. Austria, France, Germany, and Switzerland) were placed in the the upper two clusters, while countries with Southern or Post-communist welfare regimes were mostly in clusters with lower expected QoL.

Figure 3 shows the results with regard to $tr(\cdot)$. Number of chronic diseases, level of income, and age of the individuals were selected as splitting variables during tree building and in total eleven different subgroups were identified. In particular, the number of chronic diseases was shown to have a very strong effect on QoL, as it was the first splitting variable in the root node and was selected for splitting most often. The corresponding results indicate that an increasing number of chronic diseases is associated with a decreased QoL, which aligns with the findings by Heyworth et al. (2009) and Rothrock et al. (2010), who investigated the effect of chronic conditions on health-related QoL in United Kingdom (UK) and United States (US) citizens, respectively. Moreover, negative associations between the number of chronic diseases and QoL were frequently reported in the past decades (Marengoni et al., 2011; Makovski et al., 2019) and were also found in data from previous waves of SHARE (Makovski et al., 2020; Rodríguez-Bláquez et al., 2020). Moreover, household income is demonstrated to play an important role, where adults who are among the wealthier parts of the population of their respective country showed higher QoL. The positive effect of income on QoL was previously shown by Killingsworth (2021) in US citizens and von dem Knesebeck et al. (2007) in wave 1 of SHARE. Age was also selected for splitting but appeared to be only relevant for adults suffering from at least one chronic disease.

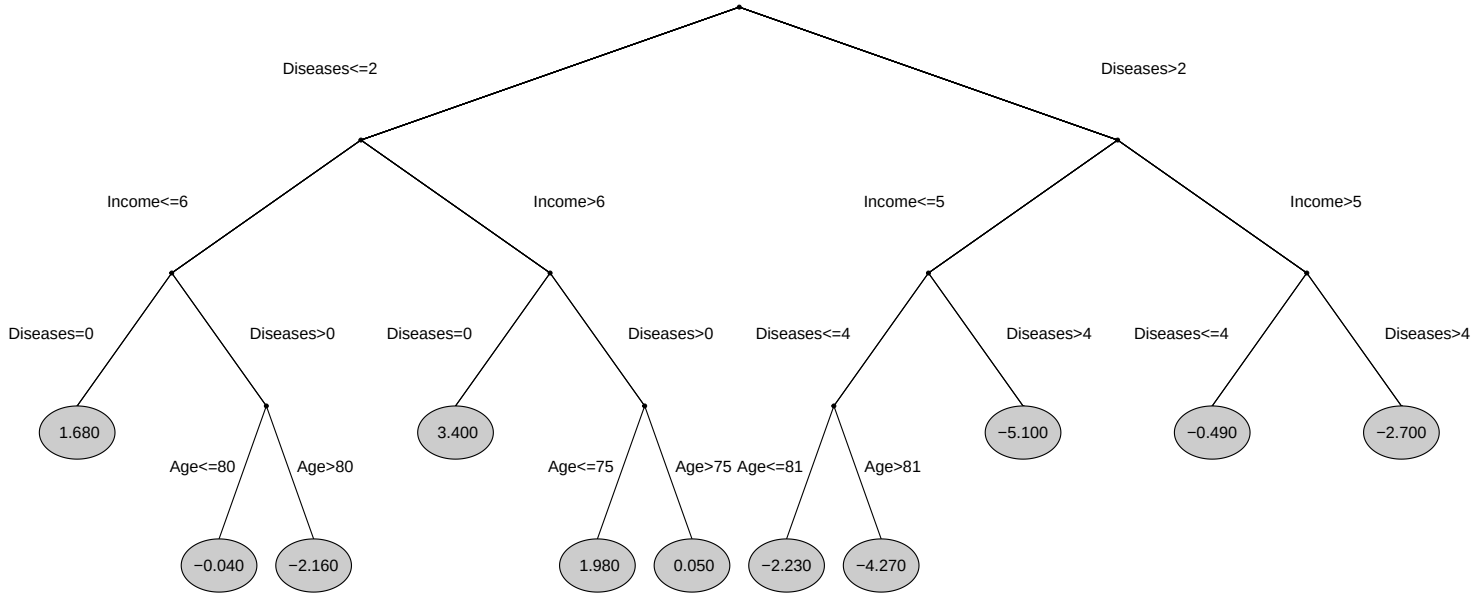


Fig. 3: Analysis of the SHARE data. Identified subgroups of individuals in $tr(\cdot)$ when fitting the TTSC model. *Diseases*, *Income*, and *Age* refer to the number of chronic diseases the person suffers from, the income decile the household falls in, and the age in years, respectively

From Figure 3 it is seen that the subgroup with the lowest expected QoL (among the people aged 50 years or older) constitutes individuals who suffer from more than four chronic diseases and are among the poorest 50 percent in terms of household income in their country. Individuals from this subgroup are expected to exhibit a by 5.10 points lower CASP score than the expected value of their country. On the other end, individuals with no chronic diseases that were among the wealthiest 40 percent of older adults from their country are shown to experience the highest QoL at a by 3.40 points increased CASP score compared to the expected value of their country.

5 Simulation study

To assess the performance and further analyze the properties of the proposed model, we considered different simulation scenarios. The simulation study was aimed to (i) investigate how the performance is affected by specific characteristics of the data, like the form of the data generating process (DGP; linear or tree-based), the number of units, the number of individuals per unit, and the number of covariates, and (ii) to compare the proposed tree-structured model (7) to alternative models.

5.1 Simulation design

We considered four simulation scenarios that were based on a DGP with predictor (1) comprising linear effects of the covariates and random unit-specific intercepts (*scenario 1*), a DPG with predictor (2) comprising tree-structured effects of the covariates and random unit-specific intercepts (*scenario 2*), a DGP with predictor (5) comprising linear effects of the covariates and clustered fixed effects of the units (*scenario 3*), and a DGP with predictor (7) composed of tree-structured effects of the covariates and clustered fixed effects of the units (*scenario 4*). Further details on the DGPs will be given in the following subsections.

In all simulation scenarios, we considered six different settings and performed 100 replications each. In the first setting (*setting 1*), we simulated data with $n = 20$ units, $n_i = 50$ observations per unit and $p = 10$ potentially informative covariates. We included six metrically scaled covariates $X_1, \ldots, X_6 \sim N(0, 1)$ and four binary covariates $X_7, \ldots, X_{10} \sim \text{Bin}(1, 0.5)$. Standard normally distributed error terms were included in the DGP. In the following we also refer to this first setting as *base setting*. In the five other settings only one parameter compared to the base setting was modified to generate the data, while all the other parameters were kept fixed. In *setting 2* and *setting 3* we modified the ratio of units compared to the observations per units, setting $n = 40/n_i = 25$ and $n = 100/n_i = 10$, respectively. We considered a higher dimensional covariate space with $p = 100$, $X_{11}, \ldots, X_{15} \sim N(0, 1)$, and $X_{16}, \ldots, X_{100} \sim \text{Bin}(1, 0.5)$ in *setting 4*. In *setting 5* the variance of the error terms was increased to $\sigma^2_\varepsilon = 2$. The last setting (*setting 6*) differs depending on the specific scenario and is described in the respective subsections.

The following models were fitted to the simulated data in each scenario:

(i) the linear mixed model (1) with linear effects of the covariates and random unit-specific intercepts (*LMM*),
(ii) a LMM with variable selection by LASSO as proposed by Groll and Tutz (2014), which applies an L_1-penalty on the linear effects of the covariates (*LMMP*),
(iii) the RE-EM tree (2) by Sela and Simonoff (2012) with tree-structured effects of the covariates and random unit-specific intercepts (*RE-EM*),
(iv) the LMM tree by Fokkema et al. (2018), which also has the form in Equation (2), but compared to RE-EM applies the framework of conditional inference trees (*LMMT*),
(v) the tree-structured FEM (5) by Berger and Tutz (2018) with linear effects of the covariates and tree-structured fixed effects of the units (*LTSC*),
(vi) a LTSC model, with variable selection applying backward selection (*LTSCB*),
(vii) the proposed tree-structured FEM (7) with tree-structured effects of the covariates and tree-structured fixed effects of the units (*TTSC*),
(viii) a model without any covariates and only a constant global intercept (*Null*), and
(ix) the true data-generating model (*Perfect*).

The LMMT model by Fokkema et al. (2018) implements a fitting procedure similar to the RE-EM tree, where the algorithm alternates between two steps: (i) Fitting the tree structure, while keeping the random effects fixed, and (ii) estimating the random effects, while keeping the tree structure fixed. Instead of the CART algorithm, the LMMT model applies conditional inference

trees (Hothorn et al., 2006). That is, in each iteration a test for parameter instability is carried out for each covariate and the covariate showing the strongest association with the outcome variable is selected for splitting (if it is significant at a predefined significance level α). The approach by Fokkema et al. (2018) is based on the framework of model-based recursive partitioning (Zeileis et al., 2008), which additionally allows that in each terminal node of the tree a separate regression model is fitted. Here, we specified an intercept-only model to ensure comparability. Note that this is conceptually equivalent to the conditional inference-based version of the RE-EM tree by Fu and Simonoff (2015).

For the random intercepts in LMM, LMMP, RE-EM, and LMMT normality was assumed. The optimal penalty parameter λ for the LASSO in LMMP was selected based on the Bayesian information criterion (BIC; Schwarz, 1978). The optimal number of splits in RE-EM, LTSC, LTSCB, and TTSC was selected based on 10-fold cross-validation with the 1SE rule (see also Section 3). The minimal bucket size (i.e. the minimum number of observation required in a node) was set to $n_{\text{mb}} = \lfloor 0.1 \cdot \sum_{i=1}^{n} n_i \rfloor$ in all tree-based models (RE-EM, LMMT, LTSC, LTSCB, and TTSC). For the LTSCB model, first the LTSC model with the optimal number of splits was fitted and subsequently covariates were excluded from the linear predictor using backward selection based on BIC, while the tree structure of the unit-specific effects was kept fixed. Note that the perfect model cannot be fitted in practice as it is unknown and serves as reference, only.

5.2 Evaluation criteria

To assess the performance of the competing models in terms of goodness-of-fit, we considered the root mean squared error (RMSE) separately for the effects of the covariates and for the unit-specific effects. The RMSE of the covariate effects was calculated by

$$\text{RMSE}_{\text{X}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\tilde{\eta}_{\text{X}}(\boldsymbol{x}_{ij}) - \tilde{\hat{\eta}}_{\text{X}}(\boldsymbol{x}_{ij})\right)^2},$$

where $\tilde{\hat{\eta}}_{\text{X}}(\boldsymbol{x}_{ij}) = \hat{\eta}_{\text{X}}(\boldsymbol{x}_{ij}) - \frac{1}{n}\sum_{i'=1}^{n}\frac{1}{n_{i'}}\sum_{j'=1}^{n_{i'}}\hat{\eta}_{\text{X}}(\boldsymbol{x}_{i'j'})$ corresponds to the covariate-specific deviation from the unit-specific expectation (also compare the adjustment of the coefficients in Section 2.2). Specifically, for the models with linear effects of the covariates (LMM, LMMP, LTSC, and LTSCB) we have that $\hat{\eta}_{\text{X}}(\boldsymbol{x}_{ij}) = \hat{\beta}_1 x_{ij1} + \cdots + \hat{\beta}_p x_{ijp}$ and for the models with tree-structured effects of the covariates (RE-EM, LMMT, and TTSC) we have that $\hat{\eta}_{\text{X}}(\boldsymbol{x}_{ij}) = \hat{tr}(\boldsymbol{x}_{ij})$. The RMSE of the unit-specific effects was calculated by

$$\text{RMSE}_{\text{I}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\tilde{\eta}_{\text{I}}(i) - \tilde{\hat{\eta}}_{\text{I}}(i)\right)^2},$$

where $\tilde{\hat{\eta}}_{\text{I}}(i) = \hat{\eta}_{\text{I}}(i) + \frac{1}{n}\sum_{i'=1}^{n}\frac{1}{n_{i'}}\sum_{j'=1}^{n_{i'}}\hat{\eta}_{\text{X}}(\boldsymbol{x}_{i'j'})$ corresponds to the expected outcome value of unit i. For models with random unit-specific intercepts (LMM, LMMP, RE-EM, and LMMT) this means $\hat{\eta}_{\text{I}}(i) = \hat{\beta}_0 + b_i$ and for models with tree-structured fixed effects (LTSC, LTSCB, and TTSC) this means $\hat{\eta}_{\text{I}}(i) = \hat{tr}_0(i)$. Of note, for TTSC $\tilde{\eta}_{\text{X}}(\cdot)$ and $\tilde{\eta}_{\text{I}}(\cdot)$ could also directly be derived from the adjusted coefficients defined in (11). The true values of $\tilde{\eta}_{\text{X}}(\cdot)$ and $\tilde{\eta}_{\text{I}}(\cdot)$ are determined analogously based on the true values.

In addition, true positive rates (TPR) and false positive rates (FPR) for the covariates were considered. The TPR is the proportion of informative covariates that were correctly identified to have an effect on the outcome variable and is given by

$$\text{TPR}_X = \frac{1}{|\{k : \vartheta_k = 1\}|}\sum_{k:\vartheta_k=1} I(\hat{\vartheta}_k = 1)\,,$$

where $\vartheta = 1$ if X_k has an effect on the outcome variable and $\vartheta_k = 0$ otherwise. The FPR specifies the proportion of noise variables that were falsely identified to have an effect on the outcome variable. It is given by

$$\text{FPR}_X = \frac{1}{|\{k : \vartheta_k = 0\}|}\sum_{k:\vartheta_k=0} I(\hat{\vartheta}_k = 1)\,.$$

5.3 Linear DGP with random unit-specific intercepts

The first scenario was based on a DGP of the form

$$y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_7 x_{ij7} + b_i + \varepsilon_{ij}\,,$$

with $\beta_1 = 0.8$, $\beta_2 = 0.4$, and $\beta_7 = 0.8$. Hence, three out of ten (or one hundred) covariates were informative. The random unit-specific intercepts b_i follow a standard normal distribution. The data sets in settings 1 to 5 were generated as described above. In setting 6, we assumed that a correlation between X_1 as well as X_2 and the random intercepts b_i is present. Specifically, a correlation of $\rho = 0.9$ was introduced by adopting the sequential procedure described in Tutz and Oelker (2017).

Table 2: Results of the simulation study: Variable selection (scenario 1). Average true positive rates (TPR) and false positive rates (FPR) for the covariates in the six different settings. The table displays the results for all models that involve variable selection. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$ and error variance $\sigma^2_\varepsilon = 1$

	Model	Setting	1	2	3	4	5	6
			Base	$n = 40$	$n = 100$	$p = 100$	$\sigma^2_\varepsilon = 2$	$\rho = 0.9$
				$n_i = 25$	$n_i = 10$			
TPR	LMMP		1.000	1.000	1.000	1.000	1.000	1.000
	RE-EM		0.787	0.763	0.807	0.827	0.647	0.683
	LMMT		0.963	0.937	0.963	1.000	0.967	0.843
	LTSCB		1.000	1.000	1.000	1.000	1.000	1.000
	TTSC		0.843	0.797	0.837	0.827	0.777	0.640
FPR	LMMP		0.003	0.004	0.004	0.001	0.004	0.004
	RE-EM		0.000	0.000	0.000	0.000	0.000	0.000
	LMMT		0.000	0.001	0.000	0.003	0.007	0.004
	LTSCB		0.003	0.004	0.003	0.008	0.003	0.000
	TTSC		0.000	0.000	0.000	0.000	0.000	0.000

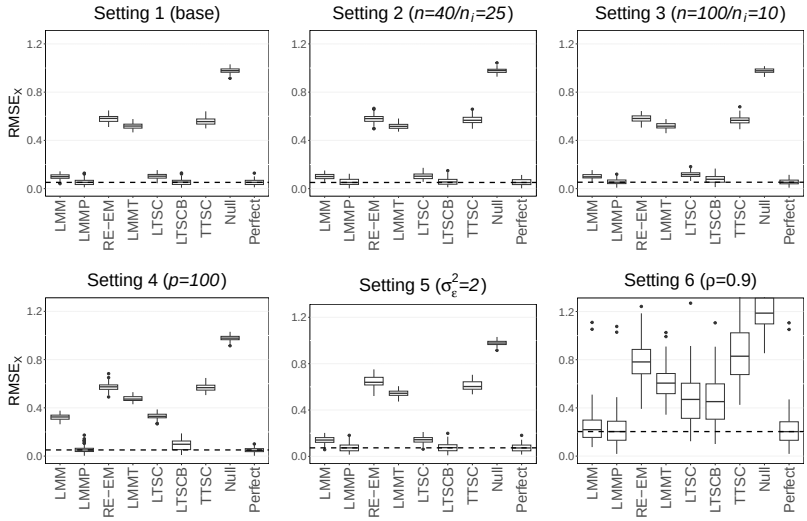


Fig. 4: Results of the simulation study: RMSE_X (scenario 1). Boxplots of the RMSE_X in the six different settings. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$ and error variance $\sigma^2_\varepsilon = 1$. The median values of the perfect model are marked by the dashed lines

The results in Table 2 indicate that all considered models were very efficient in detecting the informative covariates. The models with linear effects (LMMP and LTSCB) exhibit perfect TPRs equal to one and very low FPRs below 0.01 across all settings. Among the models with tree-structured effects LMMT yielded the highest TPRs. RE-EM and TTSC showed more conservative results, which may be due to the application of the 1SE rule. Changing the ratio of n to n_i (settings 2 and 3) and increasing the number of noise variables (setting 4) only had a minor impact on the variable selection rates. In settings 5 and 6, however, it is seen that the TPRs decreased for the tree-structured models indicating that variable selection becomes less reliable for these methods if the error variance is large or informative covariates are strongly correlated with the random intercepts. These patterns can also be observed in Figure 4, which depicts the results for RMSE_X. The models with linear effects (LMMP and LTSCB), which follow the true DGP, are shown to perform best and even similarly well to the perfect model. The corresponding models without variable selection (LMM and LTSC) performed only slightly worse throughout all settings and consistently better than the tree-structured models. In addition, all of the considered models yielded a much higher variance in RMSE_X if correlation between the informative covariates and the random intercepts was present (setting 6).

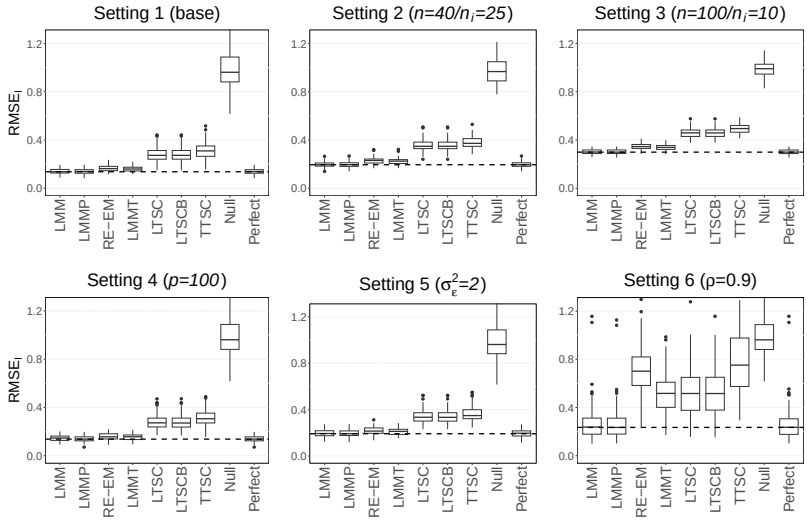


Fig. 5: Results of the simulation study: RMSE_I (scenario 1). Boxplots of the RMSE_I in the six different settings. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$ and error variance $\sigma_\varepsilon^2 = 1$. The median values of the perfect model are marked by the dashed lines

Figure 5 shows that the RMSE of the unit-specific effects were lowest for the models with random effects (LMM, LMMP, RE-EM, and LMMT) across all settings except for setting 6. Here, LMM and LMMP still performed best, but the tree-structured FEMs with linear covariate effects (LTSC and LTSCB) were beneficial compared to RE-EM and similar to LMMT. This is in line with the results obtained by Berger and Tutz (2018) for correlated covariates. Further, it underlines that if correlation between the covariates and the random intercepts is present, a correct specification of the covariate effects (that structurally aligns with the DPG) is highly important for an unbiased estimation of the unit-specific effects. The proposed TTSC model exhibited the highest RMSEs compared to all other competitors except for the Null model, which was to be expected as it aligns the least with the structure of the DPG.

5.4 Tree-structured DGP with random unit-specific intercepts

In the second scenario, the true DGP had the form

$$\begin{aligned} y_{ij} =& \gamma_1 I(x_{ij1} \leq 0 \wedge x_{ij2} \leq 0) + \gamma_2 I(x_{ij1} \leq 0 \wedge x_{ij2} > 0) + \\ & \gamma_3 I(x_{ij1} > 0 \wedge x_{ij7} = 0) + \gamma_4 I(x_{ij1} > 0 \wedge x_{ij7} = 1) + b_i + \varepsilon_{ij} \end{aligned}$$

Table 3: Results of the simulation study: Variable selection (scenario 2). Average true positive rates (TPR) and false positive rates (FPR) for the covariates in the six different settings. The table displays the results for all models that involve variable selection. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$ and error variance $\sigma^2_\varepsilon = 1$

	Model	Setting	1	2	3	4	5	6
			Base	$n = 40$	$n = 100$	$p = 100$	$\sigma^2_\varepsilon = 2$	$\rho = 0.9$
				$n_i = 25$	$n_i = 10$			
TPR	LMMP		0.987	0.983	0.980	0.967	0.930	0.997
	RE-EM		1.000	0.993	0.997	1.000	0.873	0.880
	LMMT		1.000	1.000	1.000	1.000	1.000	0.997
	LTSCB		0.950	0.940	0.943	0.987	0.833	0.930
	TTSC		0.993	0.990	1.000	0.993	0.933	0.927
FPR	LMMP		0.001	0.001	0.001	0.007	0.011	0.010
	RE-EM		0.000	0.000	0.000	0.000	0.000	0.000
	LMMT		0.001	0.001	0.002	0.001	0.019	0.004
	LTSCB		0.001	0.000	0.001	0.007	0.004	0.004
	TTSC		0.000	0.000	0.000	0.000	0.000	0.000

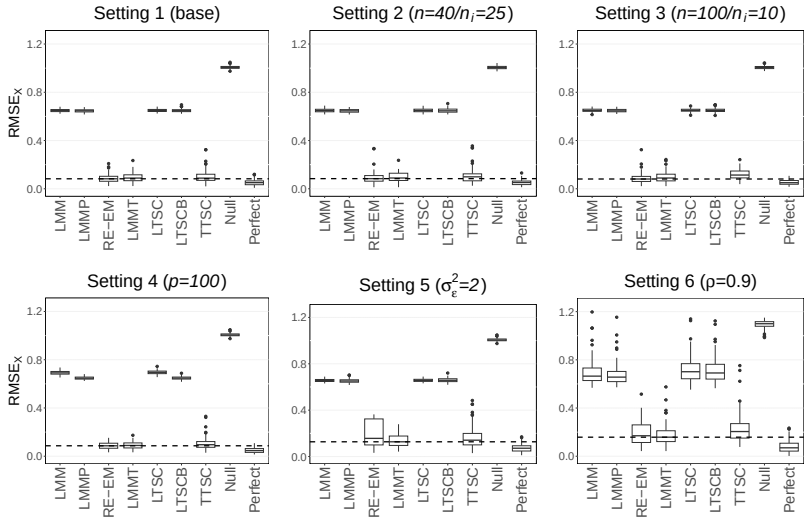


Fig. 6: Results of the simulation study: RMSE_X (scenario 2). Boxplots of the RMSE_X in the six different settings. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$ and error variance $\sigma^2_\varepsilon = 1$. The median values of the perfect model are marked by the dashed lines

with $\gamma_1 = -1.35$, $\gamma_2 = -0.45$, $\gamma_3 = 0.45$, and $\gamma_4 = 1.35$. Hence, again three out of ten (or one hundred) covariates were informative. Analogously to scenario 1, the unit-specific intercepts were standard normally distributed and for setting 6 a correlation of $\rho = 0.9$ between X_1 as well as X_2 and the random intercepts was introduced.

It is seen from Table 3 that the considered models exhibited high TPRs (> 0.8) and low FPRs (< 0.02) across all settings. The models with tree-structured effects of the covariates (RE-EM, LMMT, and TTSC) were superior to the other competitors with the highest TPRs and the lowest RMSE_X close to the perfect model (see Figure 6). As in scenario 1, RE-EM and TTSC selected no non-informative covariates (FPRs = 0). Compared to LMMT, RE-EM and TTSC yielded slightly lower TPRs and higher RMSE values in setting 5 with higher error variance. Furthermore, the results in Table 3 indicate the models with linear effects (LMMP and LTSC) were quite able to identify the informative covariates but the RMSE_X demonstrate that they were unable to capture the non-linear covariate effects. Overall, the performance suffered in

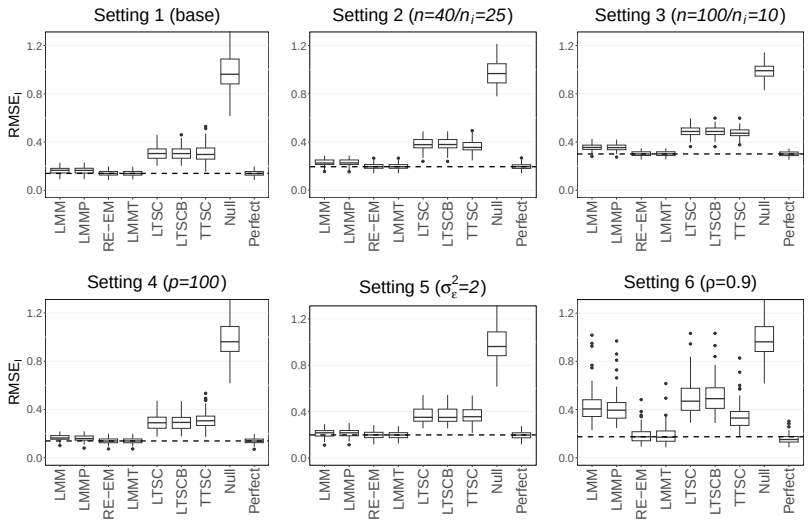


Fig. 7: Results of the simulation study: RMSE_I (scenario 2). Boxplots of the RMSE_I in the six different settings. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$ and error variance $\sigma_\varepsilon^2 = 1$. The median values of the perfect model are marked by the dashed lines

terms of variable selection and RMSE_X if correlation between the informative covariates and the random intercepts occurred (setting 6).

Similar to scenario 1, the results in Figure 7 show that the models with random effects (LMM, LMMP, RE-EM, and LMMT) were able to estimate the unit-specific effects more accurately than the tree-structured FEMs (LTSC, LTSCB, and TTSC) in settings 1 to 5. In setting 6, however, the models that capture tree-structured effects of the covariates showed superior performance compared to the linear models. In particular, TTSC yielded lower RMSE_I than LMM and LMMP. The results in scenario 2 again indicate that neither the ratio of n to n_i, the number of noise variables nor the variance of the error terms (varied in settings 2 to 5) changed the general pattern of the results.

5.5 Linear DGP with clustered unit-specific effects

The data in the third scenario were generated according to the DGP

$$y_{ij} = tr_0(i) + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_7 x_{ij7} + \varepsilon_{ij}$$

with $\beta_1 = 0.8$, $\beta_2 = 0.4$, and $\beta_7 = 0.8$. Hence, the linear predictor of the covariates coincided with scenario 1. In order to obtain clusters of units with the same intercepts, we drew a uniformly distributed random auxiliary variable $u_i \sim U(0, 1)$ for each of the n units. For settings 1 to 5, the effects of the units were then generated by

$$tr_0(i) = \beta_{01} I\left(u_i \in \left[0, \frac{1}{3}\right]\right) + \beta_{02} I\left(u_i \in \left(\frac{1}{3}, \frac{2}{3}\right]\right) + \beta_{03} I\left(u_i \in \left(\frac{2}{3}, 1\right]\right)$$

with $\beta_{01} = -1.25$, $\beta_{02} = 0$, and $\beta_{03} = 1.25$. That is, $C = 3$ clusters of units of roughly equal size were present in the data. In setting 6 we increased the number of clusters to $C = 6$ and applied the function

$$\begin{aligned} tr_0(i) =& \beta_{01} I\left(u_i \in \left[0, \frac{1}{6}\right]\right) + \beta_{02} I\left(u_i \in \left(\frac{1}{6}, \frac{2}{6}\right]\right) + \beta_{03} I\left(u_i \in \left(\frac{2}{6}, \frac{3}{6}\right]\right) + \\ & \beta_{04} I\left(u_i \in \left(\frac{3}{6}, \frac{4}{6}\right]\right) + \beta_{05} I\left(u_i \in \left(\frac{4}{6}, \frac{5}{6}\right]\right) + \beta_{06} I\left(u_i \in \left(\frac{5}{6}, 1\right]\right) \end{aligned}$$

with $\beta_{01} = -1.5$, $\beta_{02} = -0.9$, $\beta_{03} = -0.3$, $\beta_{04} = 0.3$, $\beta_{05} = 0.9$, and $\beta_{06} = 1.5$.

Table 4: Results of the simulation study: Variable selection (scenario 3). Average true positive rates (TPR) and false positive rates (FPR) for the covariates in the six different settings. The table displays the results for all models that involve variable selection. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$, $\sigma_\varepsilon^2 = 1$ and number of clusters $C = 3$

	Model	Setting	1	2	3	4	5	6
			Base	$n = 40$	$n = 100$	$p = 100$	$\sigma_\varepsilon^2 = 2$	$C = 6$
				$n_i = 25$	$n_i = 10$			
TPR	LMMP		1.000	1.000	1.000	1.000	1.000	1.000
	RE-EM		0.753	0.817	0.760	0.773	0.590	0.760
	LMMT		0.963	0.993	0.950	1.000	0.963	0.970
	LTSCB		1.000	1.000	1.000	1.000	1.000	1.000
	TTSC		0.783	0.803	0.803	0.807	0.710	0.817
FPR	LMMP		0.006	0.004	0.009	0.004	0.006	0.006
	RE-EM		0.000	0.000	0.000	0.000	0.000	0.000
	LMMT		0.000	0.000	0.000	0.002	0.007	0.000
	LTSCB		0.004	0.001	0.004	0.006	0.003	0.004
	TTSC		0.000	0.000	0.000	0.000	0.000	0.000

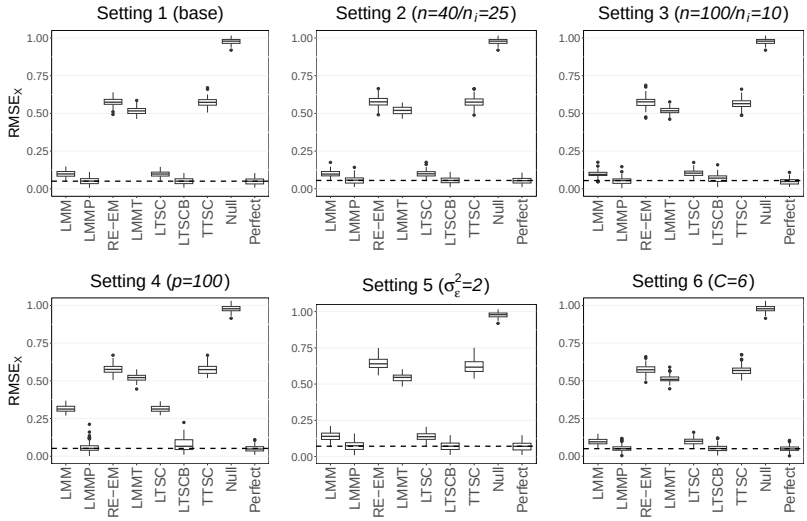


Fig. 8: Results of the simulation study: RMSE_X (scenario 3). Boxplots of the RMSE_X in the six different settings. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$, $\sigma_\varepsilon^2 = 1$ and number of clusters $C = 3$. The median values of the perfect model are marked by the dashed lines

The results shown in Table 4 and Figure 8 are fully in line with those of scenario 1 (see Table 2 and Figure 4), where the effects of the covariates also followed a linear DGP. Specifically, the linear models (LMMP and LTSCB) showed perfect TPRs with low FPRs across all settings. In addition, LMMT performed best among the tree-structured models and achieved TPRs of 0.95 or higher and decent RMSE_X in all settings. Overall, non of the considered models showed considerable differences in variable selection rates and RMSE_X compared to the base setting. Exceptions were the RMSE_X of the models without variable selection (LMM and LTSC) in setting 4 and RE-EM and TTSC in setting 5. RE-EM, in particular, strongly deteriorated in terms of TPR as the variance of the error terms increased.

Figure 9, which depicts the results of RMSE_I, shows that the tree-structured FEMs (LTSC, LTSCB, and TTSC) yielded more accurate estimates of the unit-specific effects than the models with random effects (LMM, LMMP, RE-EM, and LMMT) in settings 1, 4 and 5. As the number of units increased and the number of observations per unit decreased (settings 2 and 3), however, assuming random effects tended to be beneficial. In addition, a larger number of clusters also led

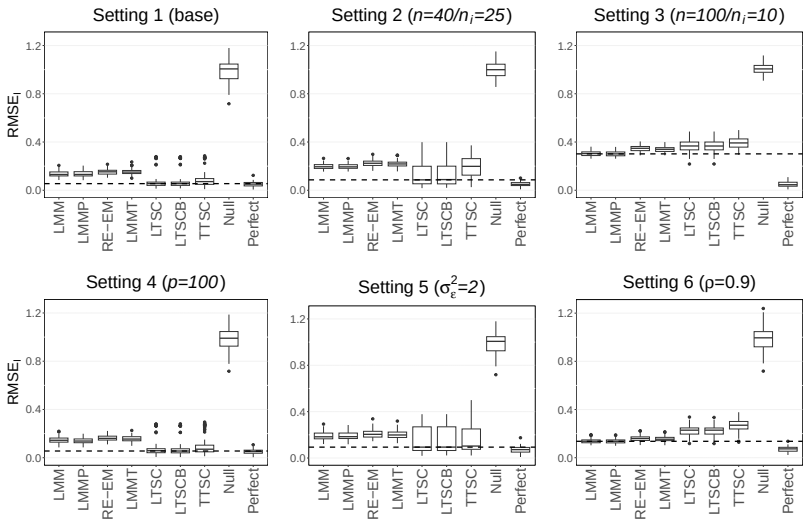


Fig. 9: Results of the simulation study: RMSE_I (scenario 3). Boxplots of the RMSE_I in the six different settings. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$, $\sigma^2_\varepsilon = 1$ and number of clusters $C = 3$. The median values of the perfect model are marked by the dashed lines

to a higher RMSE_I of the tree-structured FEMs compared to the models with random effects (setting 6). As the error variance was increased in setting 5, the tree-structured FEMs were still superior to the models with random effects, but showed much higher variability. Although the effects of the covariates followed a linear DGP, the TTSC model was not inferior in terms of RMSE_I compared to LTSC and LTSCB.

5.6 Tree-structured DGP with clustered unit-specific effects

In the fourth scenario, the true DGP had the form

$$\begin{aligned} y_{ij} =& tr_0(i) + \gamma_1 I(x_{ij1} \leq 0 \wedge x_{ij2} \leq 0) + \gamma_2 I(x_{ij1} \leq 0 \wedge x_{ij2} > 0) + \\ & \gamma_3 I(x_{ij1} > 0 \wedge x_{ij7} = 0) + \gamma_4 I(x_{ij1} > 0 \wedge x_{ij7} = 1) + \varepsilon_{ij}\,, \end{aligned}$$

where $\gamma_1 = -1.35$, $\gamma_2 = -0.45$, $\gamma_3 = 0.45$, $\gamma_4 = 1.35$. Hence, the function $tr(\cdot)$ of the covariates coincided with scenario 2. The function $tr_0(\cdot)$ of the intercepts was defined analogously to scenario 3.

The results in Table 5 and Figure 10 are comparable to the results in scenario 2 (see Table 3 and Figure 6), where the effects of the covariates also followed a tree-structured DGP. It is seen that all the considered models were able to identify the informative covariates quite well, but the RMSE_X values demonstrate that the linear models (LMM, LMMP, LTSC, and LTSCB) were unable to capture the non-linear covariate effects. The models with tree-structured effects of the covariates (RE-EM, LMMT and TTSC) yielded by far the lowest RMSE_X across all settings. While the differences between the six different settings appear small, RE-EM and TTSC had worse performance in terms of TPR and RMSE_X in setting 5 with increased variance of the error terms.

The RMSE_I shown in Figure 11 strongly coincide with the results observed in scenario 3 (see Figure 11), where we also considered a DGP with clustered unit-specific effects. Specifically, the results indicate that the tree-structured FEMs (LTSC, LTSCB, and TTSC) were beneficial if the number of units was low, the number of observations per unit was high, and there were only few clusters of units present in the data. Overall, the TTSC model was shown to perform well in these settings in terms of variable selection and MRSE on the covariate- as well as the unit-level.

To summarize the results of simulation scenarios 1 to 4, we made the following empirical key observations:

Table 5: Results of the simulation study: Variable selection (scenario 4). Average true positive rates (TPR) and false positive rates (FPR) for the covariates in the six different settings. The table displays the results for all models that involve variable selection. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$, $\sigma_\varepsilon^2 = 1$ and number of clusters $C = 3$

	Model	Setting	1	2	3	4	5	6
			Base	$n = 40$	$n = 100$	$p = 100$	$\sigma_\varepsilon^2 = 2$	$C = 6$
				$n_i = 25$	$n_i = 10$			
TPR		LMMP	0.993	0.993	0.990	0.957	0.923	0.973
		RE-EM	1.000	0.993	1.000	1.000	0.857	0.997
		LMMT	1.000	1.000	1.000	1.000	1.000	1.000
		LTSCB	0.953	0.923	0.913	1.000	0.867	0.930
		TTSC	0.983	0.983	0.993	0.987	0.877	1.000
FPR		LMMP	0.017	0.017	0.016	0.005	0.023	0.013
		RE-EM	0.000	0.000	0.000	0.000	0.000	0.000
		LMMT	0.020	0.019	0.014	0.002	0.019	0.026
		LTSCB	0.001	0.004	0.001	0.005	0.001	0.001
		TTSC	0.000	0.000	0.000	0.000	0.001	0.000

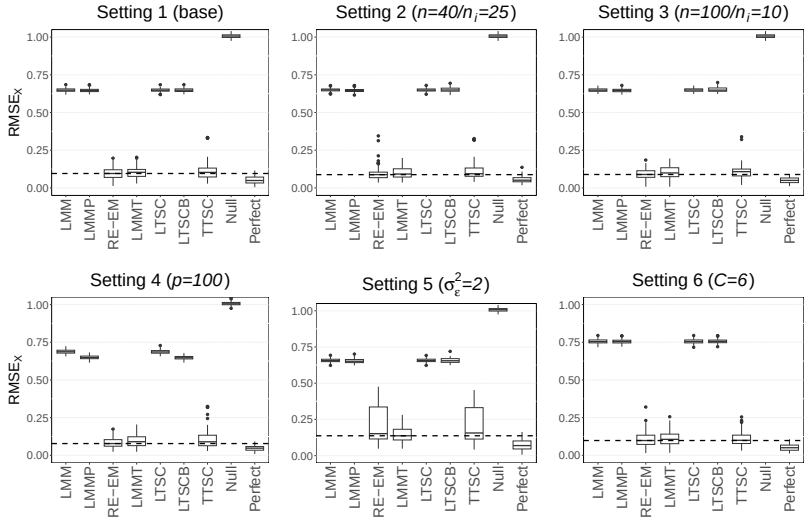


Fig. 10: Results of the simulation study: RMSE_X (scenario 4). Boxplots of the RMSE_X in the six different settings. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$, $\sigma_\varepsilon^2 = 1$ and number of clusters $C = 3$. The median values of the perfect model are marked by the dashed lines

1. All competitors showed high performance with regard to variable selection independent of the DPG.
2. Based on the RMSE, the tree-structured models were able to capture non-linear effects and interactions well.
3. Misspecification of unit-specific effects barely affects the goodness-of-fit of covariate effects.
4. Misspecification of covariate effects leads to biased unit-specific effect estimates, if correlation between the covariates and the unit-specific effects is present.
5. Tree-structured clustering is beneficial if the ratio of units to the observations per unit n/n_i is low and the number of clusters of units C is small.

6 Summary and discussion

In order to analyze QoL in the group of elderly Europeans using data of SHARE, we developed a tailored tree-structured approach. Established methods for modeling clustered data allow to combine tree-structured effects of individual-level covariates with random country-specific

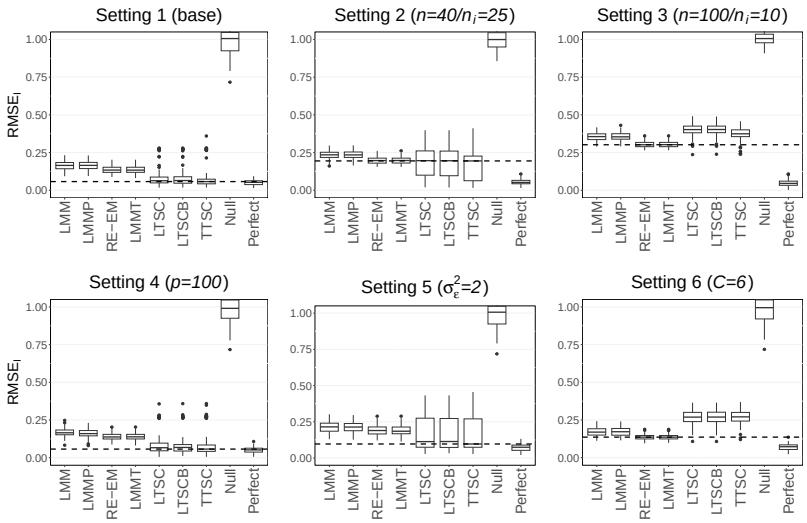


Fig. 11: Results of the simulation study: RMSE_I (scenario 4). Boxplots of the RMSE_I in the six different settings. Setting 1 serves as base setting with $n = 20$, $n_i = 50$, $p = 10$, $\sigma^2_\varepsilon = 1$ and number of clusters $C = 3$. The median values of the perfect model are marked by the dashed lines

effects (Hajjem et al., 2011; Sela and Simonoff, 2012; Fu and Simonoff, 2015; Fokkema et al., 2018), and to combine linear effects of covariates with clustered fixed country-specific effects (Berger and Tutz, 2018). A method that simultaneously includes tree-structured effects of covariates and clustered fixed country-specific effects has not been available so far. In the present paper, we fill this gap. Specifically, the proposed model extends upon tree-structured clustering, which is designed for sparse modeling of unit-specific intercepts (Berger and Tutz, 2018). We combine the tree representing unit-specific effects with a tree structure capturing effects of individual-level covariates. This second tree identifies subgroups of individuals that differ with regard to their outcome (the CASP score in SHARE). This accounts for non-linear effects and interactions between covariates, inherently performs variable selection and enables an accessible interpretation of parameters (see also the last paragraph in Section 2.2).

Our simulation study demonstrates that the proposed approach is competitive with alternative random effects-based approaches. Specifically, the proposed tree-structured FEM was shown to be advantageous if interactions between covariates were present and if there were only a few clusters of units with the same effect on the outcome. While random effects were also shown to work well in most settings and to be rather robust against violations of normality, confirming the findings in previous research (see, for example, Bell et al., 2018, and Schielzeth et al., 2020), the proposed tree-structured FEM yielded superior results in cases, where the number of units was low and the number of observations per unit was high. This is the case in SHARE with data from 28 countries and up to 3,000 observations per country. The analysis of SHARE presented in Section 4 highlights the applicability of the proposed method and confirms important findings about QoL in older adults.

While the focus in the simulation study and the application was on normally-distributed outcome variables, the proposed likelihood-based algorithm is generally applicable to differently scaled outcomes (including binary and discrete outcome variables). In addition, the predictor function is easily generalizable to an additive model of the form

$$\eta(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = tr_0(i) + tr(\boldsymbol{x}_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{z}_{ij}\,, \tag{12}$$

where $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijq})$ denotes an additional set of covariates with linear effects on the outcome. A random effects-based approach for modeling clustered data that also enables the combination of tree-structured and linear effects of the covariates was proposed by Gottard et al. (2023). Their model can be represented by an additive predictor with a linear term and three tree structures for unit-varying and unit-constant covariates as well as for interactions between unit-varying and unit-constant covariates.

In this paper, we reduced our considerations to clustered unit-specific intercepts. The proposed tree-structured algorithm, however, would also allow for clustered unit-specific effects of covariates (analogously to random slopes in random effects models). Referring to the set of covariates $\boldsymbol{z}_{ij}$, the model in Equation (5) can be extended to

$$\eta(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = tr_0(i) + \sum_{r=1}^{q} tr_r(i) z_{ijr} + tr(\boldsymbol{x}_{ij}) \,,$$

where the functions $tr_r(\cdot)$ are defined analogously to $tr_0(\cdot)$ as

$$tr_r(i) = \sum_{\ell=1}^{C_r} \beta_{r\ell} \, I(i \in N_{r\ell}) \,,$$

where $N_{r\ell}$ denotes the ℓ-th cluster of the units with respect to the effect of Z_r and $\beta_{r\ell}$ denotes the respective slope parameter. The fitting procedure described in Section 3 can easily be adapted to this case by considering the possible splits in all $q + 2$ trees in each step of the tree-building algorithm. In the first step, an order of the units $i \in \{1, \ldots, n\}$ needs to be determined with respect to each covariate, which is not necessarily the same.

If the focus is on predictive performance, the proposed model can be extended to an ensemble method. In this vein, Adler et al. (2011) investigated bootstrap-based strategies for dealing with longitudinal data in random forests, and Hajjem et al. (2012) proposed a random effects-based random forest approach for modeling clustered data.

Acknowledgements

This paper uses data from SHARE Wave 9 (DOI: 10.6103/SHARE.w9ca900) see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N° 211909, SHARE-LEAP: GA N° 227822, SHARE M4: GA N° 261982, DASISH: GA N° 283646) and Horizon 2020 (SHARE-DEV3: GA N° 676536, SHARE-COHESION: GA N° 870628, SERISS: GA N° 654221, SSHOC: GA N° 823782, SHARE-COVID19: GA N° 101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, VS 2020/0313, SHARE-EUCOV: GA N°101052589 and EUCOVII: GA N°101102412. Additional funding from the German Federal Ministry of Education and Research (01UW1301, 01UW1801, 01UW2202), the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, BSR12-04, R01_AG052527-02, R01_AG056329-02, R01_AG063944, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-eric.eu).

Statements and declarations

Competing interests: The authors have no competing interests to declare.

References

Adler, W., S. Potapov, and B. Lausen. 2011. Classification of repeated measurements data using tree-based ensemble methods. *Comput Stat* 26: 355–69. https://doi.org/10.1007/s00180-011-0249-1 .

Bell, A., M. Fairbrother, and J. Jones. 2018. Fixed and random effects models: making an informed choice. *Qual Quant* 53: 1051–74. https://doi.org/10.1007/s11135-018-0802-x .

Berger, M. 2023. *TSVC: Tree-Structured Modelling of Varying Coefficients*. R package version 1.5.3.

Berger, M. and G. Tutz. 2018. Tree-structured clustering in fixed effects models. *J Comput Graph Stat* 27: 380–392. https://doi.org/10.1080/10618600.2017.1371030 .

Berger, M., G. Tutz, and M. Schmid. 2019. Tree-structured modelling of varying coefficients. *Stat Comput* 29: 217–229. https://doi.org/10.1007/s11222-018-9804-8 .

Bergmann, M., M. Wagner, and A. Börsch-Supan. 2024. SHARE Wave 9 Methodology: From the SHARE Corona Survey 2 to the SHARE Main Wave 9 Interview. *Munich: SHARE-ERIC* .

Borrat-Besson, C., V. Ryser, and J. Goncalves. 2015. An evaluation of the CASP-12 scale used in the Survey of Health, Ageing and Retirement in Europe (SHARE) to measure Quality of Life among people aged 50+. *FORS Working Papers*. https://doi.org/10.24440/FWP-2015-00004 .

Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, and S. Zuber. 2013. Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *Int J Epidemiol*. https://doi.org/10.1093/ije/dyt088 .

Bowling, A. and P. Stenner. 2011. Which measure of quality of life performs best in older age? A comparison of the OPQOL, CASP-19 and WHOQOL-OLD. *J Epidmiol Community Health* 63: 273–280. https://doi.org/10.1136/jech.2009.087668 .

Breiman, L., J.H. Friedman, R.A. Olshen, and J.C. Stone. 1984. *Classification and Regression Trees*. Moneterey, CA Wadsworth: Taylor and Francis.

Diener, E. and C. Diener. 1995. The wealth of nations revisted: Income and quality of life. *Soc Indic Res* 36: 275–286. https://doi.org/10.1007/BF01078817 .

Doove, L.L., E. Dusseldorp, K.V. Deun, and I.V. Mechelen. 2014. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment – subgroup interactions. *Adv Data Anal Classif* 8: 403–425. https://doi.org/10.1007/s11634-013-0159-x .

European Commission and Eurostat. 2024. *Demography of Europe: 2024 edition*. Publications Office of the European Union.

Fisher, W.D. 1958. On grouping for maximum homogeneity. *J Am Stat Soc* 53: 789–798. https://doi.org/10.1080/01621459.1958.10501479 .

Fokkema, M., N. Smits, A. Zeilies, T. Hothorn, and H. Kelderman. 2018. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav Res Methods* 50: 2016–34. https://doi.org/10.3758/s13428-017-0971-x .

Frias-Goytia, G.L., C. Lojo-Seoane, S.C. Mallo, A. Nieto-Vieites, O. Juncos-Rabadàn, and A. Pereiro. 2024. A systematic review of quality of life (QoL) studies using the CASP scale in older adults. *Qual Life Res*: 1–13. https://doi.org/10.1007/s11136-024-03750-9 .

Fu, W. and J.S. Simonoff. 2015. Unbiased regression trees for longitudinal and clustered data. *Comput Stat Data Anal* 88: 53–74. https://doi.org/10.1016/j.csda.2015.02.004 .

Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Camebridge University Press.

Gottard, A., G. Vannucci, L. Grilli, and C. Rampichini. 2023. Mixed-effect models with trees. *Adv Data Anal Classif 17*(2): 431–461. https://doi.org/10.1007/s11634-022-00509-3 .

Grilli, L. and C. Rampichini. 2011. The role of sample cluster means in multi-level models: A view on endogeneity and measurment error issues. *Methodol: Eur J Res Methods Behav Soc Sci* 7: 121–33. https://doi.org/10.1027/1614-2241/a000030 .

Groll, A. and G. Tutz. 2014. Variable selection for generalized linear mixed models by l_1-penalized estimation. *Stat Comput* 24: 137–154. https://doi.org/10.1007/s11222-012-9359-z .

Grün, B. and F. Leisch. 2007. Fitting Finite Mixtures of Generalized Linear Regression in R. *Comput Stat Data Anal* 51: 5247–5252. https://doi.org/10.1016/j.csda.2006.08.014 .

Hajjem, A., F. Bellavance, and D. Larocque. 2011. Mixed effects regression trees for clustered data. *Stat Probab Lett* 81: 451–459. https://doi.org/10.1016/j.spl.2010.12.003 .

Hajjem, A., F. Bellavance, and D. Larocque. 2012. Mixed-effects random forest for clustered data. *J Comput Simul* 84: 1313–28. https://doi.org/10.1080/00949655.2012.741599 .

Heagerty, P. and B. Kurland. 2001. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 88: 973–985. https://doi.org/10.1093/biomet/88.4.973 .

Heinzl, F. and G. Tutz. 2013. Clustering in linear mixed models with approximate dirichlet process mixtures using em algorithm. *Stat Model* 13: 41–67. https://doi.org/10.1177/1471082X12471372 .

Heyworth, I.T.M., M.L. Hazell, M.F. Linehan, and T.L. Frank. 2009. How do common chronic conditions affect health-realted quality of life? *Br J Gen Pract* 59. https://doi.org/10.3399/bjgp09X453990 .

Hothorn, T., K. Hornik, and A. Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *J Comput Graph Stat* 15: 651–674. https://doi.org/10.1198/106186006X133933 .

Howel, D. 2012. Interpreting and evaluating the CASP-19 quality of life measure in older people. *Age and Ageing* 41: 612–617. https://doi.org/10.1093/ageing/afs023 .

Hyde, M., R. Wiggins, P. Higgs, and D. Blane. 2003. A measure of quality of life in early old age: The theory, development and properties of a needs satisfaction model (CASP-19). *Aging Ment Health* 7: 186–194. https://doi.org/10.1080/1360786031000101157 .

Kern, C., T. Klausch, and F. Kreuter. 2019. Tree-based machine learning methods for survey research. *J Eur Surv Res Assoc* 13: 73–93. https://doi.org/10.18148/srm/2019.v1i1.7395 .

Killingsworth, M.A. 2021. Experienced well-being rises with income, even above $75,000 per year. *Psychol Cogn Sci* 118. https://doi.org/10.1073/pnas.2016976118 .

Kim, G., G. Netuveli, D. Blane, A. Peasey, S. Malyutina, G. Simonova, R. Kubinova, A. Pajak, S. Croezen, M. M. Bobak, and H. Pikhart. 2015. Psychometric properties and confirmatory factor analysis of the CASP-19, a measure of quality of life in early old age: the hapiee study. *Aging Ment Health* 19: 595–609. https://doi.org/10.1080/13607863.2014.938605 .

Litière, S., A. Alonso, and G. Molenberghs. 2007. Type I and Type II Error Under Random Effect Misspecification in Generalized Linear Mixed Models. *Biom* 63: 1038–44. https://doi.org/10.1111/j.1541-0420.2007.00782.x .

Makovski, T.T., G.L. Coroller, P. Putrik, Y.H. Choi, M.P. Zeegers, S. Stranges, M.R. Castell, L. Huiart, and M. van den Akker. 2020. Role of clinical, functional and social factors in the association between multimorbidity and quality of life: Findings from the Survey of Health, Ageing and Retirement in Europe (SHARE). https://doi.org/10.1371/journal.pone.0240024 .

Makovski, T.T., S. Schmitz, M.P. Zeegers, S. Stranges, and M. van der Akker. 2019. Multimorbidity and quality of life: Systematic literature review and meta-analysis. *Ageing Res Rev* 53. https://doi.org/10.1016/j.arr.2019.04.005 .

Marengoni, A., S. Angleman, R. Melis, F. Mangialasche, A. Karp, A.G. aand B. Meinow, and L. Fratiglioni. 2011. Aging with multimorbidity: a systematic review of literature. *Ageing Res Rev* 10: 430–39. https://doi.org/10.1016/j.arr.2011.03.003 .

Molenberghs, G. and G. Verbeke. 2005. *Models for Discrete Longitudinal Data*. New York: Springer.

Niedzwiedz, C.L., S.V. Katikireddi, J.P. Pell, and R. Mitchell. 2014. Socioeconomic inequalities in thequality of life of older europeans in different welfare regimes. *Eur J Public Health* 24: 364–370. https://doi.org/10.1093/eurpub/cku017 .

Ripley, B.D. 1996. Pattern recognition and neural networks. *Cambridge: Cambridge University Press*. https://doi.org/10.1017/CBO9780511812651 .

Rodríguez-Bláquez, C., O. Ribeiro, A. Ayala, L. Teixeira, L. Arajúo, and M.J. Forjaz. 2020. Psychometric properties of the casp-12 scale in portugal: An analysis using share data. *Int J Environ Res Public Health* 17. https://doi.org/10.3390/ijerph17186610 .

Rothrock, N.E., R.D. Hays, K. Spritzer, S.E. Yount, W. Riley, and D. Cella. 2010. Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidmiol* 63: 1195–1204. https://doi.org/10.1016/j.jclinepi.2010.04.012 .

Schielzeth, H., N.J. Dingemanse, S. Nakagawa, D.F. Westneat, H. Allegue, C. Teplitsky, D. Réale, N. Dochtermann, L.Z. Garamszegi, and Y.F. Araya-Ajoy. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol Evol* 11: 1141–52. https://doi.org/10.1111/2041-210X.13434 .

Schwarz, G.E. 1978. Estimating the dimsension of a model. *Ann Stat* 6: 461–464. https://doi.org/10.1214/aos/1176344136 .

Sela, R.J. and J.S. Simonoff. 2012. RE-EM tree: a data mining approach for longitudinal and clustered data. *Mach Learn* 86: 169–207. https://doi.org/10.1007/s10994-011-5258-3 .

SHARE-ERIC. 2024. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 9. Release version: 9.0.0. *Data set*. https://doi.org/10.6103/SHARE.w9.900 .

Sim, J., B. Bartlam, and M. M. Bernard. 2011. The CASP-19 as a measure of quality of life in old age: evaluation of its use in a retirement community. *Qual Life Res* 20: 997–1004. https://doi.org/10.1007/s11136-010-9835-x .

Strobl, C., J. Malley, and G. Tutz. 2009. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychol Methods* 14: 323–48. https://doi.org/10.1037/a0016973 .

Tutz, G. and M.R. Oelker. 2017. Modelling clustered heterogeneity: Fixed effects, random effects and mixtures. *Int Stat Rev* 85: 204–227. https://doi.org/10.1111/insr.12161 .

Verbeke, G. and G. Molenberghs. 2000. *Linear Mixed Models for Longitudinal Data*. New York: Springer.

von dem Knesebeck, O., M. Wahrendorf, M. Hyde, and J. Siegrist. 2007. Socio-economic position and quality of lfe among older peaple in 10 European countries: results of the SHARE study. *Ageing Soc* 27: 269–84. https://doi.org/10.1017/S0144686X06005484 .

Zeileis, A., T. Hothorn, and K. Hornik. 2008. Model-based recursive pratitioning. *J Comput Graph Stat 17*(2): 492–514. https://doi.org/10.1198/106186008X319331 .

3.4 Publication 3: Confidence intervals for tree-structured varying coefficients

Spuck N, Schmid M, Monin M, Berger M. Confidence intervals for tree-structured varying coefficients. Computational Statistics and Data Analysis 2025; 207: 108142.
https://doi.org/10.1016/j.csda.2025.108142

Supplementary data can be found at:
https://doi.org/10.1016/j.csda.2025.108142

Computational Statistics and Data Analysis 207 (2025) 108142

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Confidence intervals for tree-structured varying coefficients

Nikolai Spuck [a,*,1], Matthias Schmid [a], Malte Monin [b,c], Moritz Berger [a]

[a] Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Venusberg-Campus 1, Bonn, 53127, Germany
[b] Department of Internal Medicine I, University Hospital Bonn, Venusberg-Campus 1, Bonn, 53127, Germany
[c] German Centre for Infection Research (DZIF), Partner-site Cologne-Bonn, Bonn, Germany

ARTICLE INFO

Keywords:
Varying coefficients
Tree-based modeling
Selective inference
Parametric bootstrap

ABSTRACT

The tree-structured varying coefficient (TSVC) model is a flexible regression approach that allows the effects of covariates to vary with the values of the effect modifiers. Relevant effect modifiers are identified inherently using recursive partitioning techniques. To quantify uncertainty in TSVC models, a procedure to construct confidence intervals of the estimated partition-specific coefficients is proposed. This task constitutes a selective inference problem as the coefficients of a TSVC model result from data-driven model building. To account for this issue, a parametric bootstrap approach, which is tailored to the complex structure of TSVC, is introduced. Finite sample properties, particularly coverage proportions, of the proposed confidence intervals are evaluated in a simulation study. For illustration, applications to data from COVID-19 patients and from patients suffering from acute odontogenic infection are considered. The proposed approach may also be adapted for constructing confidence intervals for other tree-based methods.

1. Introduction

Regression analysis is a powerful tool to quantify the association between an outcome variable and a set of covariates and to make inferences about the true parameter values. Classical statistical theory provides asymptotic properties for estimators of regression coefficients that allow performing hypothesis tests and constructing confidence intervals (CIs) based on their estimates. It is, however, a well known result that classical inference is invalid if the analysis involves a data-driven model selection procedure, that is, if the structure of a model's predictor function is determined in a data-driven way (e.g. by forward or stepwise variable selection; Taylor and Tibshirani, 2015), as statistical uncertainty induced by the model selection is neglected.

In this article we deal with this issue in the context of CIs for regression models with varying coefficients. This class of models first introduced by Hastie and Tibshirani (1993) generalizes the class of linear regression models, as they allow that coefficients of covariates change with the value of other variables, the so-called *effect modifiers*. Fan and Zhang (2008) and Park et al. (2015) gave comprehensive reviews on varying coefficient models and discussed several fitting approaches. In the past years varying coefficient models have been considered extensively, which has led to many extensions of the classical approach, see, for example, Wang and Hastie (2014), Buergin and Ritschard (2015, 2017), Lee et al. (2020), and Zhou and Hooker (2022). A large part of this methodology makes the basic assumption that the effect of each covariate is modified by a known set of potential effect modifiers, which is specified before model fitting. Then one determines the way the coefficients are modified. This prerequisite is relaxed by tree-structured varying coefficient (TSVC) models proposed by Berger et al. (2019), which select the effect modifiers from the whole set of available covariates

* Corresponding author.
E-mail address: spuck@imbie.uni-bonn.de (N. Spuck).
[1] Venusberg-Campus 1, Building 11, 53127 Bonn, Germany.

https://doi.org/10.1016/j.csda.2025.108142
Received 19 June 2024; Received in revised form 23 January 2025; Accepted 23 January 2025

Available online 27 January 2025

in a data-driven way and allow that the varying coefficients are caused by the interaction of several effect modifiers. The TSVC model applies recursive partitioning to identify relevant effect modifiers, which yields a separate tree T_j for the linear effect of each covariate X_j, where each leaf node contains a partition-specific coefficient. Building a tree means to find a partition of the covariate space using binary splits, which induces a piecewise constant predictor function. In TSVC models, each split refers to a coefficient and is determined by an effect modifier and a corresponding splitting rule: For a metrically or ordinally scaled effect modifier X_k, the splitting into two partitions has the form '$N_1 = N \cap \{X_k \leq c\}$ and $N_2 = N \cap \{X_k > c\}$', with regard to split point c, where N denotes the parent node.

While TSVC models may apply a concept of statistical significance to decide whether to split or not (described in detail in Berger et al., 2019), they so far do not allow for inference on the varying coefficients within the nodes of the fitted trees. Here, we fill this gap proposing a bootstrap-based approach.

When constructing CIs for tree-structured varying coefficients one needs to take into account that the tree structures and therefore the partitions that differ in terms of their effect (defined by the effect modifiers and splitting rules) were estimated from the data (here denoted by D). This is further aggravated by the fact that the trees suffer from high variability, that is, only a small change in the data D can lead to a rather different model structure (here denoted by $\mathcal{M}$), which may even strongly deviate from the true data generating process (DGP). Neglecting this uncertainty induced by the data-driven tree building procedure yields CIs that are likely to be too short. More specifically, it must be taken into account that the coefficients of interest $\beta_{jm}^{\mathcal{M}}$ (the linear effect of covariate j in node m of tree T_j) arise out of model structure $\mathcal{M}$, that is, that the model selection event $\widehat{\mathcal{M}} = \mathcal{M}$ occurred.

The TSVC model $\mathcal{M}$ considered in this paper is defined by a set of partitions $\mathcal{M} = \{\{N_{jm}, m = 1, ..., M_j\}, j = 1, ..., p\}$. Therefore, a $100(1-\alpha)\%$ CI of $\beta_{jm}^{\mathcal{M}}$ is supposed to satisfy

$$\mathbb{P}\left(\beta_{jm}^{\mathcal{M}} \in CI(\beta_{jm}^{\mathcal{M}}) \mid \widehat{\mathcal{M}} = \mathcal{M}\right) \geq 1 - \alpha\,, \tag{1}$$

which constitutes a so-called *selective inference* or *post selection inference* problem (Berk et al., 2013; Fithian et al., 2014; Lee et al., 2016). This issue has been intensively studied in linear regression models (Zhang et al., 2022). Selective CIs were proposed, among others, by Tibshirani et al. (2016) for sequential variable selection procedures, by Ruegamer and Greven (2018) after likelihood-based model selection in linear models, and by Suzumura et al. (2017) for linear models including higher-order interactions. More recently, Zhao et al. (2022) proposed a selective inference approach for LASSO-based varying coefficient models, Ruegamer et al. (2022) investigated selective inference for additive and linear mixed effect models, and Zrnic and Jordan (2023) address the selective inference problem by building on the framework of algorithmic stability. With respect to tree-structured models, Gottard et al. (2023) proposed a simple splitting of the data into training data for fitting the model and test data for conducting inference, which, however, comes at the price that only a subset of the observations is used for tree building and for subsequent inference. Loh et al. (2019) proposed bootstrap-calibrated CIs within the GUIDE regression tree framework and Neufeld et al. (2022) proposed "Tree-Values", a selective inference framework for regression trees.

In line with the principle of selective inference, we propose a parametric bootstrap approach to construct $100(1-\alpha)\%$ percentile CIs $CI_P(\beta_{jm}^{\mathcal{M}})$ for a TSVC model $\mathcal{M}$ with varying coefficients $\beta_{jm}^{\mathcal{M}}, j = 1, \ldots, p, m = 1, \ldots, M_j$, that satisfy Equation (1).

To the best of our knowledge parametric bootstrap has so far not been used in the context of selective inference. In general, our method is not restricted to the specific implementation of TSVC but could also be adapted to construct confidence intervals for parameters of other tree-based models (see also the discussion in Section 6).

The remainder of this paper is organized as follows: In Section 2, the class of TSVC models and the fitting procedure are described. Section 3 outlines our parametric bootstrap approach for constructing percentile CIs of the varying coefficients. To assess coverage proportions of the proposed CIs, we conducted a simulation study presented in Section 4. In the simulation study, we also contrast our proposal to bootstrap-calibrated CIs. In Section 5, we show the results of two applications fitting TSVC models to data of patients suffering from COVID-19 and acute odontogenic infection, respectively. Finally, our findings are summarized and discussed in Section 6.

2. Tree-structured varying coefficients

Let Y be a outcome variable of interest and $\boldsymbol{X} = (X_1, \ldots, X_p)$ be explanatory variables that are ordinally or metrically scaled, or dummy-coded representations of nominal variables. In generalized regression models it is assumed that the outcome Y given the values of covariates $\boldsymbol{X}$ follows a distribution from the exponential family. The expected outcome is related to the covariate vector in the form $\mathbb{E}(Y|\boldsymbol{X}) = g^{-1}(\eta(\boldsymbol{X}))$, where $g(\cdot)$ denotes a suitable link function and $\eta(\cdot)$ denotes the predictor function. Most frequently it is assumed that the predictor function is characterized by a linear combination of the covariates. In the more general varying coefficient model introduced in the seminal work by Hastie and Tibshirani (1993), the predictor function is given by

$$\eta(\boldsymbol{X}, \boldsymbol{Z}) = \beta_0 + \beta_1(Z_1)X_1 + \cdots + \beta_p(Z_p)X_p\,, \tag{2}$$

where $Z_1, \ldots, Z_p$ denote (additional) random variables that serve as *effect modifiers* and change the linear effects of $X_1, \ldots, X_p$ through unspecified functional forms.

The model with predictor (2) requires the effect modifiers to be specified beforehand. In practice, however, it is often unclear which variable modifies the effect of another variable. In addition, each varying coefficient may not be determined by just one variable and the effect may be driven by an interaction between several effect modifiers. To address these issues, Berger et al. (2019)

proposed the tree-structured varying coefficient (TSVC) model, which applies a recursive partitioning method to detect relevant effect modifiers. Since the effect modifiers are inherently selected by the tree building algorithm, only the set of covariates $X_1, \dots, X_p$ is considered for modeling. If effect modifiers are present, they are from this set and modify coefficients of covariates from this set. The predictor function of a TSVC model $\mathcal{M}$ is given by

$$\eta^{\mathcal{M}}(\boldsymbol{X}) = \beta_0^{\mathcal{M}} + \beta_1^{\mathcal{M}}(\boldsymbol{X}_{[-1]})X_1 + \cdots + \beta_p^{\mathcal{M}}(\boldsymbol{X}_{[-p]})X_p\,, \tag{3}$$

where $\boldsymbol{X}_{[-j]}$ denotes the set of covariates $X_1, \dots, X_p$ excluding X_j. By definition, the effect of each covariate can be modified by each other covariate except itself. The functions $\beta_j^{\mathcal{M}}(\cdot)$ are each determined by a tree structure. This means that each function $\beta_j^{\mathcal{M}}(\cdot)$ sequentially partitions the observations into disjoint subsets N_{jm}, $m = 1, ..., M_j$, based on the values of the selected effect modifiers and assigns a different regression coefficient for X_j to each partition N_{jm}. These functions can be written as

$$\beta_j^{\mathcal{M}}(\boldsymbol{X}_{[-j]}) = \sum_{m=1}^{M_j} \beta_{jm}^{\mathcal{M}}\, I(\boldsymbol{X}_{[-j]} \in N_{jm}), \tag{4}$$

where $I(\cdot)$ denotes the indicator function. Hence, the structure of a TSVC model $\mathcal{M}$ is characterized by the set of partitions $\mathcal{M} = \{\{N_{jm}, m = 1, \dots, M_j\}, j = 1, \dots, p\}$. Each coefficient is derived from binary splits successively partitioning the observations of one parental node into two child nodes (cf. Hastie et al., 2009). We start from a model with non-varying linear effects, only. Then, the first split yields model $\mathcal{M}^{[1]}$ with predictor

$$\begin{aligned}\eta^{\mathcal{M}^{[1]}}(\boldsymbol{X}_i) = \beta_0^{\mathcal{M}^{[1]}} + \beta_1^{\mathcal{M}^{[1]}} X_1 + \cdots + \left(\beta_{j1}^{\mathcal{M}^{[1]}} I(X_k \le c_k) + \beta_{j2}^{\mathcal{M}^{[1]}} I(X_k > c_k)\right) X_j \\ + \cdots + \beta_p^{\mathcal{M}^{[1]}} X_p\,,\end{aligned}$$

where c_k is the split point in effect modifier X_k selected by the algorithm regarding the effect of X_j, $\beta_{j1}^{\mathcal{M}^{[1]}}$ is the linear effect of X_j in partition $\{X_k \le c_k\}$ and $\beta_{j2}^{\mathcal{M}^{[1]}}$ is the linear effect of X_j in partition $\{X_k > c_k\}$ adjusted for the other effects in $\mathcal{M}^{[1]}$. Hence, after the first step, the varying coefficient of X_j is determined by $\beta_j^{\mathcal{M}^{[1]}}(X_k) = \beta_{j1}^{\mathcal{M}^{[1]}} I(X_k \le c_k) + \beta_{j2}^{\mathcal{M}^{[1]}} I(X_k > c_k)$. In the next step, either a different coefficient is selected for splitting or the same coefficient is further modified. If the coefficient of variable X_ℓ is split in X_r at split point c_r this yields the predictor

$$\begin{aligned}\eta^{\mathcal{M}^{[2]}}(\boldsymbol{X}) = \beta_0^{\mathcal{M}^{[2]}} + \beta_1^{\mathcal{M}^{[2]}} X_1 + \cdots + \left(\beta_{j1}^{\mathcal{M}^{[2]}} I(X_k \le c_k) + \beta_{j2}^{\mathcal{M}^{[2]}} I(X_k > c_k)\right) X_j \\ + \left(\beta_{\ell 1}^{\mathcal{M}^{[2]}} I(X_r \le c_r) + \beta_{\ell 2}^{\mathcal{M}^{[2]}} I(X_r > c_r)\right) X_\ell + \cdots + \beta_p^{\mathcal{M}^{[2]}} X_p\,,\end{aligned}$$

where $\beta_{\ell 1}^{\mathcal{M}^{[2]}}$ denotes the effect of X_ℓ in $\{X_r \le c_r\}$ and $\beta_{\ell 2}^{\mathcal{M}^{[2]}}$ denotes the effect of X_ℓ in $\{X_r > c_r\}$. That is, the varying coefficient of X_ℓ has the form $\beta_\ell^{\mathcal{M}^{[2]}}(X_r) = \beta_{\ell 1}^{\mathcal{M}^{[2]}} I(X_r \le c_r) + \beta_{\ell 2}^{\mathcal{M}^{[2]}} I(X_r > c_r)$. Further splits are performed analogously until a predefined stopping criterion is met (see below for details). In each step a so far non-varying effect turns into a varying coefficient or an already selected varying coefficient is split once more.

Sketch of the fitting procedure

In each step of the tree building algorithm, the best splitting rule from among all possible combinations of covariate X_j, respective candidate effect modifier $X_k, k \neq j$, and split point is selected, starting from a linear predictor without varying coefficients. For this, all candidate models with one additional split are evaluated and the best-performing one that yields the smallest deviance is selected. In generalized regression models the deviance is a quite natural measure of the model fit. This criterion is also equivalent to minimizing the entropy, which has been used as a splitting criterion already in the early days of tree construction (Breiman et al., 1984). Note that, in contrast to common trees, in each step of the algorithm all the observations are used to derive new estimates of the model parameters. This ensures that one obtains valid estimates of the different components together with the splitting rule.

To determine the optimal number of splits and hence the size of the trees, Berger et al. (2019) proposed an early stopping strategy based on permutation tests. This offers an approximate solution to control the global type I error rate (that is, the proportion of falsely identified covariates with varying coefficients). In this paper, we use an alternative post-pruning strategy, where a large model is grown first and is then pruned to an adequate size to avoid overfitting. Running the stepwise TSVC algorithm (with a sufficiently large number of splits) yields a sequence of nested models that are assessed with regard to their goodness of fit using a likelihood-based criterion. Subsequently, the best-performing model is selected and fitted on the whole data with the corresponding number of splits. This post-pruning strategy is less computationally intensive than early stopping using permutation tests and was previously shown to perform similarly well (Spuck et al., 2023). Note, however, that the post-pruning strategy, unlike the permutation test approach, does neither control the global type I error rate nor the probability of falsely identifying a variable as effect modifier and may therefore lead to less parsimonious models. Here, we select the optimal number of splits by minimizing the Bayesian information criterion (BIC; Schwarz, 1978). The BIC of a TSVC model $\mathcal{M}^{[s]}$ is given by

$$BIC\left(\mathcal{M}^{[s]}\right) = -2\ln\left(L_{\mathcal{M}^{[s]}}\right) + s\log(n)\,, \tag{5}$$

where $L_{\mathcal{M}^{[s]}}$ denotes the maximized value of the likelihood function of model $\mathcal{M}^{[s]}$, $s = \sum_{j=1}^{p}(M_j - 1)$ is the total number of performed splits, and n denotes the number of observations. In order to prevent the nodes from becoming too small yielding unstable coefficient estimates, a minimal bucket size constraint can be applied. The minimal bucket size constraint requires a minimum number of observations in every terminal node.

3. Selective confidence intervals

To demonstrate the selective inference problem formalized in Equation (1) that comes with the TSVC model, we consider a simple example with two metrically scaled covariates $\boldsymbol{X} = (X_1, X_2)$. Let Y be the outcome variable that follows a normal distribution, $Y \sim N\left(\mu(\boldsymbol{X}), \sigma^2\right)$, with conditional expectation

$$\mu(\boldsymbol{X}) = \beta_0 + \beta_{11}\, I(X_2 \leq c_2)\, X_1 + \beta_{12}\, I(X_2 > c_2)\, X_1 + \beta_2\, X_2\,. \tag{6}$$

Hence, there exists a varying linear effect of X_1 on Y modified by X_2. Specifically, two regions with different linear effects of X_1 are present: region $R_{11} = \{X_2 \leq c_2\}$ with linear effect β_{11} and region $R_{12} = \{X_2 > c_2\}$ with linear effect β_{12}. The linear effect of X_2 remains the same across the whole covariate space, that is, $R_2 = \{X_1 \in \mathbb{R}\}$.

Assume a TSVC model $\mathcal{M}_1$ is fitted to a sample $\mathcal{D}_1 = \{(y_i^{(1)}, \boldsymbol{x}_i^{(1)} = (x_{i1}^{(1)}, x_{i2}^{(1)})), i = 1, \ldots, n\}$, where the values of the outcome variable $y_i^{(1)}$ were drawn from the normal distribution with conditional expectation (6). If the structure of the fitted model coincides with the true DGP (one split of X_2 in the linear effect of X_1 with split point c_2), that is, $\widehat{\mathcal{M}} = \mathcal{M}_1 = \{\{R_{11}, R_{12}\}, \{R_2\}\}$, one obtains estimates $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, $\hat{\beta}_2$ of the underlying DGP parameters. However, since the structure of model $\mathcal{M}_1$ is determined based on the sample $\mathcal{D}_1$, the partitions detected by the algorithm are likely to deviate from the true underlying regions, which means $\mathcal{M}_1 \neq \{\{R_{11}, R_{22}\}, \{R_2\}\}$. For instance, assume the predictor function of model $\mathcal{M}_1$ is given by

$$\eta^{\mathcal{M}_1}(\boldsymbol{X}) = \hat{\beta}_0^{\mathcal{M}_1} + \hat{\beta}_{11}^{\mathcal{M}_1} X_1 + \hat{\beta}_{21}^{\mathcal{M}_1} I(X_1 \leq c_1)\, X_2 + \hat{\beta}_{22}^{\mathcal{M}_1} I(X_1 > c_1)\, X_2\,.$$

In this model, the interaction between X_1 and X_2 was found correctly but its form is flawed. Specifically, we have that $\mathcal{M}_1 = \{\{N_{11}\}, \{N_{21}, N_{22}\}\}$ with $N_{11} = \{X_2 \in \mathbb{R}\}, N_{21} = \{X_1 \leq c_1\}$ and $N_{22} = \{X_1 > c_1\}$. As the structure of the model is misspecified, none of the estimated coefficients $\hat{\boldsymbol{\beta}}^{\mathcal{M}_1} = (\hat{\beta}_{11}^{\mathcal{M}_1}, \hat{\beta}_{21}^{\mathcal{M}_1}, \hat{\beta}_{22}^{\mathcal{M}_1})$ matches the true underlying effects $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \beta_2)$ from the DGP (6). Instead, as outlined analogously for data-driven variable selection in linear models by Berk et al. (2013), the coefficient vector $\hat{\boldsymbol{\beta}}^{\mathcal{M}_1}$ actually is an estimate of the solution to the optimization problem

$$\begin{aligned}\boldsymbol{\beta}^{\mathcal{M}_1} = \mathbb{E}\left(\hat{\boldsymbol{\beta}}^{\mathcal{M}_1}\right) &= \underset{\boldsymbol{b}^{\mathcal{M}_1}}{\arg\max}\ \mathbb{E}\left(L_{\mathcal{M}_1}\left(\boldsymbol{b}^{\mathcal{M}_1} \mid Y, \boldsymbol{x}^{(1)}\right)\right) \\ &= \underset{\boldsymbol{b}^{\mathcal{M}_1}}{\arg\max}\ L_{\mathcal{M}_1}\left(\boldsymbol{b}^{\mathcal{M}_1} \mid \mu(\boldsymbol{x}^{(1)}), \boldsymbol{x}^{(1)}\right),\end{aligned} \tag{7}$$

where $L_{\mathcal{M}_1}(\cdot)$ denotes the likelihood function of model $\mathcal{M}_1$. Note that, following Berk et al. (2013), the expectation is evaluated only with regard to the outcome variable Y, and the values of the covariates are treated as fixed at the observed values $\boldsymbol{x}^{(1)} = (\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)})$. By definition, $\boldsymbol{\beta}^{\mathcal{M}_1}$ maximizes the likelihood of the model with the fixed structure $\mathcal{M}_1$ using the expected values of the outcome variable given the observed values of the covariates (i.e. assuming that the DGP is known). In addition, Equation (7) implies that the target effects $\beta_{jm}^{\mathcal{M}}$ in different models are generally different. For example, applying the TSVC fitting procedure to another sample $\mathcal{D}_2 = \{(y_i^{(2)}, \boldsymbol{x}_i^{(2)} = (x_{i1}^{(2)}, x_{i2}^{(2)})), i = 1, ..., n\}$ yields a new model $\mathcal{M}_2$ with a potentially different structure. Assume the predictor function of model $\mathcal{M}_2$ is given by

$$\begin{aligned}\eta^{\mathcal{M}_2}(\boldsymbol{X}) = \hat{\beta}_0^{\mathcal{M}_2} &+ \hat{\beta}_{11}^{\mathcal{M}_2}\, I(X_2 \leq c_2)\, X_1 + \hat{\beta}_{12}^{\mathcal{M}_2}\, I(X_2 \leq c_2)\, X_1 \\ &+ \hat{\beta}_{21}^{\mathcal{M}_2}\, I(X_1 \leq c_1) X_2 + \hat{\beta}_{22}^{\mathcal{M}_2}\, I(X_1 > c_1)\, X_2\,.\end{aligned}$$

Even though the estimated coefficients $\hat{\beta}_{21}^{\mathcal{M}_1}$ and $\hat{\beta}_{21}^{\mathcal{M}_2}$ of X_2 refer to same partition N_{21}, they do not estimate the same effect, because model structures $\mathcal{M}_1$ and $\mathcal{M}_2$ do not match and were adjusted for differently structured effects of X_1.

More generally, given a data set $\mathcal{D} = \{(y_i, \boldsymbol{x}_i = (x_{i1}, ..., x_{ip}), i = 1, ..., n\}$ and a TSVC model with selected structure $\mathcal{M} = \{\{N_{jm}, m = 1, ..., M_j\}, j = 1, ..., p\}$ fitted to the data, the corresponding coefficient vector $\hat{\boldsymbol{\beta}}^{\mathcal{M}}$ estimates the *best approximating varying linear coefficients* given the selected model structure $\mathcal{M}$ defined by

$$\boldsymbol{\beta}^{\mathcal{M}} = \mathbb{E}\left(\hat{\boldsymbol{\beta}}^{\mathcal{M}}\right) = \underset{\boldsymbol{b}^{\mathcal{M}}}{\arg\max}\ L_{\mathcal{M}}\left(\boldsymbol{b}^{\mathcal{M}} \mid \mu(\boldsymbol{x}), \boldsymbol{x}\right). \tag{8}$$

Our objective is to construct selective CIs for the coefficients $\beta_{jm}^{\mathcal{M}}$ as defined in Equation (8) that satisfy Equation (1). For this purpose, the distribution of $\boldsymbol{y}$ conditional on the model selection event $\widehat{\mathcal{M}} = \mathcal{M}$ needs to be considered. Note that if the structure of the selected model coincides with the structure of the DGP, the best approximating varying linear coefficients match the true effects of the DGP. In linear regression models with LASSO penalization, Lee et al. (2016) found that if the selection event $\widehat{\mathcal{M}} = \mathcal{M}$ can be characterized by a set of inequalities $\boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b}$, where $\boldsymbol{A}$ and $\boldsymbol{b}$ must not depend on $\boldsymbol{y}$, $\widehat{\mathcal{M}} = \mathcal{M}$ constitutes a *linear selection event* and exact statistical

inference of the coefficients conditional on the selection event can be performed. Specification of the selection event for TSVC models, however, would require a vast number of inequalities. The main reason is that the TSVC algorithm involves the fitting of several trees, which is considerably more complex than fitting of a single tree or a predictor function with interactions of predefined order (scenarios investigated by Neufeld et al., 2022 and Suzumura et al., 2017, respectively). Specifically, in the first iteration of the TSVC algorithm, the event of selecting one splitting rule is characterized by $p(p-1)n$ inequalities, assuming p continuous covariates with n possible split points each. Each inequality specifies that the maximal likelihood value of a model that results from one of the possible splitting rules (i.e. a combination of covariate X_j, effect modifier $X_k, k \neq j$, and split point) is lower than the maximal likelihood value of the model with the selected splitting rule. Overall, $\mathcal{O}(np^2S)$ inequalities are required to describe the selection of one particular sequence of nested TSVC models $\mathcal{M}^{[s]}, s = 1, \ldots, S$, and an optimal model $\mathcal{M}$ out of it. Since there are cases where the same model structure $\mathcal{M}$ can arise from a number of different sequences of nested models (e.g. when the same splits are performed in a different order), the conditioning set that characterizes the selection event $\widehat{\mathcal{M}} = \mathcal{M}$ is a union of sets defined by these inequalities.

To tackle this complex mechanism, we propose a parametric bootstrap approach tailored to the selective inference problem at hand (described in detail in Sections 3.1 and 3.2). Basically, given a TSVC model $\mathcal{M}$ fitted to data $\mathcal{D}$ with coefficients $\beta_{jm}^{\mathcal{M}}$, we (i) generate samples $\mathcal{D}_b$ by drawing new values of the outcome variable Y while keeping the covariate values fixed using a parametric bootstrap scheme and (ii) calculate estimates of $\beta_{jm}^{\mathcal{M}}$ from the TSVC model $\mathcal{M}_b$ fitted to $\mathcal{D}_b$ in order to construct percentile CIs based on these estimates.

3.1. Calculating bootstrap effect estimates for given $\mathcal{M}$

To construct a CI for the coefficient $\beta_{jm}^{\mathcal{M}}$ from a given model with structure $\mathcal{M}$ satisfying Equation (1), we compute estimates for $\beta_{jm}^{\mathcal{M}}$ from a set of bootstrap samples $\mathcal{D}_b$, $b = 1, \ldots, B$ (see Section 3.2 for details on the applied bootstrap sampling scheme). A naive approach, which simply enforces the structure of the original model $\mathcal{M}$ on each sample, would imply that the model structure is predefined and neglect the uncertainty induced by the data-driven tree building procedure. To account for this uncertainty, we first apply the TSVC fitting procedure to the samples $\mathcal{D}_b$, resulting in B different models $\mathcal{M}_b$ of potentially different form. For each b the predictor function of model $\mathcal{M}_b$ is given by

$$\eta^{\mathcal{M}_b}(\boldsymbol{X}) = \hat{\beta}_0^{\mathcal{M}_b} + \hat{\beta}_1^{\mathcal{M}_b}(\boldsymbol{X}_{[-1]})X_1 + \ldots + \hat{\beta}_p^{\mathcal{M}_b}(\boldsymbol{X}_{[-p]})X_p$$

with

$$\hat{\beta}_j^{\mathcal{M}_b}(\boldsymbol{X}_{[-j]}) = \sum_{m=1}^{M_j^{(b)}} \hat{\beta}_{jm}^{\mathcal{M}_b} \, I(\boldsymbol{X}_{[-j]} \in N_{jm}^{(b)}).$$

Secondly, we determine an estimate of the coefficient of interest $\beta_{jm}^{\mathcal{M}}$ from the original model based on bootstrap model $\mathcal{M}_b$ by averaging the node-specific effect estimates $\hat{\beta}_{jm}^{\mathcal{M}_b}$ with regard to the corresponding partition N_{jm} from the original model yielding

$$\bar{\beta}_{jm}^{(b)} = \frac{1}{|N_{jm}|} \sum_{i:\boldsymbol{x}_i \in N_{jm}} \hat{\beta}_j^{\mathcal{M}_b}(\boldsymbol{x}_{i[-j]}). \tag{9}$$

This means, for each covariate X_j each observation is assigned to one of the subsets N_{jm} that was identified by the original model $\mathcal{M}$, and subsequently the average value of the function $\hat{\beta}_j^{\mathcal{M}_b}(\cdot)$ from model $\mathcal{M}_b$ across the observations in N_{jm} is calculated. Therefore, Equation (9) defines an estimate of $\beta_{jm}^{\mathcal{M}}$ for bootstrap sample $\mathcal{D}_b$ that accounts for the uncertainty induced by the data-driven tree building.

Finally, a $100(1-\alpha)\%$ percentile CI for $\beta_{jm}^{\mathcal{M}}$ is constructed as

$$CI_P\left(\beta_{jm}^{\mathcal{M}}\right) = \left[\bar{\beta}_{jm}^{\alpha/2}; \bar{\beta}_{jm}^{1-\alpha/2}\right], \tag{10}$$

where $\bar{\beta}_{jm}^{q}$ denotes the $100q$-th percentile of the set of bootstrap estimates $\bar{\beta}_{jm}^{(1)}, \ldots, \bar{\beta}_{jm}^{(B)}$.

3.2. Parametric bootstrap procedure

By the definition in Equation (9) one can determine bootstrap estimates of the coefficients of interest $\beta_{jm}^{\mathcal{M}}$. Yet, calculating these estimates does in itself not condition on the model selection event $\widehat{\mathcal{M}} = \mathcal{M}$. To take this into account, we mimic the conditioning by applying a parametric bootstrap scheme, which is based on the original model $\mathcal{M}$.

For each observation i, the value $y_i^{(b)}$ in bootstrap sample $\mathcal{D}_b$ is drawn from the conditional distribution of $Y \mid \boldsymbol{X} = \boldsymbol{x}_i$ given the fitted TSVC model $\mathcal{M}$, following the parametric bootstrap sampling scheme described in Efron and Tibshirani (1993). That is, the new outcome values $y_i^{(b)}$ are generated from a distribution with expectation

$$\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}_i) = g^{-1}\left(\eta^{\mathcal{M}}(\boldsymbol{x}_i)\right), \tag{11}$$

where $\eta^{\mathcal{M}}(\cdot)$ is the predictor function of the original TSVC model fitted to $\mathcal{D}$. Of note, we keep the values of the covariates $\boldsymbol{X}$ fixed at the observed values $\boldsymbol{x}_i$, which is in line with the definition of the best approximating varying linear coefficients in Equation (8). Drawing from the conditional distribution also ensures that a bootstrap estimate can be calculated in each sample even if a node in the original model is very small (unlike with a nonparametric bootstrap scheme where it is not ensured that a bootstrap sample contains observations from every node).

If a Gaussian TSVC model is considered, $g(\cdot)$ is the identity link and the new outcome values of the bootstrap samples are drawn using

$$y_i^{(b)} \sim N\left(\eta^{\mathcal{M}}(\boldsymbol{x}_i), \hat{\sigma}_\varepsilon^2\right),$$

where $N(\cdot,\cdot)$ denotes the normal distribution and $\hat{\sigma}_\varepsilon^2$ is the residual variance of model $\mathcal{M}$. For a binary logistic TSVC model, the new outcome values are generated as

$$y_i^{(b)} \sim \text{Bin}\left(1, \text{logit}\left(\eta^{\mathcal{M}}(\boldsymbol{x}_i)\right)\right),$$

where $\text{Bin}(\cdot,\cdot)$ denotes the binomial distribution. In general, this approach allows to generate B bootstrap samples $\mathcal{D}_b = \{(y_i^{(b)}, \boldsymbol{x}_i), i = 1,\dots n\}$ for any kind of generalized TSVC model.

To summarize, given a TSVC model $\mathcal{M}$ fitted to data $\mathcal{D}$, we propose to perform the following steps to construct $100(1-\alpha)\%$ CIs for the coefficients $\beta_{jm}^{\mathcal{M}}$:

1. **Bootstrap sampling:** Generate B bootstrap samples $\mathcal{D}_b = \{(y_i^{(b)}, \boldsymbol{x}_i), i = 1,\dots n\}$ by sampling new outcome values from the conditional distribution of the outcome variable with expectation (11).
2. **Model fitting:** Apply the TSVC fitting procedure described in Section 2 to each sample $\mathcal{D}_b$ in order to obtain a model $\mathcal{M}_b$ for each sample.
3. **Calculating bootstrap estimates:** Determine the estimates $\tilde{\beta}_{jm}^{(b)}$ as defined in Equation (9) for each model $\mathcal{M}_b$.
4. **Percentile intervals:** Construct percentile CIs by computing the $100(\alpha/2)$-th and $100(1-\alpha/2)$-th percentiles of the set of bootstrap estimates $\tilde{\beta}_{jm}^{(1)}, \dots, \tilde{\beta}_{jm}^{(B)}$ as described in Equation (10).

The proposed CI method is available in the R add-on package TSVC version 1.7.2 by Berger (2025).

4. Simulation study

To assess coverage proportions of the proposed parametric bootstrap percentile CIs, we considered different simulation scenarios. The aims of the simulation study were (i) to evaluate how the coverage of the proposed CIs is affected by the structure of the DGP, (ii) to investigate the effect of sample size and noise in the DGP on the coverage proportions, and (iii) to compare coverage proportions of the proposed CIs to those based on alternative types of CIs (e.g. simple asymptotic normal distribution-based Wald intervals). The scenarios were based on a linear DGP without varying effects (scenario 1), a tree-structured varying effect DGP (scenario 2), a tree-structured varying effect DGP where effect modifiers were prespecified before model fitting (scenario 3), and a tree-structured varying effect DGP with additional noise variables (scenario 4). Further details on the DGPs will be given in the following subsections. In each replication of the three scenarios, a TSVC model was fitted to the data, where the maximal number of splits was set to $S = 5$ and the BIC was used to determine the optimal number of splits. For the resulting coefficients of interest, 90% as well as 95% CIs were constructed using the following methods:

(i) simple asymptotic normal distribution-based Wald type CIs (*Wald*),
(ii) our proposed parametric bootstrap percentile CIs (*Parametric percentile*), and
(iii) Wald type CIs, where an adjusted α-level is determined via bootstrap calibration to account for the uncertainty induced by the tree building (*Bootstrap calibration*; Loh et al., 2019).

The $100(1-\alpha)\%$ Wald type CIs are calculated as

$$CI_W\left(\beta_{jm}^{\mathcal{M}}\right) = \left[\hat{\beta}_{jm}^{\mathcal{M}} + z_{\alpha/2}\text{SE}(\hat{\beta}_{jm}^{\mathcal{M}}); \hat{\beta}_{jm}^{\mathcal{M}} + z_{1-\alpha/2}\text{SE}(\hat{\beta}_{jm}^{\mathcal{M}})\right], \tag{12}$$

where z_q denotes the $100q$-th percentile of the standard normal distribution and $\text{SE}(\hat{\beta}_{jm}^{\mathcal{M}})$ denotes the standard error of the coefficient estimate (not including the uncertainty induced by the model selection). The bootstrap calibration method introduced by Loh et al. (2019) is designed to construct CIs for coefficients of regression models that were fitted on subgroups identified by a GUIDE regression tree. We applied a version of their algorithm adapted to TSVC models. For a detailed description of the approach, see the Supplementary Material. The proposed parametric percentile CIs and the bootstrap-calibrated CIs were constructed based on $B = 1000$ bootstrap samples, respectively.

In the three simulation scenarios we considered a continuous outcome variable and either two or three covariates with a potential effect on the outcome, where $X_1, X_2 \sim N(0,1)$ and $X_3 \sim \text{Bin}(1, 0.5)$. We considered sample sizes of $n \in \{200, 500, 1000\}$ and DGPs with normally distributed error terms with standard deviations of $\sigma_\varepsilon \in \{1, 2\}$.

Table 1
Average number of splits based on 5000 replications when fitting a TSVC model for a linear DGP (scenario 1). The first row shows the average total number of splits per replication in the TSVC models. Below the average number of splits per replication is reported separately for each combination of covariate and effect modifier. The column *DGP* contains the true number of splits.

Covariate	Effect modifier	DGP	n	200		500		1000	
			σ_ε	1	2	1	2	1	2
—	—	**0**		**0.63**	**0.63**	**0.34**	**0.34**	**0.22**	**0.22**
X_1	X_2	0		0.32	0.32	0.18	0.18	0.11	0.11
X_2	X_1	0		0.31	0.31	0.17	0.17	0.11	0.11

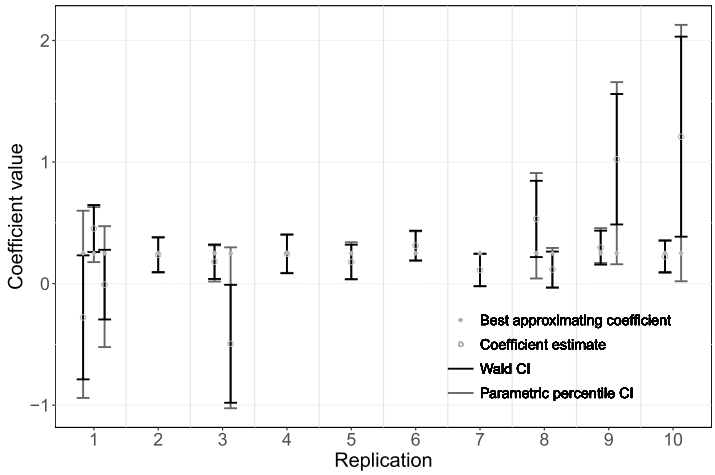


Fig. 1. Fitted tree-structured varying coefficients and 95% CIs for a linear DGP (scenario 1). The figure shows the estimated effects $\hat{\beta}_{jm}^{\mathcal{M}_r}$ of the original model fitted to the data D_r, the best approximating linear coefficients $\beta_{jm}^{\mathcal{M}_r}$ as defined in Equation (8), corresponding 95% Wald type and parametric percentile CIs for the varying linear coefficients of X_1 for 10 exemplary replications. The number of coefficients of X_1 in the true linear DGP is 1. The underlying data were drawn from a linear DGP with $n = 200$ and $\sigma_\varepsilon = 1$.

Coverage proportions were calculated based on $R = 5000$ replications. For the varying linear coefficients of a covariate X_j, the average coverage proportion was calculated as

$$C_j = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{M_j^r}\sum_{m=1}^{M_j^r} I\left(\beta_{jm}^{\mathcal{M}_r} \in CI(\beta_{jm}^{\mathcal{M}_r})\right),$$

where $\mathcal{M}_r$ denotes the TSVC model fitted in the r-th replication and M_j^r denotes the number of coefficients of X_j in model $\mathcal{M}_r$. The average coverage proportion across all covariates is then given by

$$C_{\text{av}} = \frac{1}{p}\sum_{j=1}^{p} C_j .$$

4.1. Linear DGP

In the first scenario, X_1 and X_2 were included as covariates and potential effect modifiers in the TSVC fitting procedure. The DGP of the first scenario was given by

$$y_i = 0.25\, x_{i1} + \varepsilon_i,\ i = 1, \ldots, n, \tag{13}$$

which means that X_1 has a simple non-varying linear effect and X_2 is non informative. The proportions of variance explained by X_1 were approximately 0.06 ($\sigma_\varepsilon = 1$) and 0.02 ($\sigma_\varepsilon = 2$).

Table 1 shows that the number of splits performed by the TSVC fitting procedure increased with lower sample size but appeared to be unaffected by the standard deviation of the error term. Hence, the structure of the model tended to be closer to the DGP with larger sample sizes. Of note, the average number of falsely performed splits was nearly equal for the informative covariate X_1 and the non-informative covariate X_2.

Exemplary results of the fitted CIs are depicted in Fig. 1. The figure illustrates that the proposed parametric percentile CIs coincided with the simple Wald type CIs in cases where no splits were performed (i.e. if there was only one coefficient of X_1, as, for example,

Table 2
Coverage proportions of 95% CIs based on 5000 replications for a linear DGP (scenario 1). For each CI method the first row shows the coverage proportion averaged across all coefficients (C_{av}). Below coverage proportions are reported separately for each covariate (C_j). Coverage proportions for 90% CIs are given in Table S1 in the Supplementary Material. Coverage proportions of 95% and 90% CIs using early stopping by permutation tests are reported in Tables S14 and S15 in the Supplementary Material.

CI method	Covariate	n	200		500		1000	
		σ_ε	1	2	1	2	1	2
Wald	—		**.833**	**.833**	**.875**	**.875**	**.900**	**.900**
	X_1		.832	.832	.874	.874	.903	.903
	X_2		.835	.835	.877	.877	.896	.896
Parametric percentile	—		**.951**	**.951**	**.949**	**.949**	**.949**	**.949**
	X_1		.950	.950	.949	.949	.951	.951
	X_2		.952	.952	.949	.949	.947	.947
Bootstrap calibration	—		**.883**	**.883**	**.902**	**.902**	**.916**	**.916**
	X_1		.880	.880	.900	.900	.919	.919
	X_2		.885	.885	.903	.903	.914	.914

in replication 2). In cases with more than one coefficient (for example in replication 1), the parametric percentile CIs were wider and were therefore more likely to cover the best approximating coefficient. An evaluation of the average widths of the CIs and corresponding standard deviations is presented in Tables S2 and S3 in the Supplementary Material. From Fig. 1 it is also seen, that unlike the Wald type CIs, the parametric percentile CIs were not necessarily symmetric around the coefficient estimate. Note that, due to the linear DGP without varying coefficients, the best approximating coefficient equals the true effect from the DGP independent of the selected TSVC model structure.

From Table 2 it is seen that the proposed parametric percentile CIs yielded coverage proportions very close to the nominal level across all settings (with varying n and σ_ε). The coverage proportions of the Wald type CIs were far too low but increased with higher sample size. This is likely due to the lower number of splits performed for larger samples (see Table 1) and the fact that the Wald type CIs are valid and yield the desired coverage if no splits are performed (i.e. if the coefficients are simply non-varying). The bootstrap-calibrated CIs showed improved coverage proportions compared to the Wald type CIs but performed considerably worse than our proposed CIs (coverage proportions < 0.92). Analogous results were observed for 90% CIs (see Supplementary Material Table S1). With early stopping by permutation tests (see Supplementary Material Tables S14 and S15) the coverage proportions of the parametric percentile CIs remained unaffected, while the bootstrap-calibrated CIs were too conservative and showed to be highly sensitive to the pruning method.

Of note, even with this simple underlying linear DGP without varying effects, neglecting the fact that constructing CIs for TSVCs is a selective inference problem (e.g. by applying a naive Wald type CI) may yield highly anti-conservative results with low coverage.

4.2. Varying effect DGP

The data in the second scenario was generated by

$$y_i = 0.5\, I(x_{i2} \leq 0.5 \wedge x_{i3} = 1)\, x_{i1} - I(x_{i2} > 0.5)\, x_{i1} + \varepsilon_i\,,\ i = 1, \ldots, n\,. \tag{14}$$

Here, a varying effect of X_1 that is determined by a tree structure with three terminal nodes was present. The covariates X_2 and X_3 did not have a linear effect but served only as effect modifiers for X_1. The proportions of variance explained by the covariate and the effect modifiers were 0.26 ($\sigma_\varepsilon = 1$) and 0.08 ($\sigma_\varepsilon = 2$).

Table 3 shows that the average number of splits performed by TSVC decreased with sample size and noise. It is also seen that the proportion of splits performed for the linear effect of X_1 increased with sample size, which suggests that the structure of the fitted models aligned more closely with the structure of the DGP if n was large. Furthermore, the table shows that X_2 was more likely to be selected as an effect modifier than X_3, which reflects the tendency of the tree building towards the selection of continuous over binary splitting variables.

Fig. 2 as well as Supplementary Material Tables S5 and S6 illustrate that the parametric bootstrap approach yielded wider CIs than the normal distribution-based Wald type approach (which resulted in much better coverage proportions, see Table 4). Note that, in this scenario, the best approximating coefficients differed between replications depending on the structure of the fitted TSVC model. In most of the depicted replications the three regions with coefficients $\beta_{11} = 0$, $\beta_{12} = 0.5$, and $\beta_{13} = 1$ appear to be identified quite well. There were, however, also cases where too many (replications 4 and 8) or too few splits in the coefficient of X_1 were performed (replication 7).

Table 4 shows that the coverage proportions of the proposed parametric percentile CIs were slightly conservative but approached the nominal level of 95% for larger sample size and lower noise. Coverage proportions for the coefficients of X_1 exceeded the nominal level the most whereas the CIs for the coefficients of X_3 (eighth row in Table 4) were close to the nominal level across all settings. This may be due to the fact that X_3 was a binary variable and therefore allowed only one split point when selected as effect modifier. While the bootstrap calibration approach outperformed the normal distribution-based Wald type approach, both resulted in insufficiently low coverage across all settings (< 0.87 and < 0.92 on average). Coverage of the Wald type CIs increased with larger sample size, but

Table 3
Average number of splits based on 5000 replications when fitting a TSVC model for a varying effect DGP (scenario 2). The first row shows the average total number of splits per replication in the TSVC models. Below the average number of splits per replication is reported separately for each combination of covariate and effect modifier. The column *DGP* contains the true number of splits.

Covariate	Effect modifier	DGP	n	200		500		1000	
			σ_ϵ	1	2	1	2	1	2
—	—	**2**		**3.55**	**3.05**	**3.04**	**2.73**	**2.87**	**2.66**
X_1	—	**2**		**2.47**	**1.43**	**2.42**	**2.00**	**2.47**	**2.22**
	X_2	1		1.50	1.03	1.41	1.14	1.43	1.23
	X_3	1		0.97	0.40	1.02	0.86	1.03	1.00
X_2	—	**0**		**0.31**	**0.42**	**0.18**	**0.20**	**0.12**	**0.13**
	X_1	0		0.29	0.39	0.18	0.19	0.12	0.13
	X_3	0		0.02	0.03	0.01	0.01	0.00	0.01
X_3	—	**0**		**0.77**	**1.20**	**0.43**	**0.53**	**0.28**	**0.30**
	X_1	0		0.42	0.72	0.23	0.31	0.15	0.16
	X_2	0		0.35	0.48	0.21	0.22	0.13	0.14

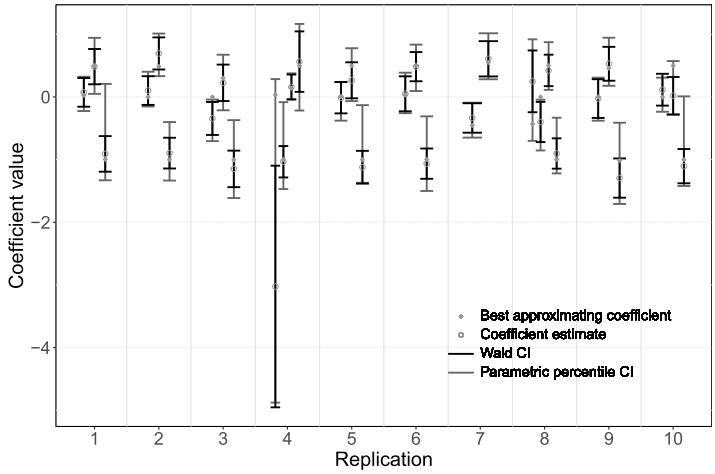


Fig. 2. Fitted tree-structured varying coefficients and 95% CIs for a varying effect DGP (scenario 2). The figure shows the estimated effects $\hat{\beta}_{jm}^{\mathcal{M}_r}$ of the original model fitted to the data $\mathcal{D}_r$, the best approximating linear coefficients $\beta_{jm}^{\mathcal{M}_r}$ as defined in Equation (8), corresponding 95% Wald type and parametric percentile CIs for the varying linear coefficients of X_1 for 10 exemplary replications. The number of coefficients of X_1 in the true varying effect DGP is 3. The underlying data were drawn from a linear DGP with $n = 200$ and $\sigma_\epsilon = 1$.

no differences with regard to n and σ_ϵ were apparent for bootstrap calibration. The coverage proportions of the 90% CIs exhibited an overall similar pattern (see Supplementary Material Table S2).

4.3. Varying effect DGP with known effect modifiers

In the third scenario, the varying effect DGP from Equation (14) was applied again. However, compared to scenario 2 it is now assumed that the effect modifiers were known before model fitting, i.e. it was specified that only X_2 and X_3 are considered as potential effect modifiers in the TSVC fitting procedure. This type of scenario is common in applications, where prior knowledge is available or certain shapes of interactions between variables are scientifically not meaningful (see for example the application to real-world data in Section 5.2). Note that, while knowing the effect modifiers simplified the model selection problem, the TSVC algorithm still needed to detect which coefficients are modified by which effect modifier and the corresponding splitting rule.

The average number of splits shown in Table 5 was lower than in the second scenario and closer to the true number of 2 splits across all settings. As in the previous scenarios, fewer splits were performed if the sample size was large and the level of noise was high. In addition, it is seen that nearly none of the splits were performed in the coefficient of X_2, where the only available splitting option was X_3.

Table 6 shows that the proposed parametric percentile CIs yielded coverage proportions close to the nominal level for the coefficients of X_1 and X_3 (sixth and eighth row) and rather conservative coverage proportions for the coefficients of X_2 (seventh row) across all settings. The fact that X_1 was no longer considered as a potential effect modifier for X_2 and X_3 led to substantially improved coverage proportions for the coefficients of X_1 compared to the results observed in the previous scenario (see Table 4). The Wald type and bootstrap-calibrated CIs again tended to yield insufficient coverage proportions but they were much closer to nominal level

Table 4
Coverage proportions of 95% CIs based on 5000 replications for a varying effect DGP (scenario 2). For each CI method the first row shows the coverage proportion averaged across all coefficients (C_{av}). Below coverage proportions are reported separately for each covariate (C_j). Coverage proportions for 90% CIs are given in Table S2 in the Supplementary Material.

CI method	Covariate	n	200		500		1000	
		σ_ε	1	2	1	2	1	2
Wald	—		**.795**	**.795**	**.843**	**.852**	**.865**	**.867**
	X_1		.824	.807	.862	.873	.879	.883
	X_2		.797	.801	.849	.857	.868	.872
	X_3		.764	.777	.817	.825	.849	.846
Parametric percentile	—		**.968**	**.971**	**.966**	**.970**	**.963**	**.965**
	X_1		.981	.984	.979	.985	.972	.977
	X_2		.971	.975	.971	.972	.971	.972
	X_3		.951	.955	.948	.952	.947	.948
Bootstrap calibration	—		**.901**	**.911**	**.901**	**.914**	**.911**	**.908**
	X_1		.922	.916	.914	.934	.918	.917
	X_2		.896	.913	.899	.909	.911	.809
	X_3		.886	.903	.890	.899	.903	.899

Table 5
Average number of splits based on 5000 replications when fitting a TSVC model for a varying effect DGP with known effect modifiers (scenario 3). The first row shows the average total number of splits per replication in the TSVC models. Below the average number of splits per replication is reported separately for each combination of covariate and effect modifier. The column *DGP* contains the true number of splits.

Covariate	Effect modifier	DGP	n	200		500		1000	
			σ_ε	1	2	1	2	1	2
—	**—**	**2**		**2.95**	**2.39**	**2.64**	**2.45**	**2.60**	**2.37**
X_1	**—**	**2**		**2.64**	**2.06**	**2.47**	**2.29**	**2.49**	**2.26**
	X_2	1		1.61	1.35	1.44	1.29	1.45	1.24
	X_3	1		1.03	0.71	1.03	0.99	1.04	1.01
X_2	X_3	0		0.01	0.01	0.01	0.01	0.00	0.00
X_3	X_2	0		0.30	0.32	0.16	0.16	0.11	0.11

Table 6
Coverage proportions of 95% CIs based on 5000 replications for a varying effect DGP with known effect modifiers (scenario 3). For each CI method the first row shows the coverage proportion averaged across all coefficients (C_{av}). Below coverage proportions are reported separately for each covariate (C_j). Coverage proportions for 90% CIs are given in Table S3 in the Supplementary Material.

CI method	Covariate	n	200		500		1000	
		σ_ε	1	2	1	2	1	2
Wald	—		**.873**	**.870**	**.903**	**.903**	**.915**	**.915**
	X_1		.862	.856	.895	.897	.906	.910
	X_2		.899	.899	.920	.919	.934	.931
	X_3		.858	.855	.893	.892	.906	.903
Parametric percentile	—		**.952**	**.957**	**.955**	**.957**	**.956**	**.956**
	X_1		.948	.961	.951	.955	.949	.951
	X_2		.970	.970	.971	.972	.972	.972
	X_3		.939	.940	.944	.943	.946	.946
Bootstrap calibration	—		**.923**	**.928**	**.928**	**.930**	**.934**	**.931**
	X_1		.925	.925	.928	.928	.929	.928
	X_2		.929	.935	.927	.936	.946	.944
	X_3		.914	.924	.918	.927	.925	.922

of 95% compared to scenario 2. The bootstrap-calibrated CIs achieved proportions around 0.93 if the sample size was large and the level of noise was low. Similar results were observed for the 90% CIs (see Supplementary Material Table S3). Average widths of the CIs are given in Supplementary Material Tables S8 and S9. Wald type CIs were shorter than the parametric percentile CIs throughout all settings. The bootstrap-calibrated and the parametric percentile CIs were similarly wide, indicating that the symmetric shape of the bootstrap-calibrated CIs may be the cause of insufficient coverage proportions.

Table 7
Average number of splits based on 5000 replications when fitting a TSVC model for a varying effect DGP with additional noise variables (scenario 4). The first row shows the average total number of splits per replication in the TSVC models. Below the average number of splits per replication is reported separately for each combination of covariate and effect modifier. The column *DGP* contains the true number of splits.

Covariate	Effect modifier	DGP	n	200		500		1000	
			σ_ε	1	2	1	2	1	2
—	—	**2**		**4.39**	**4.09**	**3.70**	**3.54**	**3.11**	**3.11**
X_1	—	**2**		**2.48**	**1.84**	**2.44**	**2.25**	**2.25**	**2.26**
	X_2	1		1.46	1.19	1.42	1.27	1.24	1.25
	X_3	1		1.02	0.64	1.02	0.98	1.01	1.01
X_2	X_3	0		0.01	0.01	0.01	0.01	0.00	0.00
X_3	X_2	0		0.21	0.25	0.14	0.14	0.11	0.10
Noise	—	**0**		**0.24**	**0.29**	**0.16**	**0.16**	**0.11**	**0.11**
	X_2	0		0.21	0.25	0.14	0.15	0.10	0.10
	X_3	0		0.03	0.03	0.02	0.02	0.01	0.01

Table 8
Coverage proportions of 95% CIs based on 5000 replications for a varying effect DGP with additional noise variables (scenario 4). For each CI method the first row shows the coverage proportion averaged across all coefficients (C_{av}). Below coverage proportions are reported separately for each covariate (C_j). Coverage proportions for 90% CIs are given in Table S3 in the Supplementary Material.

CI method	Covariate	n	200		500		1000	
		σ_ε	1	2	1	2	1	2
Wald	—		**.832**	**.819**	**.877**	**.871**	**.877**	**.894**
	X_1		.863	.848	.890	.887	.890	.914
	X_2		.768	.760	.821	.821	.821	.842
	X_3		.820	.804	.872	.866	.872	.888
	Noise		.840	.826	.883	.877	.884	.899
Parametric percentile	—		**.944**	**.941**	**.951**	**.951**	**.951**	**.952**
	X_1		.957	.955	.949	.955	.950	.962
	X_2		.980	.977	.988	.985	.988	.988
	X_3		.969	.968	.972	.976	.972	.971
	Noise		.933	.930	.943	.943	.943	.943
Bootstrap calibration	—		**.928**	**.933**	**.921**	**.924**	**.924**	**.924**
	X_1		.944	.947	.931	.940	.932	.931
	X_2		.913	.922	.910	.916	.908	.908
	X_3		.925	.931	.920	.924	.924	.924
	Noise		.928	.933	.922	.923	.925	.925

4.4. Varying effect DGP with additional noise variables

In the fourth scenario, we again considered the DGP in Equation (14). In order to investigate the performance of the proposed CIs in a higher dimensional scenario, we included seven additional noise variables $X_4, \dots, X_6 \sim N(0,1)$ and $X_7, \dots, X_{10} \sim \text{Bin}(1, 0.5)$. In such scenarios with a larger number of covariates, it is usually not meaningful to assume that each covariate can be modified by each other covariate. Otherwise, the interpretability of the TSVC model would strongly suffer. Therefore, as in scenario 3, we only allowed X_2 and X_3 as the potential effect modifiers.

It is seen from Table 7 that the average number of splits increased with the number of covariates (cf. Table 5 in scenario 3). Analogously to the previous scenarios, the number of splits decreased with sample size, and almost no splits in the effect of X_2 were performed. While the true splits with regard to X_1 were selected quite well, there was also a substantial proportion of falsely selected splits in the noise variables.

Table 8 shows that the average coverage proportions of the parametric percentile CIs were close to the nominal level across all settings. The parametric percentile CIs referring to the potential effect modifiers X_2 and X_3 yielded rather conservative results, whereas the coverage proportions for the noise variables was below the nominal level if the sample size was low ($n = 200$) but approached the 95% level as the sample size increased. Similarly to the previous scenarios, the Wald type CIs yielded highly anti-conservative coverage proportions but improved with increasing sample size. The bootstrap-calibrated CIs exhibited larger but still anti-conservative coverage proportions that changed only little across the different settings. An overall similar pattern was observed for the 90% CIs (see Table S10 in the Supplementary Material). The parametric percentile and bootstrap calibration approach yielded CIs of similar width (see Tables S11 and S12 in the Supplementary Material), which, taking together all the results, supports the finding that the symmetrically constructed bootstrap-calibrated CIs are inappropriate for TSVC models.

Table 9
Runtimes of the parametric percentile approach. Median runtimes in minutes for the four different simulation scenarios based on 100 replications in the setting with sample size $n = 200$ and standard deviation $\sigma_\varepsilon = 1$. CIs were constructed based on $B = 1000$ bootstrap samples. The first and third quartile of the runtimes are given in brackets. The calculations were performed on a high performance computing cluster that consists of 7 nodes with a total of 720 AMD Epyc CPUs and 2.2 terabytes RAM for general purpose computing. GPU nodes are one 35 Intel Xeon Gold 6 150 CPUs, 500 GB RAM, two NVIDIA V100, and one NVIDIA P40. Roughly 162 terabytes of hard drive space are available for storing project data. Connections between nodes use 10 gigabits per second Ethernet links.

Simulation scenario	Run time (in minutes)
1	15 [14, 27]
2	62 [56, 104]
3	25 [21, 42]
4	78 [70, 167]

4.5. Runtime

In the final part of the simulation study, we investigated the computing times of the proposed parametric percentile CIs. For this, we considered 100 replications of the setting with sample size $n = 200$ and standard deviation $\sigma_\varepsilon = 1$ for each of the four simulation scenarios described in Sections 4.1 to 4.4. Table 9 shows that the distributions of the runtimes for each scenario are right-skewed. This reflects that in some iterations a relatively high number of splits was performed, which increased runtime. It is also seen that the runtime of the algorithm differed strongly between the four simulation scenarios. Overall, runtimes increased with the number of covariates: Scenario 1 with only $p = 2$ covariates exhibited the lowest median runtime while the highest median runtime was observed in simulation scenario 4 with $p = 10$. On the other hand, the computational complexity can be decreased substantially by prespecification of the effect modifiers: Scenario 2, where all covariates served as potential effect modifiers, showed runtimes that were about twice as high as in scenario 3, where X_1 was excluded from the set of possible effect modifiers. Limiting the set of possible effect modifiers leads to a lower number of available splitting rules and thereby a lower number of candidate models to be fitted in each iteration of the tree-building procedure.

Overall, runtimes of the proposed algorithm depend on the number of covariates p, the number of potential effect modifiers, the maximal number of splits S, and, of course, the number of bootstrap samples B.

5. Applications

To illustrate the proposed parametric bootstrap approach for constructing CIs for TSVCs, two applications to real-world patient data were considered. The results are described in the following.

5.1. Patients with COVID-19

We considered data from a retrospective study in patients with PCR-confirmed COVID-19 that were admitted to the infectious disease department of the University Hospital Bonn between March 2020 and November 2021. A main objective of the study was to investigate the effect of treatment with the monoclonal antibody combination casirivimab/imdevimab (CVIV) on the need for oxygen support in the further course of the disease. We analyzed data from $n = 238$ patients hospitalized within five days after infection. For more details on the study, see Huebner et al. (2023). The characteristics of the patients included in our analysis are: Sex (0: female, 1: male), age in years, whether the patient suffered from another respiratory disease (0: no, 1: yes), and treatment with CVIV (0: no, 1: yes).

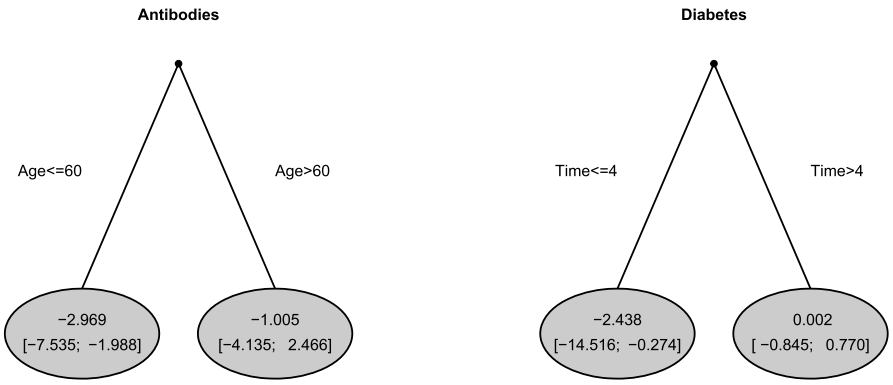
Huebner et al. (2023) analyzed the data using propensity score-weighted logistic regression. A need for oxygen support was shown to be significantly less frequent following treatment with CVIV (at error level $\alpha = 0.05$). Exploratory analyses indicated higher age as one of most relevant risk factors for requiring oxygen support in COVID-19 patients.

Our objective was to detect possible interactions between the four variables and we allowed each variable to be modified by each other variable. In order to do so, we fitted a logistic TSVC model with binary outcome 'need for oxygen support' (yes/no) to the data, where the BIC was used to select the optimal number of splits, the maximal number of splits considered was $S = 5$ and we set the minimal bucket size to $n_{\text{mb}} = 20$. Then we applied the proposed parametric bootstrap approach to obtain percentile CIs of the odds ratios based on $B = 1000$ bootstrap samples. For comparison, we also calculated asymptotic normal distribution-based Wald type CIs.

The results in Table 10 and Fig. 3a show that one split in the treatment effect with regard to age at split point 60 years was performed. According to the coefficient estimates, patients of age 60 years or younger benefited more from the CVIV treatment than patients older than 60 years. The Wald type CIs implied significant effects of age and CVIV treatment in both identified age groups at error level $\alpha = 0.05$. The proposed parametric percentile CIs were much wider and indicated only a significant effect of treatment with CVIV for the group of patients aged 60 years or younger but no significant treatment effect for patients older than 60 years. Sex and the presence of another respiratory disease both showed no evidence for an effect.

Table 10
Effect estimates, odds ratios, and 95% CIs of the odds ratios from the logistic TSVC model fitted to the COVID-19 data.

Covariate	Partition	β	exp(β)	Wald type CI	Parametric percentile CI
Sex	—	0.399	1.491	[0.826; 2.709]	[0.814; 2.987]
Age	—	0.022	1.023	[1.004; 1.041]	[0.983; 1.053]
Respiratory disease	—	-0.028	0.972	[0.426; 2.305]	[0.420; 3.131]
CVIV	Age ≤ 60	-2.969	0.051	[0.012; 0.158]	[0.000; 0.137]
	Age > 60	-1.005	0.366	[0.170; 0.785]	[0.016; 11.771]



(a) Effect of antibodies on need for oxygen support (b) Effect of diabetes on LOS in hospital

Fig. 3. Varying coefficients detected in the real-world applications. The figures show the tree-structured representation of the varying effects of antibody treatment on need for oxygen support and of diabetes on LOS in hospital for patients suffering from COVID-19 and acute odontogenic infection, respectively.

5.2. Patients with acute odontogenic infection

In a second application, we analyzed data from a retrospective study investigating hospitalized patients with abscess of odontogenic origin conducted between 2012 and 2017 by the Department of Oral and Cranio-Maxillo and Facial Plastic Surgery at the University Hospital Bonn. Patients with an acute odontogenic infection suffer from pain, swelling, erythema and hyperthermia. If not treated at an early stage, such infections may spread into deep neck spaces and lead to perilous complications by menacing anatomical structures, such as major blood vessels, the upper airway and the mediastinum (Biasotto et al., 2004). The primary objective of the study was to identify risk factors that are associated with a prolonged length of stay (LOS) in the treatment of severe odontogenic infections. LOS was recorded in days ($t = 1, \ldots, 18$). Here data from 303 patients that underwent surgical treatment in terms of incision and drainage of the abscess were considered. Intravenous antibiotics were administered during the operation and for the length of inpatient treatment. Further details on the study can be found in Heim et al. (2019). Characteristics of the patients relevant for modeling were: age in years, spreading of the infection focus into facial spaces (0: no, 1: yes), and the presence of diabetes mellitus type 2 (0: no, 1: yes).

Puth et al. (2020) analyzed the data using a logistic discrete hazard model with tree-structured varying coefficients (see the Supplementary Material for more details on the logistic discrete hazard model). Specifically, they allowed for the coefficients of all covariates to be modified by t (time since admission), which was considered as the only potential effect modifier. The number of splits performed was determined using a permutation test (Berger et al., 2019). Their findings indicated that the effect of diabetes is modified by t, where patients suffering from diabetes are much less likely to be discharged within the first four days since admission but after four days the effect of diabetes vanishes.

We analyzed the data analogously allowing only t as potential effect modifier, except using BIC to obtain the optimal number of splits with a maximal number of splits of $S = 5$ and a minimal bucket size of $n_{\text{mb}} = 20$. With the adapted strategy we were able to reproduce the fitted model structure by Puth et al. (2020). Then we applied Wald type CIs and the proposed parametric bootstrap approach to obtain 95% CIs of the coefficients based on $B = 1000$ bootstrap samples. As detailed in the Supplementary Material the exponential coefficients refer to the ratios of the continuation ratios.

The results are shown in Table 11 and Fig. 3b. The parametric percentile CIs were again much wider than the Wald type CIs and indicated a significant effect of spreading of the infection focus into facial spaces on the time to discharge but no significant effect of age (in contrast to the Wald type CI). Importantly, both CI methods indicated that diabetes significantly decreased the probability of being discharged within the first 4 days since admission, whereas no effect of diabetes after day 4 was shown, confirming the findings by Puth et al. (2020).

Table 11
Effect estimates, exponential coefficients, and 95% CIs of the exponential coefficients from the logistic discrete hazard TSVC model fitted to the odontogenic infection data.

Covariate	Partition	β	exp(β)	Wald type CI	Parametric percentile CI
Age	—	-0.008	0.992	[0.983; 0.999]	[0.983; 1.001]
Spreading	—	-0.939	0.391	[0.255; 0.584]	[0.204; 0.614]
Diabetes	$t \leq 4$	-2.438	0.087	[0.004; 0.409]	[0.000; 0.760]
	$t > 4$	0.002	1.000	[0.578; 1.695]	[0.430; 2.159]

6. Summary and discussion

TSVC models are flexible tools for generalized regression that allow the linear effects of the covariates to vary with the effect modifiers and apply a tree building procedure to inherently detect relevant effect modifiers. Constructing CIs for TSVCs is a selective inference problem as statistical inference is performed after model selection. In this vein, we proposed a parametric bootstrap approach tailored to the complex selection mechanism of TSVC.

The applications to real-world data from COVID-19 patients and from patients suffering from acute odontogenic infection showed that the proposed CIs may differ strongly from naive Wald type CIs and lead to different conclusions when assessing statistical significance of the coefficients. Both, the effect of CVIV in the group of elderly patients and the effect of diabetes within the first four days of hospitalization are highly clinically meaningful. This highlights that accounting for the selective inference problem is essential when statistical inference on the parameters of a TSVC model is of interest. In the simulation study, our approach yielded coverage proportions close to the nominal level for the linear DGP whereas the simple Wald type CI and the bootstrap calibration approach by Loh et al. (2019) showed insufficient coverage. Low coverage proportions of bootstrap-calibrated CIs are also in line with findings in previous papers (Neufeld et al., 2022). The results of simulation scenario 3 (where the effect modifiers were prespecified before model fitting) also demonstrate that the performance of the CI methods depends on the complexity of the selection problem. In more complex scenarios, where the effect modifiers are not known beforehand (scenario 2), the proposed approach showed rather conservative results for the coefficients of continuous covariates, whereas Wald type and bootstrap-calibrated CIs yielded coverage proportions that were far too low. Overall, naive Wald type CIs are markedly out of target if there is strong uncertainty of the predictor-response relationship and there is only weak evidence of possible interactions (and the associated split points) given a set of covariates. For example, the large discrepancy between the Wald type CI and the parametric percentile CI in the COVID-19 application indicates that there is only weak evidence for the split at 60 years of age. On the other hand, if the group of elderly above 60 years of age clearly benefited less from the treatment, there would be less uncertainty in the interaction effect and the Wald type CI for CVIV: Age$>$60 may be closer to the parametric percentile CI.

The parametric bootstrap procedure offers an approximate solution for conditioning on the model selection event. As further refinement, weighted percentiles could be applied, where the bootstrap estimates from models with tree structures that are more similar to the originally fitted TSVC model are given more weight in the percentile calculation of the proposed algorithm (see step 4 in the algorithm described in Section 3.2).

TSVC models can be fitted using the eponymous R add-on package (Berger, 2025). While the implementation generally allows that the effect of each covariate is modified by the other variables, the package also enables to flexibly incorporate prior knowledge about the model structure. For instance, if the effect modifiers are known beforehand (as in simulation scenario 3 and 4 and the application to the odontogenic infection data), this can be specified in the arguments of the TSVC modeling function. It is also possible to declare covariates having fixed non-varying linear effects, and covariates that serve as effect modifiers, only.

Berger et al. (2019) proposed to apply permutation tests as an early stopping criterion for the tree building in TSVC models. In each iteration, a test is performed to assess whether a further split should be performed or not. That is, these tests allow for inference on the difference between the coefficients in two nodes but not on the coefficients themselves. Our approach fills this important gap and allows to quantify uncertainty of the parameter estimates and to assess statistical significance of the varying coefficients in each node. Of note, the proposed CI approach can be used in combination with any stopping criterion for the tree building in TSVC, including permutation tests and minimal node size tuning. Here, we selected the optimal number of splits based on the BIC. Alternatively, the Akaike information cirterion (AIC) or the cross-validated predictive log-likelihood can be applied.

The proposed CI approach may easily be extended to construct confidence intervals for parameters of other tree-based models, which is a promising focus for future research. Examples include probability estimates in the leaf nodes of a classification tree, parametric models in model-based recursive partitioning (Zeileis et al., 2008), and survival trees (Schmid et al., 2016; Spuck et al., 2023). In order to perform the parametric bootstrap, it is required to make an appropriate distributional assumption. Distributional regression within the scope of tree-based models has currently been studied by several authors, see, among others, Schlosser et al. (2019) and Weinhold et al. (2020).

Acknowledgements

Support by the German Research Foundation (DFG), grant BE 7543/1-1, is gratefully acknowledged. Many thanks to Nils Heim for providing the data of the patients suffering from odontogenic infection.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2025.108142.

References

Berger, M., 2025. TSVC: tree-structured modelling of varying coefficients. https://CRAN.R-project.org/package=TSVC. R package version 1.7.2.
Berger, M., Tutz, G., Schmid, M., 2019. Tree-structured modelling of varying coefficients. Stat. Comput. 29, 217–229. https://doi.org/10.1007/s11222-018-9804-8.
Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., 2013. Valid post-selection inference. Ann. Appl. Stat. 41 (2). https://doi.org/10.1214/12-aos1077.
Biasotto, M., Pellis, T., Cadenaro, M., Bevilacqua, L., Berlot, G., Di Lenarda, R., 2004. Odontogenic infections and descending necrotising in mediastinitis: case report and review of the literature. Int. Dent. J. 54 (2), 97–102. https://doi.org/10.1111/j.1875-595X.2004.tb00262.x.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, J.C., 1984. Classification and Regression Trees. Taylor and Francis, Moneterey, CA Wadsworth.
Buergin, R., Ritschard, G., 2015. Tree-based varying coefficient regression for longitudinal ordinal responses. Comput. Stat. Data Anal. 86, 65–80. https://doi.org/10.1016/j.csda.2015.01.003.
Buergin, R., Ritschard, G., 2017. Coefficient-wise tree-based varying coefficient regression with vcrpart. J. Stat. Softw. 80 (6). https://doi.org/10.18637/jss.v080.i06.
Efron, B., Tibshirani, R., 1993. An Introduction to Bootstrap. Chapman & Hall, New York.
Fan, J., Zhang, W., 2008. Statistical methods with varying coefficient models. Stat. Interface 1, 179–195. https://doi.org/10.4310/SII.2008.v1.n1.a15.
Fithian, W., Sun, D., Taylor, J., 2014. Optimal inference after model selection. arXiv:1410.2597.
Gottard, A., Vannucci, G., Grilli, L., Rampichini, C., 2023. Mixed-effect models with trees. Adv. Data Anal. Classif. 17 (2), 431–461. https://doi.org/10.1007/s11634-022-00509-3.
Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. J. R. Stat. Soc., Ser. B, Stat. Methodol. 55, 757–779. https://doi.org/10.1111/j.2517-6161.1993.tb01939.x.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second edition. Springer, New York.
Heim, N., Berger, M., Wiedemeyer, V., Reich, R., Martini, M., 2019. A mathematical approach improves the predictability of length of hospitalization due to acute odontogenic infection. A retrospective invetigation of 303 patients. J. Maxillofac. Surg. 47 (2), 334–340. https://doi.org/10.3844/jmssp.2019.354.365.
Huebner, Y.R., Spuck, N., Berger, M., Schlabe, S., Rieke, G.J., Breitschwerdt, S., van Bremen, K., Strassburg, C.P., Gonzalez-Carmona, M.A., Wasmuth, J.-C., Rockstroh, J.K., Boesecke, C., Monin, M.B., 2023. Antiviral treatment of covid-19: which role can clinical parameters play in therapy evaluation? Infection 51, 1855–1861. https://doi.org/10.1007/s15010-023-02081-0.
Lee, J., Li, G., Wilson, J.D., 2020. Varying-coefficient models for dynamic networks. Comput. Stat. Data Anal. 152. https://doi.org/10.1016/j.csda.2020.107052.
Lee, J.D., Sun, D.L., Sun, Y., Taylor, J.E., 2016. Exact post-selection inference, with application to the lasso. Ann. Stat. 44 (3). https://doi.org/10.1214/15-aos1371.
Loh, W.-Y., Man, M., Wang, S., 2019. Subgroups form regression trees with adjustment for prognostic effect and post selection inference. Stat. Med. 38 (4). https://doi.org/10.1002/sim.7677.
Neufeld, A.C., Gao, L.L., Witten, D.M., 2022. Tree-values: selective inference for regression trees. J. Mach. Learn. Res. 23, 1–43.
Park, B.U., Mammen, E., Lee, Y.K., Lee, E.R., 2015. Varying coefficient regression models: a review and new developments. Int. Stat. Rev. 83, 36–64. https://doi.org/10.1111/insr.12029.
Puth, M.-T., Tutz, G., Heim, N., Muenster, E., Schmid, M., Berger, M., 2020. Tree-based modeling of time-varying coefficients in discrete time-to-event models. Lifetime Data Anal. 26 (3), 545–572. https://doi.org/10.1007/s10985-019-09489-7.
Ruegamer, D., Greven, S., 2018. Selective inference after likelihood- or test-based model selection in linear models. Stat. Probab. Lett. 140, 7–12. https://doi.org/10.1016/j.spl.2018.04.010.
Ruegamer, D., Baumann, P.F.M., Greven, S., 2022. Selective inference for additive and linear mixed models. Comput. Stat. Data Anal. 167. https://doi.org/10.1016/j.csda.2021.107350.
Schlosser, L., Hothorn, T., Stauffer, R., Zeileis, A., 2019. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. Ann. Appl. Stat. 13, 1564–1589. https://doi.org/10.1214/19-AOAS1247.
Schmid, M., Kuechenoff, H., Hoerauf, A., Tutz, G., 2016. A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. Stat. Med. 35 (5), 734–751. https://doi.org/10.1002/sim.6729.
Schwarz, G.E., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464. https://doi.org/10.1214/aos/1176344136.
Spuck, N., Schmid, M., Heim, N., Klarmann-Schulz, U., Hoerauf, A., Berger, M., 2023. Flexible tree-structured regression models for discrete event times. Stat. Comput. 33 (20). https://doi.org/10.1007/s11222-022-10196-x.
Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K., Takeuchi, I., 2017. Selective inference for sparse high-order interaction models. In: International Conference on Machine Learning. PMLR, pp. 3338–3347.
Taylor, J., Tibshirani, R.J., 2015. Statistical learning and selective inference. Proc. Natl. Acad. Sci. 112 (25), 7629–7634. https://doi.org/10.1073/pnas.1507583112.
Tibshirani, R.J., Taylor, J., Lockhart, R., Tibshirani, R., 2016. Exact post-selection inference for sequential regression procedures. J. Am. Stat. Assoc. 111, 600–620. https://doi.org/10.1080/01621459.2015.1108848.
Wang, J., Hastie, T., 2014. Boosted varying-coefficient regression models for product demand prediction. J. Comput. Graph. Stat. 23 (2), 361–382. https://doi.org/10.1080/10618600.2013.778777.
Weinhold, L., Schmid, M., Mitchell, R., Maloney, K.O., Wright, M.N., Berger, M., 2020. A random forest approach for bounded outcome variables. J. Comput. Graph. Stat. 29, 639–658. https://doi.org/10.1080/10618600.2019.1705310.
Zeileis, A., Horthorn, T., Hornik, K., 2008. Model-based recursive pratitioning. J. Comput. Graph. Stat. 17 (2), 492–514. https://doi.org/10.1198/10618600SX319331.
Zhang, D., Khalili, A., Asgharian, M., 2022. Post-model-selection inference in linear regression models: an integrated review. Stat. Surv. 16. https://doi.org/10.1214/22-ss135.
Zhao, Q., Small, D.S., Ertefaie, A., 2022. Selective inference for effect modification via the lasso. J. R. Stat. Soc., Ser. B, Stat. Methodol. 84, 382–413. https://doi.org/10.1111/rssb.12483.
Zhou, Y., Hooker, G., 2022. Decision tree boosted varying coefficient models. Data Min. Knowl. Discov. 36, 2237–2271. https://doi.org/10.1007/s10618-022-00863-y.
Zrnic, T., Jordan, M.I., 2023. Post-selection inference via algorithmic stability. Ann. Stat. 51, 1666–1691. https://doi.org/10.1214/23-AOS2303.

4 Discussion with references

The articles in this dissertation enhance the applicability of tree-based regression in biomedical research by introducing novel flexible tree-based models tailored to different data scenarios and by developing a framework for selective inference for TSVCs.

The application of the survival tree in *Publication 2* (Elahi et al., 2023) illustrated the advantages of tree-based modeling for the analysis of biomedical data. Specifically, the survival tree selected three characteristics of the OSCC patients (type of flap, age, and whether an early complication occurred) that affected the LOS most strongly from a set of more than fifteen covariates, which illustrates the inherent variable selection in tree-based models. In addition, the tree-based modeling approach facilitated the detection of important interactions between time and the characteristics of the patients. Type of flap and age were shown to affect the probability of being discharged from hospital only within the first twelve days since admission, whereas the occurrence of complications affected LOS only after day twelve. The graphical representation as a hierarchical tree (see *Publication 2*, Fig. 1) allows for an intuitive interpretation of the results. Furthermore, the results of the simulation studies confirmed that tree-based modeling is highly effective at selecting relevant covariates and capturing interaction effects, in particular, in high-dimensional settings, see *Publication 1* (Spuck et al., 2023) and the *Unpublished Manuscript* (Spuck et al., 2025 a). On the other hand, the proposed tree-based models performed worse than the linear competitors in scenarios with linear data generating processes (Spuck et al., 2023; 2025 a).

Notably, all proposed models can be fitted within the framework of TSVC by applying the eponymous R add-on package (Berger, 2025). Tree-based modeling of varying coefficients was further explored, for example, by Bürgin and Ritschard (2015) for longitudinal ordinal data, by Puth et al. (2020) in the context of time-varying coefficients in time-to-event analysis, and by Zakrisson and Lindholm (2025) in combination with cyclic gradient boosting. The additive structure of the predictor functions and the likelihood-based TSVC fitting procedure make the proposed models easily generalizable and allow extensions that combine the tree structures with linear terms. This may be particularly advantageous in biomedical applications, where

prior knowledge on the effects of some of the covariates is available. Due to the likelihood-based TSVC model fitting procedure, the proposed methods can be applied with outcome variables on a variety of different scales and with a number of different link functions.

In Spuck et al. (2023), various criteria to limit the complexity of the tree-based models and avoid overfitting were compared. The post-pruning strategy yielded models with a predictive performance similar to the rather conservative permutation test approach (Berger et al., 2019). Therefore, less computationally intensive post-pruning strategies based on cross-validation (Spuck et al., 2025 a) and an information criterion (Spuck et al., 2025 b) were adapted in the following projects. Note, however, that the post-pruning strategy is unable to guarantee a prespecified error level and may lead to less parsimonious models.

It is a well-known result that tree-based models suffer from high variability and are prone to overfitting (Hastie et al., 2009). To overcome these challenges and improve predictive performance, the proposed models can be extended to ensemble methods. Tree-based ensemble methods were studied previously, for example, by Schmid et al. (2020) and Moradian et al. (2022) in the context of discrete time-to-event data and by Hajjem et al. (2012) and Speiser et al. (2019) in the context of clustered data.

The simulation results in *Publication 3* (Spuck et al., 2025 b) confirm the selective inference problem with tree-based approaches (cf. Neufeld et al., 2022). In particular, asymptotic normal distribution-based CIs for TSVCs yielded insufficient coverage proportions across all simulation scenarios. The proposed CI method resulted in coverage proportions closer to the nominal level with more conservative results as the complexity of the model selection problem increased. Of note, the proposed approach was designed for TSVC models, but can easily be adapted to construct CIs for parameters from other tree-based models, for instance, model-based recursive partitioning (Zeileis et al., 2008), and models for discrete time-to-event analysis (Schmid et al., 2016; Spuck et al., 2023), or clustered data (Spuck et al., 2025 a), which appears promising to explore in future research. However, unlike other approaches for selective inference (Tibshirani et al., 2016; Neufeld et al., 2022), the proposed bootstrap approach is not based on classical statistical theory and should be viewed as an approximate

solution for complex tree-based models, where no asymptotic approach is currently available. The ability of the proposed tree-based methods to handle very high-dimensional data is limited by computational cost. The CI method in Spuck et al. (2025 b) is particularly demanding due to the repeated application of the complex TSVC fitting procedure, where the number of required optimizations of the likelihood function grows with the number of covariates.

4.1 Conclusion

In summary, this cumulative dissertation increases the applicability of tree-based regression approaches with a focus on biomedical research. A major contribution is the extension of the catalog of available tree-based models for discrete time-to-event and clustered data. Furthermore, the dissertation outlines a novel approach for inference on parameters from tree-based models and thereby helps overcome one of the major drawbacks of tree-based approaches.

4.2 References

Berger M. TSVC: Tree-Structured Modelling of Varying Coefficients. R package version 1.7.2, 2025. URL: `https://CRAN.R-project.org/package=TSVC`

Berger M, Tutz G, Schmid M. Tree-structured modelling of varying coefficients. Statistics and Computing 2019; 29: 217–229

Bürgin R, Ritschard G. Tree-based varying coefficient regression for longitudinal ordinal responses. Computational Statistics and Data Analysis 2015; 86: 65–80

Elahi F, Spuck N, Berger M, Kramer FJ, Heim N. Mathematical approach improves predictability of length of hospitalisation due to oral squamous cell carcinoma: a retrospective investigation of 153 patients. British Journal of Oral and Maxillofacial Surgery 2023; 61: 605–611

Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. Journal of Computation and Simulation 2012; 84: 1313–1328

Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York: Springer, 2009

Moradian H, Yao W, Larocque D, Simonoff D, Frydman H. Dynamic estimation with random forests for discrete-time survival. Canadian Journal of Statistics 2022; 50: 533–548

Neufeld AC, Gao LL, Witten DM. Tree-Values: Selective Inference for Regression Trees. Journal of Machine Learning and Research 2022; 23: 1–43

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

Schmid M, Küchenoff H, Hörauf A, Tutz G. A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. Statistics in Medicine 2016; 35: 734–751

Schmid M, Welchowski T, Wright MN, Berger M. Discrete-time survival forests with Hellinger distance. Data Mining and Knowledge Discovery 2020; 34: 812–832

Speiser JL, Wolf BJ, Chung D, Karvellas CJ, Koch DG, Durkalski VL. BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. Chemometrics and Intelligent Laboratory Systems 2019; 185: 122–134

Spuck N, Schmid M, Berger M. Flexible tree-structured regression for clustered data with an application to quality of life in older adults. Uploaded to arXiv 2025 a; 2501.12787

Spuck N, Schmid M, Heim N, Klarmann-Schulz U, Hörauf A, Berger M. Flexible tree-structured regression models for discrete event times. Statistics and Computing 2023; 33: 1–21

Spuck N, Schmid M, Monin M, Berger M. Confidence intervals for tree-structured varying coefficients. Computational Statistics and Data Analysis 2025 b; 207: 108142

Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R. Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association 2016; 111: 600–620

Zakrisson H, Lindholm M. A tree-based varying coefficient model. Computational Statistics 2025: Published online

Zeileis A, Hothorn T, Hornik K. Model-based recursive pratitioning. Journal of Computational and Graphical Statistics 2008; 17: 492–514

Complete publication list

Elahi F, Spuck N, Berger M, Kramer FJ, Heim N. Mathematical approach improves predictability of length of hospitalisation due to oral squamous cell carcinoma: a retrospective investigation of 153 patients. British Journal of Oral and Maxillofacial Surgery 2023; 61: 605–611

Huebner Y, Spuck N, Berger M, Schlabe S, Rieke G, Breitschwerdt S, van Bremen K, Strassburg CP, Gonzalez-Carmona MA, Wasmuth JC, Rockstroh JK, Boesecke C, Monin M. Antiviral treatment of COVID-19: which role can clinical parameters play in therapy evaluation? Infection 2023; 51: 1855–1861

Massoth G, Vorhofer E, Spuck N, Mikus M, Mini N, Strassberger-Nerschbach N, Wittmann M, Neumann C, Schindler E. Impact of the mother's voice on sedation need and stress during cardiologic examination of children (SMUSS study): a prospective, interventional, randomised, controlled, monocentric study. Cardiology in the Young 2024; 34: 2437–2444

Orth HM, Flasshove C, Berger M, Hattenhauer T, Biederbick KD, Mispelbaum R, Klein U, Stemler J, Fisahn M, Doleschall AD, Baermann BN, Koenigshausen E, Tselikmann O, Killer A, de Angelis C, Gliga S, Stegbauer J, Spuck N, Silling G, Rockstroh JK, Strassburg CP, Brossart P, Panse JP, Jensen BEO, Luedde T, Boesecke C, Heine A, Cornely OA, Monin MB. Early combination therapy of COVID-19 in high-risk patients. Infection 2024; 52: 877–889

Spuck N, Schmid M, Heim N, Klarmann-Schulz U, Hörauf A, Berger M. Flexible tree-structured regression models for discrete event times. Statistics and Computing 2023; 33: 1–21

Spuck N, Schmid M, Monin M, Berger M. Confidence intervals for tree-structured varying coefficients. Computational Statistics and Data Analysis 2025 b; 207: 108142

Thol F, Warwas B, Spuck N, Kramer FJ, Heim N. Microbial spectrum and resistance of odontogenic abscesses - microbiological analysis using next generation sequencing. Clinical Oral Investigations 2024; 29: 8

Unpublished manuscripts

Spuck N, Schmid M, Berger M. Detection of nonlinearity, discontinuity and interactions in generalized regression models: Uploaded to arXiv 2023; 2310.20409

Spuck N, Schmid M, Berger M. Flexible tree-structured regression for clustered data with an application to quality of life in older adults. Uploaded to arXiv 2025 a; 2501.12787

5 Acknowledgments

First, I would like to thank my supervisor Prof. Dr. Matthias for his helpful guidance and constructive feedback on my projects. In addition, I am highly grateful to PD Dr. Moritz Berger for all of his support, his open-minded approach to challenges, and the many exciting scientific discussions we had. I would also like to thank the members of my PhD committee, Prof. Dr. Andreas Groll, Prof. Dr. Robert Finger, and Prof. Dr. Michael Wagner, for their valuable input. Many thanks to all of my (present and past) colleagues from IMBIE for a great work atmosphere, many enjoyable conversations, and fun times at conferences. I am particularly grateful to Jenny and David for being such great and supportive office mates. Special thanks go to Charlotte, David, and Moritz for their feedback and proofreading of my thesis and to Annika, Charlotte, and Hannah for their help with its formalities.

Finally, I am immensely grateful to my family and friends who supported and encouraged me and were there for me whenever I needed them. Thank you for everything.