

Statistical learning for multivariate distributional regression with complex dependencies

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

Annika Lisa Strömer

from Weener

2025

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

First reviewer: Prof. Dr. Andreas Mayr
Second reviewer: Prof. Dr. Nadja Klein

Day of oral examination: 27.11.2025

From the Institute of Medical Biometry, Informatics and Epidemiology

Table of Contents

List of abbreviations	5
1 Abstract	6
2 Introduction and aims with references	8
2.1 Thesis outline	10
2.2 Advanced modeling approaches	11
2.2.1 Multivariate trajectories in obesity surgery	11
2.2.2 Multivariate distributional regression	12
2.2.3 Dependent censoring in time-to-event data	12
2.3 Enhanced variable selection and complexity reduction	13
2.4 References	15
3 Publications	18
3.1 Publication A: Multivariate trajectories of weight and mental health and their prognostic significance six years after obesity surgery	18
3.2 Publication B: Boosting multivariate structured additive distributional regression models	32
3.3 Publication C: Modelling dependent censoring in time-to-event data by boosting copula regression	56
3.4 Publication D: Deselection of base-learners for statistical boosting — with an application to distributional regression	80
3.5 Publication E: Enhanced variable selection for boosting sparser and less complex models in distributional copula regression	99
4 Discussion with references	123
4.1 Conclusion	126
4.2 References	127

5	Acknowledgements	132
6	Statement	133

List of abbreviations

EDE-Q	Eating disorder examination – questionnaire
GAMs	Generalized additive models
GAMLSS	Generalized additive models for location, scale and shape
GLMs	Generalized linear models
HRQoL	Health-related quality of life
LCMMs	Latent class linear mixed models
PHQ-D	Patient health questionnaire – depression
PRAC	Psychosocial registry for obesity surgery
STAR	Structured additive regression

1 Abstract

Large, complex datasets are becoming increasingly important in biomedical research. Such datasets typically feature a high number of variables per subject, multiple outcomes and complex dependency structures. While they provide new opportunities to examine scientific questions in greater detail, they also pose major statistical challenges. Addressing these challenges requires advanced methods that can handle high dimensionality, capture dependencies between correlated outcomes and provide interpretable results.

This cumulative dissertation develops statistical frameworks for multivariate distributional regression and variable selection techniques, enabling the analysis of complex biomedical data while balancing flexibility, interpretability and efficiency. It comprises five publications covering methodological advances and applications in diverse biomedical contexts.

The first project demonstrates the value of advanced multivariate modeling for uncovering clinically relevant patterns in complex longitudinal data. Using latent class linear mixed models (LCMMs), unobserved patient subgroups are identified with distinct five-year trajectories in weight, depressive symptoms, eating disorder psychopathology and health-related quality of life (HRQoL) after obesity surgery. The results show that physical and psychological changes can evolve differently over time and may vary in sustainability, underscoring the need for joint models that capture both interdependencies and heterogeneity. The second project develops a model-based boosting approach for multivariate distributional regression within the framework of generalized additive models for location, scale and shape (GAMLSS). This method enables simultaneous modeling of all distribution parameters – including dependence parameters – of arbitrary parametric multivariate outcomes as functions of covariates. It incorporates data-driven variable selection and scales to high-dimensional settings where the number of covariates exceeds the number of observations ($p \gg n$). Building on this, the third project tackles the issue of dependent censoring in survival analysis, a challenging scenario where the common assumption of independent censoring does not hold. In such cases, censoring may be related to the patient's health status; for instance, patients in poorer condition may withdraw from a study earlier. The work proposes a novel model-based boosting method

using distributional copula regression to jointly model the marginal distributions of event and censoring times as well as their dependence, as functions of covariates. The fourth and fifth papers address the challenge of improving interpretability in model-based boosting, particularly for high-dimensional biomedical data. While boosting provides flexibility, it may result in overly complex models by including covariates with negligible importance. The fourth paper proposes a deselection approach for univariate (distributional) regression that removes irrelevant predictors with only a minor impact on the prediction of the model, yielding simpler and more interpretable models without compromising predictive performance. The fifth paper extends this approach to distributional copula regression, enabling not only the removal of variables with minor importance but also the determination of whether specific parameters require covariate effects. This controls model complexity and enhances interpretability. This dissertation includes five research articles published in peer-reviewed international journals (*Publication A – E*).

2 Introduction and aims with references

Large and complex data sets are playing an increasingly important role in many areas of biomedical research. Recent technical developments in molecular medicine and genetics, combined with the emergence of new digital measurement instruments, have enabled researchers to generate vast amounts of data, often with a high number of variables per subject – even in traditional clinical studies and epidemiological registries (Jordan et al., 2024; Dattangire and Biradar, 2024). The resulting data dimensionality creates new opportunities to investigate complex scientific questions that were previously difficult to study due to limitations in data availability, while simultaneously introducing major statistical challenges. Hence, there is an ongoing need for advanced statistical methods capable of handling not only high dimensionality but also the complex dependencies that may arise among multiple, potentially correlated outcomes.

Classical regression models, including generalized linear models (GLMs; McCullagh and Nelder 1989), generalized additive models (GAMs; Hastie and Tibshirani, 1990) and structured additive regression models (STAR; Brezger and Lang, 2006), are designed to estimate the conditional mean of a univariate response variable as a function of covariates. While these approaches are well-established and widely used across many fields, they may not be sufficient to address the emerging challenges mentioned above. In many biomedical applications, such as health-related quality of life (HRQoL) analysis, it is essential to explore not just insights obtained through mean regression, but also how the entire distribution, including variability, changes with patient characteristics (Ferrari and Cribari-Neto, 2004; Hunger et al., 2011; Hunger et al., 2012). Distributional regression models address these needs by modeling the full conditional distribution, enabling richer insights than mean regression alone. Among these, generalized additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005) are a well-known example, allowing covariates to influence multiple distribution parameters, including location, scale and shape, thus enabling detailed characterization of distributional properties for outcomes such as HRQoL. Despite these advantages, GAMLSS and similar models are typically restricted to univariate settings with one outcome

variable, and thus can not capture dependencies among multiple correlated outcomes. Such dependencies are common and relevant in biomedical research and arise in diverse forms: one example where multivariate modeling could be particularly valuable is in long-term clinical studies, where physical and psychological health outcomes are often measured in parallel after medical interventions, with improvements that may occur simultaneously yet follow distinct temporal patterns (Dawes et al., 2016; Kalarchian et al., 2019; Hilbert et al., 2022). Another example is the analysis of multiple biomarkers or spatially correlated health outcomes. For instance, in child malnutrition research, indicators such as stunting, underweight and wasting are related and frequently co-occur, suggesting a need for joint modeling to identify shared risk factors (Tesfaw and Fenta, 2021).

To further increase flexibility and move beyond univariate responses, multivariate modeling frameworks have been developed, extending traditional regression methods to model multiple outcomes jointly. For example, the work by Klein et al. (2015) builds on the GAMLSS framework to allow for flexible, joint modeling of multivariate response distributions. The approach enables the simultaneous estimation of marginal distributions and the dependencies between outcomes as functions of covariates. These multivariate distributions, such as the multivariate Gaussian, are often used when the outcomes are of the same type and well-behaved, making them particularly suitable in biomedical contexts. However, some applications, such as the study of risk factors for adverse birth outcomes (Klein et al., 2019), involve diverse types of outcomes (e.g., combinations of binary and continuous responses) as well as dependencies that are asymmetric, nonlinear or concentrated in the tails of the joint distribution. Copula-based models provide a flexible and interpretable alternative. By decoupling the specification of marginal distributions from the dependence structure (Sklar, 1959; Nelsen, 2006), copulas enable the construction of joint models that accommodate diverse types of outcomes and dependence structures.

The curse of dimensionality frequently results in data situation where the number of covariates exceeds the number of observations ($p \gg n$). These datasets are commonly referred to as high-dimensional datasets and pose further challenges for statistical modeling. In such

situations, regression methods must not only be flexible but also incorporate variable selection to identify the most relevant covariates. Statistical learning approaches have been developed to address these challenges by combining flexible modeling with mechanisms for variable selection and regularization. Such methods are particularly effective for analyzing high-dimensional datasets, allowing for the extraction of meaningful patterns while controlling overfitting and noise (James et al., 2022). In this context, statistical boosting has emerged as a powerful tool (Bühlmann and Hothorn, 2007; Mayr et al., 2014). It offers a modular framework for high-dimensional regression that integrates flexible predictor specifications, structured additive modeling and data-driven variable selection, including within the GAMLSS framework (Mayr et al., 2012; Thomas et al., 2018). This makes boosting particularly attractive for analyzing large-scale data, where both interpretability and predictive accuracy are essential. Combining boosting algorithms with advanced regression frameworks is therefore a flexible and effective solution to model modern, high-dimensional biomedical data.

2.1 Thesis outline

This cumulative dissertation aims to develop modeling techniques for multivariate distributional regression with complex dependencies, with an emphasis on high-dimensional data and variable selection, and to apply these methods in a biomedical context. It comprises five published articles, organized into two thematic sections: (i) advanced modeling approaches (*Publication A – C*) and (ii) enhanced variable selection and complexity reduction (*Publication D and E*).

The first article examines the joint multivariate trajectories of long-term improvements in weight and mental health outcomes over time among patients who have undergone obesity surgery (*Publication A*). The following two articles are embedded in the boosting framework: the second article introduces a flexible approach to multivariate distributional regression (*Publication B*), and the third article develops a copula-based method to appropriately model dependent censoring in survival analysis (*Publication C*). The fourth and fifth article focus on variable selection and complexity reduction in boosting, proposing a deselection procedure for reducing

the number of variables in univariate (distributional) regression (*Publication D*) and its extension to more complex models, enabling both variable and model simplification (*Publication E*).

A complete list of all publications from my work at the Institute of Medical Biometry, Informatics, and Epidemiology, and the Institute of Biometry and Statistics (University of Marburg) during my PhD years is provided in the Appendix.

2.2 Advanced modeling approaches

2.2.1 Multivariate trajectories in obesity surgery

While the physical benefits of obesity surgery are well-documented, there is much less information on how various aspects of patients' well-being, including weight, psychopathology, and quality of life, change together over time. Traditional methods often analyze these outcomes separately (Courcoulas et al., 2018; Voorwinde et al., 2022), which can overlook essential connections between them.

In *Publication A*, long-term changes in weight, depressive symptoms (Patient Health Questionnaire – Depression, PHQ-D), eating disorder psychopathology (Eating Disorder Examination – Questionnaire, EDE-Q), and HRQoL after obesity surgery are investigated using data from the ongoing Psychosocial Registry for Obesity Surgery (PRAC) study (Hilbert et al., 2022). The longitudinal outcomes from baseline to five years post-surgery are jointly modeled using latent class linear mixed models (LCMMs), while accounting for incomplete follow-up due to ongoing recruitment. Three distinct patient subgroups are identified, reflecting low, medium and high sustainability of improvements. These findings indicate that improvements across different outcomes do not always occur in parallel and that the sustainability of benefits varies between subgroups, underscoring the complex and multifaceted nature of recovery after obesity surgery.

2.2.2 Multivariate distributional regression

Section 2.2.1 emphasizes the need for multivariate models to capture joint outcome patterns accurately. To move beyond modeling only mean trajectories and to incorporate full distributional characteristics as well as dependence structures, multivariate distributional regression provides a flexible framework. It allows each parameter of the multivariate distribution to be modeled as a function of covariates (Klein et al., 2015). While this flexibility extends modeling capabilities, it also presents constraints, particularly in high-dimensional settings.

To address this, a novel statistical boosting approach is introduced in *Publication B*. The proposed method models all distribution parameters simultaneously while performing data-driven variable selection to identify the most relevant predictors for each parameter. Therefore, it eliminates the need to manually pre-specify which covariates affect which parameters, simplifying model building and making the approach particularly suitable for high-dimensional settings. The approach focuses on modeling and investigating bivariate regression models with emphasis on the most common parametric distributions in biomedical research: the bivariate Bernoulli distribution for binary outcomes, the bivariate Poisson distribution for count data, and the bivariate Gaussian distribution for continuous outcomes (Marshall and Olkin, 1985; Kocherlakota and Kocherlakota, 1992; Kotz et al., 2000). The flexibility and scalability of this framework are demonstrated in a simulation study and across diverse biomedical applications, including the analysis of genetic predispositions in the UK Biobank, health service utilization in Australia, and childhood malnutrition in Nigeria. By jointly modeling all distribution parameters and their associations as flexible functions of covariates, including linear, nonlinear, and interaction effects, boosting multivariate distributional regression provides an interpretable solution for complex data.

2.2.3 Dependent censoring in time-to-event data

While classical bivariate models (see Section 2.2.2) allow for the joint analysis of two fully observed outcomes, survival analysis presents a distinct challenge: some individuals may not experience the event of interest within the study period or may be lost to follow-up before its

occurrence, resulting in censored observations as an inherent feature of survival data. Such incomplete observation fundamentally distinguishes survival data from standard bivariate settings.

A central challenge is the often questionable assumption that survival and censoring times are conditionally independent given covariates. For instance, if patients withdraw from a study due to worsening health, methods that assume independent censoring may yield biased results (Huang and Zhang, 2008). To address dependent censoring, copula-based methods have gained popularity for modeling the joint distribution of survival and censoring times (Zheng and Klein, 1995; Rivest and Wells, 2001). However, identifiability is challenging because right-censoring prevents the observation of both times for the same subject. Recent advancements show that identifiability is achievable with parametric copulas and marginals, even without prior knowledge of the copula (Czado and Van Keilegom, 2023; Deresa et al., 2022). Nonetheless, most existing copula-based approaches are limited in their capacity to model the dependence between survival and censoring times as a function of covariates or to accommodate high-dimensional data.

To address these limitations, *Publication C* introduces a model-based boosting approach for dependent censoring via distributional copula regression. By specifying parametric forms for both the margins and the copula, the framework enables flexible and interpretable modeling, allowing all distributional components, including the dependence parameter, to vary with covariates and to be estimated jointly. Unlike existing methods that often assume a fixed copula or do not permit covariate effects on the dependence structure, this approach provides greater flexibility to capture complex, covariate-dependent relationships. Integration within the boosting framework ensures scalability and facilitates data-driven variable selection for all model parameters, making it well-suited for modern, high-dimensional biomedical datasets.

2.3 Enhanced variable selection and complexity reduction

While boosting algorithms offer great flexibility for modeling a broad range of univariate and multivariate regression models (as introduced in Sections 2.2.2 and 2.2.3), they often tend to

select too many variables with negligible effect sizes (Hans et al., 2023; Jobst et al., 2024). This tendency arises because their tuning procedures prioritize predictive performance rather than enforcing sparsity. However, the inclusion of redundant variables is particularly undesirable in high-dimensional settings, where sparse and interpretable models are crucial for both practical application and scientific findings. This challenge becomes even more pronounced in models involving multiple distribution parameters or multivariate outcomes, where the complexity of the model increases and the need for interpretability and sparsity becomes even more critical. In these cases, managing overall model complexity involves more than just selecting variables; it also requires assessing whether each parameter should depend on covariates or remain constant. For example, when only the mean parameter shows meaningful covariate dependence, the complexity of a full GAMLSS may be avoidable, favoring the use of simpler models like GLMs or GAMs instead.

To address this, *Publication D* introduces a deselection framework that relies on the attributable risk reduction of individual base-learners. This approach systematically removes covariates that have negligible impacts on model performance, resulting in sparser and more interpretable models without compromising prediction accuracy. The deselection procedure is controlled by a threshold parameter, which represents the minimum total risk reduction that should be attributed to a corresponding base-learner to avoid deselection (e.g., 1%, see Fig. 3, *Publication D*). Building on this, *Publication E* extends the deselection approach to multivariate distributional copula regression, where the method automatically determines whether specific parameters (e.g., copula dependence) require covariate effects or can be simplified to a constant. This data-driven complexity control is critical for high-dimensional biomedical data, where overparameterization risks overfitting and may obscure scientific insights.

2.4 References

- Brezger A, Lang S. Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, 2006; 50 (4): 967–991
- Bühlmann P, Hothorn T. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, 2007; 22 (4): 477–505
- Courcoulas AP, King WC, Belle SH, Berk P, Flum DR, Garcia L, Gourash W, Horlick M, Mitchell JE, Pomp A, Pories WJ, Purnell JQ, Singh A, Spaniolas K, Thirlby R, Wolfe BM, Yanovski SZ. Seven-Year Weight Trajectories and Health Outcomes in the Longitudinal Assessment of Bariatric Surgery (LABS) Study. *JAMA Surgery*, 2018; 153 (5): 427–434
- Czado C, Van Keilegom I. Dependent censoring based on parametric copulas. *Biometrika*, 2023; 110 (3): 721–738
- Dattangire R, Biradar D. Leveraging Big Data for Disease Surveillance and Public Health Interventions. *International Journal of Global Innovations and Solutions (IJGIS)*, 2024
- Dawes AJ, Maggard-Gibbons M, Maher AR, Booth MJ, Miake-Lye I, Beroes JM, Shekelle PG. Mental Health Conditions Among Patients Seeking and Undergoing Bariatric Surgery: A Meta-analysis. *JAMA*, 2016; 315 (2): 150–163
- Deresa N, Van Keilegom I, Antonio K. Copula-based inference for bivariate survival data with left truncation and dependent censoring. *Insurance: Mathematics and Economics*, 2022; 107: 1–21
- Ferrari S, Cribari-Neto F. Beta Regression or Modeling Rates and Proportions. *Journal of Applied Statistics*, 2004; 31: 799–815
- Hans N, Klein N, Faschingbauer F, Schneider M, Mayr A. Boosting distributional copula regression. *Biometrics*, 2023; 79 (3): 2298–2310
- Hastie T, Tibshirani R. Generalized Additive Models. London: ChapmanHall, 1990
- Hilbert A, Staerk C, Strömer A, Mansfeld T, Sander J, Seyfried F, Kaiser S, Dietrich A, Mayr A. Nonnormative Eating Behaviors and Eating Disorders and Their Associations With Weight Loss and Quality of Life During 6 Years Following Obesity Surgery. *JAMA Network Open*, 2022; 5 (8): e2226244

- Huang X, Zhang N. Regression survival analysis with an assumed copula for dependent censoring: A sensitivity analysis approach. *Biometrics*, 2008; 64 (4): 1090–1099
- Hunger M, Baumert J, Holle R. Analysis of SF-6D Index Data: Is Beta Regression Appropriate? *Value in Health*, 2011; 14 (5): 759–767
- Hunger M, Döring A, Holle R. Longitudinal beta regression models for analyzing health-related quality of life scores over time. *BMC Medical Research Methodology*, 2012; 12 (1): 144
- James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: With applications in R. New York, NY: Springer, 2022
- Jobst D, Möller A, Groß J. Gradient-Boosted Generalized Linear Models for Conditional Vine Copulas. *Environmetrics*, 2024; 35: e2887
- Jordan R, Celeste R, Bernabe E, Schwendicke F. Big Data in Epidemiology: Brave New World? *Journal of Dental Research*, 2024; 103 (11): 1047–1050
- Kalarchian MA, King WC, Devlin MJ, Hinerman A, Marcus MD, Yanovski SZ, Mitchell JE. Mental disorders and weight change in a prospective study of bariatric surgery patients: 7 years of follow-up. *Surgery for Obesity and Related Diseases*, 2019; 15 (5): 739–748
- Klein N, Kneib T, Klasen S, Lang S. Bayesian Structured Additive Distributional Regression for Multivariate Responses. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 2015; 64 (4): 569–591
- Klein N, Kneib T, Marra G, Radice R, Rokicki S, McGovern ME. Mixed binary-continuous copula regression models with application to adverse birth outcomes. *Statistics in Medicine*, 2019; 38 (3): 413–436
- Kocherlakota S, Kocherlakota K. Bivariate Discrete Distributions. New York: Dekker, 1992
- Kotz S, Balakrishnan N, Johnson N. Continuous Multivariate Distributions: Models and Applications, Volume 1, Second Edition. Wiley Series in Probability and Statistics, 2000
- Marshall AW, Olkin I. A Family of Bivariate Distributions Generated by the Bivariate Bernoulli Distribution. *Journal of the American Statistical Association*, 1985; 80 (390): 332–338
- Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms. *Methods of information in medicine*, 2014; 53 (06): 419–427

- Mayr A, Fenske N, Hofner B, Kneib T, Schmid M. Generalized additive models for location, scale and shape for high dimensional data – a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2012; 61 (3): 403–427
- McCullagh P, Nelder JA. *Generalized Linear Models*. London: ChapmanHall, 1989
- Nelsen RB. *An introduction to copulas*. New York: Springer, 2006
- Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 2005; 54 (3): 507–554
- Rivest LP, Wells MT. A Martingale Approach to the Copula-Graphic Estimator for the Survival Function under Dependent Censoring. *Journal of Multivariate Analysis*, 2001; 79 (1): 138–155
- Sklar M. *Functions de Repartition an Dimension Set Leursmarges*. Publications de L'Institut de Statistique de L'Universite de Paris, 1959; 8: 229–231
- Tesfaw LM, Fenta HM. Multivariate logistic regression analysis on the association between anthropometric indicators of under-five children in Nigeria: NDHS 2018. *BMC Pediatrics*, 2021; 21 (1): 193
- Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B. Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, 2018; 28: 673–687
- Voorwinde V, Hoekstra T, Monpellier VM, Steenhuis IH, Janssen IM, van Stralen MM. Five-year weight loss, physical activity, and eating style trajectories after bariatric surgery. *Surgery for Obesity and Related Diseases*, 2022; 18 (7): 911–918
- Zheng M, Klein JP. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 1995; 82 (1): 127–138

3 Publications

3.1 Publication A: Multivariate trajectories of weight and mental health and their prognostic significance six years after obesity surgery

Hilbert A, Strömer A, Staerk C, Schreglmann B, Mansfeld T, Sander J, Seyfried F, Kaiser S, Stroh C, Dietrich A, Schmidt R, Mayr A. Multivariate trajectories of weight and mental health and their prognostic significance six years after obesity surgery. *International Journal of Eating Disorders* 2025; 1–13.









<https://doi.org/10.1002/eat.24527>

Supplementary information can be found at:

<https://doi.org/10.1002/eat.24527>

ORIGINAL ARTICLE OPEN ACCESS

Multivariate Trajectories of Weight and Mental Health and Their Prognostic Significance 6 Years After Obesity Surgery

Anja Hilbert¹  | Annika Strömer^{1,2}  | Christian Staerk^{3,4}  | Ben Schreglmann¹  | Thomas Mansfeld⁵  | Johannes Sander⁶  | Florian Seyfried⁷  | Stefan Kaiser⁸  | Christine Stroh⁹  | Arne Dietrich¹⁰  | Ricarda Schmidt¹  | Andreas Mayr² 

¹Integrated Research and Treatment Center AdiposityDiseases, Behavioral Medicine Research Unit, Department of Psychosomatic Medicine and Psychotherapy, University of Leipzig Medical Center, Leipzig, Germany | ²Department of Medical Biometrics, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany | ³IUF—Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany | ⁴Department of Statistics, TU Dortmund University, Dortmund, Germany | ⁵Department of General Surgery, Asklepios Clinic, Hamburg, Germany | ⁶Obesity Clinic, Schön Klinik Hamburg Eilbek, Hamburg, Germany | ⁷Department of General, Visceral, Transplant, Vascular and Pediatric Surgery, University Hospital, University of Würzburg, Würzburg, Germany | ⁸Department of Visceral, Pediatric and Vascular Surgery, Hospital Konstanz, Konstanz, Germany | ⁹Department of Obesity and Metabolic Surgery, Municipal Hospital Gera, Gera, Germany | ¹⁰Department of Surgery, Clinic for Visceral, Transplantation, Thoracic and Vascular Surgery, University Hospital Leipzig, Leipzig, Germany

Correspondence: Anja Hilbert (anja.hilbert@medizin.uni-leipzig.de)

Received: 17 March 2025 | **Revised:** 7 August 2025 | **Accepted:** 7 August 2025

Action Editor: Ruth Striegel Weissman

Funding: This work was supported by the German Federal Ministry of Education and Research (Grant 01EO1501, Dr. Hilbert) and internal funds of the Behavioral Medicine Research Unit, Department of Psychosomatic Medicine and Psychotherapy, University of Leipzig Medical Center. The funder had no role in study design, data collection, data analysis, reporting of this study, and submission for publication.

Keywords: bariatric surgery | multivariate trajectory modeling | psychopathology | quality of life | weight loss

ABSTRACT

Objective: Obesity surgery (OS) results in substantial, albeit heterogeneous, long-term improvements in weight and mental health, with unclear trajectories and their associations. This study examined multivariate trajectories of weight, psychopathology, and health-related quality of life (HRQOL) after OS, and their prospective association with long-term health outcomes.

Method: In the prospective multicenter Psychosocial Registry of Obesity Surgery, $N=856$ patients were classified into multivariate trajectory classes using latent class linear mixed models, based on assessments of weight, depression, eating disorder psychopathology, and HRQOL at baseline and annually 1–5 years following OS. The prognostic significance of trajectory classes for 6-year follow-up was examined. Multivariate trajectory modeling was compared with univariate weight trajectory modeling for concordance and prognostic significance.

Results: We identified three trajectory classes of *low* (I, 2.8%), *medium* (II, 89.1%), and *high* (III, 8.1%) *sustainability* 1–5 years after OS, indicating high (I) or gradual deterioration (II) or further improvement (III) after initial improvement of indicators. The *low sustainability* class (I) reached nadir improvements earliest. Consistently, trajectory classes were prospectively associated with differential clinically significant improvement in weight and mental health at the 6-year follow-up. Multivariate trajectory modeling was discordant with univariate weight trajectory modeling and showed greater predictive value for health outcomes at the 6-year follow-up.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *International Journal of Eating Disorders* published by Wiley Periodicals LLC.

Discussion: Patients who achieve nadir improvements in weight and mental health early may require clinical attention to prevent long-term relapse. Monitoring changes in the first years after OS appears essential to identify patients in need of additional intervention, ideally using indicators beyond weight, such as mental health.

Summary

- This study analyzed combined physical and mental health changes over 5 years after obesity surgery, identifying three patient groups with low, medium, or high long-term improvement.
- Considering multiple health indicators provided better outcome prediction than weight alone.
- Low improvement was linked to a higher risk of eating disorders.
- Monitoring of multiple factors appears crucial to identifying at-risk patients needing extra care.
- Future research should explore these patterns over longer periods.

1 | Introduction

Obesity surgery (OS) is currently the most efficacious and sustainable intervention for severe obesity (i.e., obesity class 3 body mass index [BMI] ≥ 40 kg/m² or class 2 BMI ≥ 35 kg/m² with obesity-related comorbidities; National Institute for Health and Clinical Excellence 2006), an increasingly prevalent health disorder (Hales et al. 2018; Ward et al. 2019; Williamson et al. 2020). The commonly applied surgical procedures, laparoscopic Roux-en-Y gastric bypass (RYGB) and laparoscopic sleeve gastrectomy (SG), lead to a total weight loss of 20%–35% over 5–10 years of follow-up (O'Brien et al. 2019; van Rijswijk et al. 2021), with weight regain mostly beginning around 2 years after reaching nadir weight. In the long term, a significant minority of patients experience poor weight loss (Arterburn et al. 2020) and less improvement of the adverse physical obesity-related sequelae (e.g., type 2 diabetes mellitus; Adams et al. 2017; Courcoulas et al. 2018; Puzziferri et al. 2014).

Regarding mental health, OS produces substantial improvements in mental health, for example, in symptoms of depression (Dawes et al. 2016; Kalarchian et al. 2019) and eating disorders as well as health-related quality of life (HRQOL; Andersen et al. 2015; Kalarchian et al. 2019; Kolotkin and Andersen 2017); however, deteriorations after initial improvement have been documented (Devlin et al. 2018; Hilbert et al. 2022; Kalarchian et al. 2019). Weight loss and improvements in mental health seem to co-occur following OS (Hilbert et al. 2022; Nielsen et al. 2022), but their associations over time, prognostic relevance, and baseline correlates have yet to be explored (Hindle et al. 2017; Youssef et al. 2020). This is particularly relevant given the multifactorial nature of obesity presenting with physical and mental health-related sequelae, including functional impairment at different severity levels (Sharma and Kushner 2009).

While available long-term follow-up studies usually report outcomes following OS cross-sectionally by timepoints (Reis et al. 2023), initial studies have begun to explore post-surgical weight trajectories through a longitudinal perspective from pretreatment over time. Using univariate person-centered longitudinal latent structure analyses, prospective (Courcoulas et al. 2018; Slurink et al. 2024) and retrospective (Lent et al. 2018; Voorwinde et al. 2022) studies covering 3–7 years of follow-up mostly identified three trajectories with varying degrees of weight loss and regain (range: 3–7 trajectories). This evidence was limited by a high loss to follow-up and concentration on RYGB only, considering that weight trajectories following SG showed steeper weight regain (Shen et al. 2024). Only a few studies have addressed trajectories of mental health in addition to that of weight. For example, in their prospective linear growth mixture modeling of 3 years following RYGB ($N = 420$), Slurink et al. (2024) identified three different weight trajectories (high and moderate weight loss with weight regain, low weight loss without weight regain) and, separately, four or five trajectories of physical or mental HRQOL, respectively, similar to previous HRQOL modeling (Youssef et al. 2020). However, associations of HRQOL trajectories with weight trajectories were not analyzed. Using latent class growth mixture modeling on retrospective 5-year data following OS, mainly RYGB ($N = 2785$), Voorwinde et al. (2022) identified five weight trajectories (average weight loss/fairly stable, above average weight loss/partial regain, low response, rapid weight loss/regain, continued weight loss). Separately, they modeled eating behavior and physical activity, resulting in three trajectories per variable that showed partial associations with weight trajectories.

Thus, while heterogeneity in weight trajectories after OS is increasingly being elucidated, their relation to HRQOL and mental health trajectories and their prognostic relevance remain unclear. Mechanistically, sustained weight loss following OS may involve long-term improvements in HRQOL and mental health (Hilbert et al. 2022; Nielsen et al. 2022). In contrast, unresolved psychopathology at follow-up—such as eating disorder symptoms and, less consistently, depression—was associated with reduced long-term weight loss and weight regain (Devlin et al. 2018; Freire et al. 2021; Hilbert et al. 2022; Kalarchian et al. 2019), potentially by promoting increased energy intake or reduced physical activity. Postoperative eating disorder symptoms, depressive symptoms, and HRQOL have been found to be interrelated.

Overall, while changes in weight, HRQOL, and psychopathology often co-occur after OS, it is uncertain whether their trajectories align, highlighting the need for multivariate rather than the previously applied univariate trajectory analyses. It is further largely unclear whether specific multivariate change trajectories have a differential prognostic relevance for long-term physical and mental health outcomes and can

improve identification of individuals at risk compared to univariate weight trajectory analysis. Thus, this study sought to investigate change trajectories in weight, psychopathology, and HRQOL following OS using multivariate trajectory analysis, allowing the identification of trajectory classes, their sociodemographic and clinical baseline correlates, and their prospective association with long-term health outcomes. For sensitivity analysis, multivariate trajectory modeling was compared with univariate trajectory modeling of weight for concordance and prognostic significance.

2 | Methods

2.1 | Participants

This study is part of the ongoing prospective Psychosocial Registry for Obesity Surgery (PRAC) study (Hilbert et al. 2022) implemented at six surgical treatment centers in Germany upon approval by the local ethics committees and registration in the German Clinical Trials Register (DRKS00006749). Written informed consent was obtained from all patients prior to enrollment. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline was followed for this study (von Elm et al. 2007).

Inclusion criteria were age ≥ 18 years and planned OS; exclusion criteria were lack of German language skills and non-compliance with study procedures. Study-specific inclusion criteria consisted of the surgical procedures RYGB or SG; complete baseline data on the four indicator variables used for trajectory analysis; at least one follow-up assessment 1–5 years following surgery; and enrollment between 03/2012 and 01/2023. Assessment time points were at baseline (T0, pre-surgery) and 1–6 years (T2–T7) post-operatively (T1 at 6 months not reported in this study because depression was not assessed).

A total of 1144 adult volunteers were enrolled in PRAC, of which 217 were excluded (surgery not received: 196; no RYGB or SG: 11; dropout before baseline: 10) leaving 927 eligible patients providing baseline data on the four trajectory variables, of whom $N=856$ (92.34%) had at least one follow-up assessment. During follow-up, 17.29% (148/856; T2: 38; T3: 27; T4: 23; T5: 28; T6: 21; T7: 11) were lost due to withdrawal of consent (126/148) or death (22/148). From the total baseline sample of $N=856$ (70.86%), 567 patients received RYGB (66.24%), 289 patients received SG (33.76%), and 2938 follow-up assessments were available (T2: 754; T3: 629; T4: 533; T5: 438; T6: 329; T7: 255). Notably, in addition to study dropout (17.29%), further missingness resulted from the study's ongoing repeated-measures data collection.

2.2 | Measures for Multivariate Trajectory Analysis

Body weight and height were measured objectively using calibrated equipment (for details on the imputation of missing objective body weight data at follow-up using subjective body weight, see the Data Analytic Plan). In addition to weight, three well-established, validated measures were selected, covering

different aspects of mental health with clinical relevance, based on a multidimensional definition of health (World Health Organization 1946). Depressive symptoms were assessed over the last 2 weeks by the 9-item Patient Health Questionnaire (PHQ-D; Gräfe et al. 2004; Spitzer et al. 1999) (0 = *not at all*; 3 = *almost every day*), with higher sum scores (0–27) indicative of a more severe level of depression (Cronbach's $\alpha=0.85$, 95% CI 0.82–0.86; McDonald's ω total = 0.88, 95% CI 0.80–0.99). Eating disorder psychopathology was assessed using the Eating Disorder Examination-Questionnaire (EDE-Q; Fairburn and Beglin 2008; Hilbert and Tuschen-Caffier 2016), covering restraint, eating concern, weight concern, and shape concern with 22 items (0 = *characteristic was not present*; 6 = *characteristic was present every day/in extreme form*). A mean global score was derived, with higher scores indicating greater eating disorder psychopathology (Cronbach's $\alpha=0.87$, 95% CI 0.86–0.88; McDonald's ω total = 0.91, 95% CI 0.89–0.97). HRQOL was determined using the 31-item Impact of Weight on Quality of Life-Lite (IWQOL) questionnaire total sum score (Kolotkin et al. 2001; Mueller et al. 2011) (recoded as 0 = *worst* to 100 = *best*; Cronbach's $\alpha=0.95$, 95% CI 0.95–0.96; McDonald's ω total = 0.97, 95% CI 0.95–0.99). For correlations among the trajectory variables, see Table S1.

2.3 | Outcome Measures

Outcomes were determined at T7 and included the continuous measures used for the trajectory analysis. In addition, established indicators of clinically significant change were used. For weight loss and clinically significant weight loss, percentage total body weight loss from baseline (%TBWL) and presence versus absence of %TBWL ≥ 20 were determined. In addition, percentage alterable weight loss (%AWL = $100 \times (\text{baseline BMI} - \text{follow-up BMI}) / (\text{baseline BMI} - 13)$) and presence versus absence of clinically significant %AWL ≥ 35 were calculated as new measures to estimate outcome independent of baseline BMI (van de Laar et al. 2018). For depressive symptoms, a PHQ-D score < 10 versus ≥ 10 was used to determine absence versus presence of clinically significant, moderate depression (Gräfe et al. 2004). For eating disorder psychopathology, EDE-Q global scores $< 95^{\text{th}}$ percentile versus $\geq 95^{\text{th}}$ percentile of normative population means (Hilbert et al. 2012) were used as indicators for the absence versus presence of clinically significant eating disorder psychopathology. (For HRQOL, validated cut-offs of meaningful change of IWQOL scores in obesity surgery or norms were unavailable.) In addition, adverse surgery-related outcomes including bariatric reoperations and complications with the surgical procedure until T7, as well as improvement in obesity-related comorbidities between baseline and T7, were assessed interview-based using three items from the Bariatric Analysis and Reporting Outcome System (BAROS; Oria and Moorehead 2009) that were dichotomized (present, absent).

2.4 | Baseline Correlates

Sex (male, female), age (years), education (low < 12 years, high ≥ 12 years of school education), and surgical procedure (RYGB, SG) were examined as baseline correlates of the multivariate

trajectories; as were baseline scores of the continuous and dichotomous outcomes described above.

2.5 | Data Analytic Plan

To derive multivariate trajectory classes in a data-driven manner, latent class linear mixed models (LCMMs) were utilized to jointly model the longitudinal outcomes of weight, PHQ-D, EDE-Q, and IWQOL over the first 5 years after surgery. Weight was represented as relative weight loss to baseline over time (i.e., %TBWL), while PHQ-D, EDE-Q, and IWQOL scores were modeled as absolute differences (change scores) from baseline over time. If objective body weight was missing at follow-up but subjective body weight was available, the objective body weight was imputed using linear regression, estimating the objective weight based on the subjective body weight. For all other outcomes, we used all available information at the corresponding timepoints in the linear mixed models without imputation. This approach is robust to missing data under various mechanisms, including the Missing at Random (MAR) assumption, and is well-suited for prospective longitudinal designs (Twisk et al. 2013).

LCMMs were chosen to account for heterogeneity between patients by categorizing them into unobserved groups (latent trajectory classes) with distinct longitudinal class-specific trajectories. Outcome- and class-specific trajectories were modeled on all available data using flexible link functions based on splines, with random effects incorporated to adjust for repeated measurements and multiple outcomes. Model selection involved fitting models with varying numbers of latent classes and selecting the final number of classes based on minimizing the Akaike and Bayesian information criteria (AIC/BIC), maximizing entropy (i.e., class separation), and considering interpretation. Trajectory class membership of patients was determined based on posterior class probabilities. Mean and individual trajectories for all outcomes were graphically displayed, stratified by trajectory class.

To compare resulting classes from the LCMM regarding differences in baseline correlates and 6-year outcomes, odds ratios (ORs) with two-sided 95% confidence intervals (CIs) were calculated for categorical variables, and η^2 measures with two-sided 95% CIs were reported for continuous variables using all available data. Dunn's test was used for multiple post hoc comparisons. For sensitivity analysis, a univariate LCMM was computed solely for %TBWL, and the resulting classes were compared to those obtained from the LCMM for multiple outcomes. The comparison metric included predictive R^2 values derived from linear models for each outcome with class as the predictor variable, as well as Cohen's κ and interclass correlation. Furthermore, mean and individual trajectories stratified by trajectory class were graphically displayed for all participants with available data at the final timepoint (T7) as well as for those with complete data across all timepoints (complete cases).

All statistical tests were conducted using two-sided $\alpha=0.05$. Effect size interpretation was based on Cohen (small, medium, large: R^2 , 0.02, 0.13, 0.26; η^2 , 0.01, 0.06, 0.14; OR, 1.46, 2.50, 4.14;

Chen et al. 2010; Cohen 1988). Analyses were performed via the statistical programming environment R version 4.0.4 and the lcmm add-on package version 2.0.0.

3 | Results

The sample of $N=856$ patients in OS was predominantly middle-aged, female, with a low level of education, presented with class 3 obesity, and underwent surgery at the University Medical Center Leipzig (Table 1). Across follow-up, 13.32% (114/856) reported a bariatric reoperation.

3.1 | Multivariate Trajectory Groups

Data-driven LCMM on changes in body weight, depression, eating disorder psychopathology, and HRQOL from T0 to T6 identified three distinct multivariate trajectory classes of *low* (I), *medium* (II), and *high* (III) *sustainability* (Figure 1; Table S2), reflecting early relapse (I), gradual deterioration (II), or continued improvement (III) after initial post-surgical improvements from T0 to T2 (Table S3).

Descriptively, in the *low sustainability* class (I), 24 (2.80%) patients reached their nadir weight at T2, followed by a steep increase by T6. Similarly, their T2 postoperative improvement in depression, eating disorder psychopathology, and HRQOL deteriorated substantially by T6. The *medium sustainability* class (II) consisted of 763 (89.14%) patients who reached their nadir weight at T3 after a slight continued weight loss from T2 to T3, followed by a gradual weight regain from T2 to T6. Patients in this class maintained their T2 improvement in eating disorder psychopathology and HRQOL through T6, while experiencing a slight increase in depression. The *high sustainability* class (III) included 69 (8.06%) patients who reached their nadir weight at T3 after considerable sustained weight loss between T2 and T3. These patients showed the greatest improvements in all parameters at T2 and experienced further improvement in eating disorder psychopathology and HRQOL from T2 to T6, but a slight increase in depression during this time.

The multivariate trajectory classes identified in the total intent-to-treat sample showed a similar distribution and course in the subsample with available 6-year follow-up data ($N=225$) and in the complete-case subsample ($N=164$; Figure S1). Figures S2–S4 illustrate the individual trajectories within each multivariate trajectory class for the total sample and the respective subsamples.

3.2 | Baseline Correlates

Regarding baseline correlates, the trajectory classes did not differ significantly by sex, age, and education nor by weight or BMI (Table 2; mostly less than small to small effects). However, they differed in mental health (medium to large effects): Depressive symptoms and eating disorder psychopathology, and clinically significant levels thereof, were lower in the *low* and *medium sustainability* classes than in the *high sustainability* class, and HRQOL was lower in the

TABLE 1 | Sample characteristics.

	Total (N = 856)	Roux-en-Y gastric bypass (N = 567)	Sleeve gastrectomy (N = 289)
	No. (%)	No. (%)	No. (%)
Gender			
Female	587 (68.57%)	406 (71.60%)	181 (62.63%)
Male	269 (31.43%)	161 (28.40%)	108 (37.37%)
Age, years			
Median (IQR)	47.00 (37.00, 55.00)	48.00 (37.00, 55.00)	47.00 (36.00, 55.00)
Mean (SD)	46.39 (11.67)	46.61 (11.38)	45.97 (12.21)
Education ^a			
High	117 (16.98%)	74 (16.59%)	43 (17.70%)
Low	572 (83.02%)	372 (83.41%)	200 (82.30%)
Missing	167	121	46
Body weight, kg			
Median (IQR)	137.30 (122.40, 157.20)	133.80 (120.00, 149.20)	153.00 (128.30, 174.70)
Mean (SD)	141.56 (28.42)	135.60 (23.32)	153.80 (33.95)
Body mass index, kg/m ²			
Median (IQR)	47.52 (42.55, 52.81)	46.41 (42.17, 50.69)	50.61 (43.44, 57.52)
Mean (SD)	48.40 (8.10)	46.41 (6.33)	51.68 (9.97)
Weight status ^b			
Obesity class 1	11 (1.29%)	10 (1.77%)	1 (0.35%)
Obesity class 2	100 (11.68%)	66 (11.68%)	32 (11.07%)
Obesity class 3	745 (87.23%)	489 (86.55%)	256 (88.58%)
Treatment center			
Leipzig	683 (79.79%)	471 (83.07%)	212 (73.36%)
Other	173 (21.21%)	96 (16.93%)	77 (26.64%)

Note: Zero missing values for each variable, with the exception of education.

Abbreviation: IQR, interquartile range.

^aSchool education: high, ≥ 12 years or higher; low, < 12 years.

^bObesity class 1, body mass index 30.0–34.9 kg/m²; class 2, 35.0–39.9 kg/m²; class 3, ≥ 40.0 kg/m².

high sustainability class than in the *low sustainability* class, whereas the *medium sustainability* class had higher HRQOL than the other two classes.

3.3 | Clinical Outcomes at 6-Year Follow-Up

When prospectively relating trajectory classes derived from T0 to T6 to weight loss outcome at T7, %TBWL and %AWL and their dichotomized variants %TBWL $\geq 20\%$ and %AWL $\geq 35\%$ consistently differed by class and were significantly lower in the *low* and *medium sustainability* classes than in the *high sustainability* class, while weight and BMI were higher (Table 2; mostly medium to large effects).

Regarding psychological outcomes, trajectory classes differentially related to depressive symptoms and HRQOL (large

effects), but not to clinically significant, moderate depression (PHQ-D ≥ 10) at T7 (less than small to small effects); these outcomes were more favorable in the *high sustainability* class than in the *medium* and *low sustainability* classes (Table 2). Eating disorder psychopathology differed among all trajectory classes (large effect), with higher scores in those with *low* than *medium sustainability* and in those with *medium* than *high sustainability*. A clinically significant eating disorder psychopathology (EDE-Q $\geq 95^{\text{th}}$ percentile) at T7 was more likely in the *low* than *medium* and *high sustainability* classes (large or small effect, respectively), whereas the differences in the *medium* and *high sustainability* classes did not reach statistical significance (less than small effect). Trajectory classes did not differ significantly in patient-reported improvement of physical comorbidity from baseline through T7 (less than small effect). Descriptively, the proportion of patients reporting improvement was a quarter higher in the *high* than *low sustainability* class.

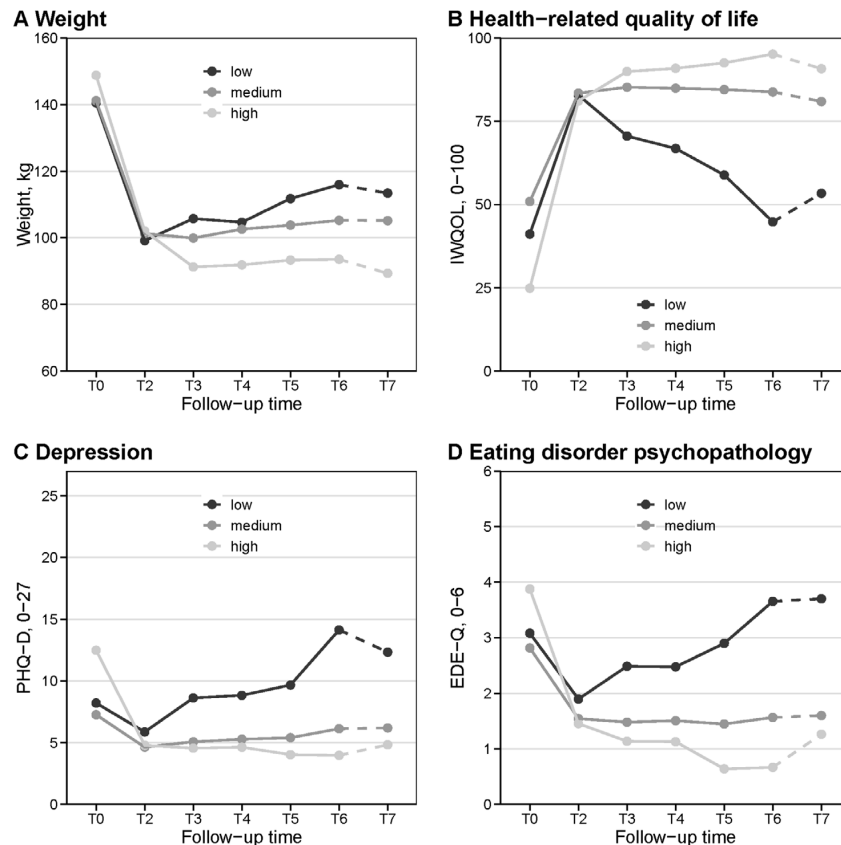


FIGURE 1 | Multivariate trajectory classes of low, medium, and high sustainability determined via latent class linear mixture modeling conducted from baseline to the first 5 years after obesity surgery (T0–T6) and applied to the 6-year follow-up (T7; $N=856$). Displayed are means from valid cases. EDE-Q, Eating Disorder Examination-Questionnaire; IWQOL, Impact of Weight on Quality of Life-Lite; PHQ-D, Patient Health Questionnaire-Depression.

3.4 | Surgical Correlates

Trajectory classes were unrelated to surgical procedure, bariatric reoperations, or complications following the initial operation up to T7 (Table 2; mostly less than small effects). When the 3 trajectory classes were applied separately to RYGB and SG patients, a similar distribution of patients in the *low*, *medium*, and *high sustainability* classes was found (RYGB: 14, 2.47%; 509, 89.77%; 44, 7.76%; SG: 10, 3.46%; 254, 87.89%; 25, 8.65%). Descriptively, the most notable difference occurred in the weight trajectory following SG (Figure S5): the *low sustainability* class had the lowest nadir weight, but steepest weight regain.

3.5 | Sensitivity Analysis: Univariate LCMM for Weight

A univariate LCMM for relative weight loss from baseline only (i.e., %TBWL) also identified three distinct trajectory classes from T0–T6, with a *low sustainability* class including 7 (0.84%) patients, a *medium sustainability* class including 826 (98.92%) patients, and a *high sustainability* class including 2 (0.24%)

patients. The univariate and the multivariate class solutions were neither concordant nor correlated in their classification of patients with values close to zero ($\kappa=0.02$, $z=0.93$, $p=0.35$; intraclass correlation coefficient = 0.04, $p=0.15$). The univariate model explained significantly less variance in the prediction of clinical outcomes at 6-year follow-up (T7) than the multivariate model in %TBWL (predictive R^2 : 0.16% vs. 7.64%), depressive symptoms (predictive R^2 : 2.32% vs. 18.42%), eating disorder psychopathology (predictive R^2 : 0.08% vs. 18.00%), and HRQOL (predictive R^2 : 0.65% vs. 25.54%; $p<0.001$). Thus, the univariate model mostly involved less than small predictive effects across outcomes including %TBWL, whereas the multivariate model mostly involved moderately sized predictive effects.

4 | Discussion

OS results in substantial improvements in weight and associated sequelae, but many patients face challenges with weight recidivism, relapse in psychopathology, and HRQOL impairment over the long term. The contributing factors are likely multifactorial, but not well understood, and may be mutually reinforcing. However, research on long-term outcomes after OS has

TABLE 2 | Baseline sociodemographic and clinical correlates and 6-year clinical outcomes of multivariate trajectories of low, medium, and high sustainability (N=856).

	Class 1: Low sustainability (n = 24)	Class 2: Medium sustainability (n = 763)	Class 3: High sustainability (n = 69)	
	No. (%) or mean (SD)	No. (%) or mean (SD)	No. (%) or mean (SD)	Effect size
Baseline correlates				
Sex, female	16 (66.67%)	519 (68.02%)	52 (75.36%)	1 vs. 2: OR = 1.06, 95% CI, 0.39 to 2.68 3 vs. 2: OR = 0.70, 95% CI, 0.37 to 1.25 1 vs. 3: OR = 0.66, 95% CI, 0.22 to 2.10
Sex, male	8 (33.33%)	244 (31.98%)	17 (24.64%)	
Age, years	44.33 (10.45)	46.39 (11.84)	47.07 (10.09)	$\eta^2 = 1.15\text{e-}03$, 95% CI, 0.00 to 0.01
Education, low	16 (76.19%)	501 (82.95%)	55 (85.94%)	1 vs. 2: OR = 1.52, 95% CI, 0.43 to 4.46 3 vs. 2: OR = 0.80, 95% CI, 0.34 to 1.69 1 vs. 3: OR = 0.53, 95% CI, 0.13 to 2.30
Education, high	5 (23.81%)	103 (17.05%)	9 (14.06%)	
Education total	21	604	64	
Education missing	3	159	5	
Body weight, kg	140.50 (25.46)	141.15 (28.40)	148.78 (32.09)	$\eta^2 = 5.30\text{e-}03$, 95% CI, 0.00 to 0.02
Body mass index, kg/m ²	47.81 (8.03)	41.14 (8.01)	51.45 (8.49)	$\eta^2 = 0.01$, 95% CI, 0.00 to 0.03
PHQ-D, 0–27	8.21 (5.30) ^a	7.26 (4.93) ^a	12.49 (4.70) ^b	$\eta^2 = 0.08$, 95% CI, 0.05 to 0.12
PHQ-D ≥ 10	8 (33.33%)	188 (27.49%)	51 (73.91%)	1 vs. 2: OR = 1.32, 95% CI, 0.48 to 3.33 3 vs. 2: OR = 7.45, 95% CI, 4.15 to 13.92 1 vs. 3: OR = 5.54, 95% CI, 1.87 to 17.77
PHQ-D total (N = 777)	24	684	69	
PHQ-D missing	0	79	0	
EDE-Q, 0–6	3.08 (0.96) ^a	2.82 (0.98) ^a	3.88 (0.84) ^b	$\eta^2 = 0.09$, 95% CI, 0.05 to 0.13
EDE-Q $\geq 95\%$	18 (75.00%)	469 (68.37%)	66 (95.65%)	1 vs. 2: OR = 1.39, 95% CI, 0.52 to 4.33 3 vs. 2: OR = 10.16, 95% CI, 3.27 to 51.10 1 vs. 3: OR = 7.13, 95% CI, 1.37 to 48.36
EDE-Q total (N = 779)	24	686	69	
EDE-Q missing	0	77	0	

(Continues)

TABLE 2 | (Continued)

	Class 1: Low sustainability (<i>n</i> = 24)	Class 2: Medium sustainability (<i>n</i> = 763)	Class 3: High sustainability (<i>n</i> = 69)	Effect size
	No. (%) or mean (SD)	No. (%) or mean (SD)	No. (%) or mean (SD)	
IWQOL, 0–100	41.12 (19.91) ^a	50.89 (21.14) ^b	24.87 (13.08) ^c	$\eta^2 = 0.12$, 95% CI, 0.08 to 0.16
IWQOL total (<i>N</i> = 778)	24	685	69	
IWQOL missing	0	78	0	
Surgical correlates				
Surgical procedure, RYGB	14 (58.33%)	509 (66.71%)	44 (63.77%)	1 vs. 2: OR = 1.43, 95% CI, 0.56 to 3.52 3 vs. 2: OR = 1.14, 95% CI, 0.65 to 1.95 1 vs. 3: OR = 0.80, 95% CI, 0.28 to 2.33
Surgical procedure, SG	10 (41.67%)	254 (33.29%)	25 (36.23%)	
Reoperations over 6 years	5 (20.83%)	99 (12.98%)	10 (14.49%)	1 vs. 2: OR = 2.37, 95% CI, 0.50 to 5.03 3 vs. 2: OR = 1.14, 95% CI, 0.50 to 2.34 1 vs. 3: OR = 0.65, 95% CI, 0.17 to 2.72
Complications over 6 years	6 (27.27%)	160 (32.27%)	16 (32.00%)	1 vs. 2: OR = 0.77, 95% CI, 0.24 to 2.13 3 vs. 2: OR = 0.97, 95% CI, 0.48 to 1.86 1 vs. 3: OR = 1.25, 95% CI, 0.37 to 4.66
Complications total (<i>N</i> = 561)	22	489	50	
Complications missing	2	274	19	
Clinical outcomes at 6 years				
Body weight, kg	113.42 (22.04) ^a	105.16 (23.06) ^a	89.34 (18.65) ^b	$\eta^2 = 0.05$, 95% CI, 0.01 to 0.10
%TBWL	21.03 (11.37) ^a	25.03 (10.90) ^a	35.40 (10.75) ^b	$\eta^2 = 0.08$, 95% CI, 0.02 to 0.14
%TBWL $\geq 20\%$	4 (40.00%)	149 (67.12%)	22 (95.65%)	1 vs. 2: OR = 0.33, 95% CI, 0.07 to 1.43 3 vs. 2: OR = 10.72, 95% CI, 1.67 to 449.96 1 vs. 3: OR = 28.06, 95% CI, 2.51 to 1564.35
Body mass index, kg/m ²	38.88 (6.85) ^a	35.89 (7.03) ^a	31.92 (6.40) ^a	$\eta^2 = 0.03$, 95% CI, 0.00 to 0.09
%AWL	28.49 (14.66) ^a	34.50 (14.93) ^a	48.26 (13.93) ^b	$\eta^2 = 0.07$, 95% CI, 0.02 to 0.14
%AWL $\geq 35\%$	2 (20.00%)	109 (49.10%)	19 (82.61%)	1 vs. 2: OR = 0.26, 95% CI, 0.03 to 1.35 3 vs. 2: OR = 4.90, 95% CI, 1.56 to 20.42 1 vs. 3: OR = 16.73, 95% CI 2.27 to 220.65

(Continues)

TABLE 2 | (Continued)

	Class 1: Low sustainability (<i>n</i> = 24)	Class 2: Medium sustainability (<i>n</i> = 763)	Class 3: High sustainability (<i>n</i> = 69)	Effect size
	No. (%) or mean (SD)	No. (%) or mean (SD)	No. (%) or mean (SD)	
Body weight total (<i>N</i> = 255)	10	222	23	
Body weight missing	14	541	46	
PHQ-D, 0–27	12.33 (6.42) ^a	6.19 (5.53) ^a	4.83 (3.45) ^b	$\eta^2 = 0.18$, 95% CI, 0.10 to 0.27
PHQ-D ≥ 10	4 (44.44%)	48 (25.13%)	2 (8.70%)	1 vs. 2: OR = 2.37, 95% CI, 0.45 to 11.52 3 vs. 2: OR = 0.28, 95% CI, 0.03 to 1.24 1 vs. 3: OR = 0.13, 95% CI, 0.01 to 1.19
PHQ-D total (<i>N</i> = 223)	9	191	23	
PHQ-D missing	15	572	46	
EDE-Q, 0–6	3.70 (1.58) ^a	1.60 (1.26) ^b	1.26 (0.96) ^c	$\eta^2 = 0.18$, 95% CI, 0.09 to 0.27
EDE-Q $\geq 95\%$	7 (77.78%)	59 (30.57%)	3 (13.04%)	1 vs. 2: OR = 7.86, 95% CI, 1.55 to 79.71 3 vs. 2: OR = 0.34, 95% CI, 0.06 to 1.22 1 vs. 3: OR = 0.05, 95% CI, 0 to 0.41
EDE-Q total (<i>N</i> = 225)	9	193	23	
EDE-Q missing	15	570	46	
IWQOL, 0–100	53.36 (25.85) ^a	80.92 (19.93) ^a	90.76 (13.33) ^b	$\eta^2 = 0.26$, 95% CI, 0.16 to 0.34
IWQOL total (<i>N</i> = 224)	9	192	23	
IWQOL missing	15	571	46	
Somatic comorbidity, improved	6 (75.00%)	170 (85.43%)	19 (100%)	1 vs. 2: OR = 0.51, 95% CI, 0.09 to 5.44
Somatic comorbidity total (<i>N</i> = 226)	8	199	19	
Somatic comorbidity missing	16	564	50	

Note: Zero missing values for sex, age, surgical procedure, and reoperations.

Abbreviations: %AWL, percentage alterable weight loss; %TBWL, percentage total body weight loss; EDE-Q, Eating Disorder Examination-Questionnaire; IWQOL, Impact of Weight on Quality of Life-Lite; PHQ-D, Patient Health Questionnaire-Depression; RYGB, Roux-en-Y gastric bypass; SG, sleeve gastrectomy.

p < 0.05; The superscript letters a, b, c indicate significant Dunn's posthoc tests.

traditionally focused on single variables, particularly weight, examined cross-sectionally as separate snapshots at single time points or, more recently, in emerging weight trajectory analyses. An advantage of the multivariate trajectory modeling used in this study is that longitudinal data on multiple clinically relevant variables were considered jointly to generate individual trajectories grouped into latent trajectory classes, allowing an understanding of how different outcomes change postsurgically in relation to each other. Importantly, the identified multivariate trajectory class solution derived over 5 years of follow-up was found to be superior in predicting clinical outcomes, including weight at 6 years of follow-up, compared to discordant univariate weight trajectory modeling. It was unrelated to baseline weight and inversely related to baseline mental health.

Using multivariate latent class linear mixed modeling, we identified three latent trajectory classes, similar to previous research on univariate weight trajectories (Lent et al. 2018; Shen et al. 2024; Slurink et al. 2024). Descriptively, patients in all trajectory classes achieved similar levels of improvement at 1 year after OS, but exhibited specific trajectories thereafter, showing: *low sustainability* (I), including weight regain, relapse beyond baseline levels of depression and eating disorder psychopathology, and HRQOL impairment; *medium sustainability* (II), with slight weight regain and maintenance or slight deterioration in psychopathology and HRQOL; and *high sustainability* (III), with slight weight regain, but continued improvement in psychopathology and HRQOL. Notably, individual trajectories within the medium sustainability class showed within-class variability (see Figure S2), which is common in latent trajectory modeling and reflects the clinical diversity typically observed in large patient groups, thus warranting caution in interpreting mean trends. Taken together, the results suggest that different indicators do not change in parallel, but show distinctive courses, with most similarity of change in the nonindependent psychopathological and social impairment variables (Table S1).

Consistent with previous weight trajectory analyses spanning 5–7 years following RYGB, the vast majority (~90%) of patients showed moderate long-term weight loss and some degree of weight regain (Courcoulas et al. 2018; Voorwinde et al. 2022). Although patients were assigned to multiple classes rather than one class with *medium sustainability*, this pattern lends support to the validity of our multivariate trajectory class distribution. In addition, these previous studies also identified two smaller trajectory classes, comprising approximately 10% of patients, characterized by a low weight loss response or sustained high weight loss, resembling our *low* and *high sustainability* classes, respectively.

Further consistent with most previous weight trajectory analyses (Courcoulas et al. 2018; Lent et al. 2018; Shen et al. 2024; Voorwinde et al. 2022), a later nadir weight was associated with greater long-term weight loss. While patients with a *low sustainability* trajectory reached the nadir of all indicators at 1-year follow-up, patients with a *medium sustainability* trajectory reached the nadir of all indicators at 2-year follow-up. In contrast, patients with a *high sustainability* trajectory reached their weight nadir at 2-year follow-up, while their psychopathology and HRQOL continued to improve throughout follow-up. Ultimately, the trajectories resulted in low, medium, and high

improvements at 5-year follow-up, in line with weight outcomes in univariate weight trajectory modeling (Courcoulas et al. 2018; Lent et al. 2018; Shen et al. 2024; Voorwinde et al. 2022), suggesting that the timepoint of maximal improvement of multiple indicators is critical to identify the trajectory class within which a patient belongs. A comparison of the trajectory classes by surgical procedures supported that the weight at nadir may be less critical for determining long-term success than the timing; in fact, after SG, the *low sustainability* class achieved the lowest nadir weight of all trajectory groups at 1-year follow-up, followed by the steepest weight regain (Shen et al. 2024). Notable also was that the *high sustainability* class reached the largest improvements at 5 years across surgical procedures, although these patients had displayed worst baseline levels in indicators of mental health, also found in some (Courcoulas et al. 2018; Lent et al. 2018) but not other (Shen et al. 2024; Voorwinde et al. 2022) univariate weight trajectory modeling studies.

Predictor analyses supported these descriptive results at 6-year follow-up, suggesting prognostic relevance of the trajectory classes for long-term weight, BMI, %TBWL, %AWL, depression, eating disorder psychopathology, and HRQOL with mostly medium to large effect sizes. Furthermore, they highlighted differences in clinically significant improvement, for example, of %TBWL $\geq 20\%$ or %AWL $\geq 35\%$ (van de Laar et al. 2018), and eating disorder psychopathology below clinical cut-offs at 6 years following OS, especially when comparing the *high sustainability* class to the other classes with up to large effect sizes. Although most outcomes descriptively showed highest impairment in the *low sustainability* trajectory ($n = 24$), this class did not differ statistically from the *medium sustainability* trajectory, except for (clinically significant) eating disorder psychopathology with large effect sizes. Thus, the *low sustainability* class was especially characterized by elevated eating disorder psychopathology at 6-year follow-up. These results suggest that the multivariate trajectory modeling in our study has prognostic significance, which is particularly relevant in light of the inconsistent evidence on long-term outcome prediction after OS (El Ansari and Elhag 2021), although the small size of the *low* and *high sustainability* classes underscores the need for replication of the multivariate trajectory modeling in independent samples. In addition, the stability of the findings and predictive validity should be tested over a longer time period. The increase in eating disorder psychopathology over follow-up in the *low sustainability* class permits speculation that the majority of patients in this class (re)developed an eating disorder syndrome after OS, which may have impaired long-term weight maintenance (Devlin et al. 2018; Hilbert et al. 2022).

Regarding predictors of trajectory classes, although baseline sociodemographic correlates and surgical procedures were not prospectively associated with trajectory class membership, several patterns warrant note: Descriptively, the *low sustainability* class included the highest proportion of individuals with high education and patients receiving SG, whereas the *high sustainability* class included the highest proportion of female patients and individuals with low education, aspects previously associated with unfavorable versus favorable weight trajectories (Courcoulas et al. 2018; Keith Jr et al. 2018; Lent et al. 2018; Shen et al. 2024). Trajectory classes were unrelated to baseline weight or BMI but showed a medium-sized inverse relationship

with psychopathological and social impairments, with the highest baseline impairments in the *high sustainability* class, which included the largest proportion of patients with clinically significant depression and eating disorder psychopathology. Thus, OS may have yielded the greatest relief from impairments alongside the most substantial and sustained weight loss, although causal inference requires experimental designs with control for confounding. This effect may partly reflect the exclusion of patients with unstable psychological conditions (e.g., severe mental disorders) from OS, in line with evidence-based obesity treatment guidelines (Deutsche Adipositas Gesellschaft 2024).

Further, the divergence in psychological symptom trajectories despite similar weight regain between the *low* and *medium sustainability* classes suggests that factors beyond weight regain have contributed to increased depressive and disordered eating symptoms over time. Potential contributors include psychosocial stressors (e.g., postsurgical lifestyle adjustment difficulties), maladaptive coping (e.g., emotional eating), or adverse life events (e.g., separation). Moreover, preexisting vulnerabilities (e.g., low self-esteem) may not have been fully captured at baseline. Future research should investigate diverse psychosocial factors over time to elucidate postsurgical trajectories. Of note, trajectory classes did not differ significantly by patient-reported complications following surgery and revisional surgery. Previous research had found an association of revisional surgery with weight outcomes, but not with mental health outcomes (Courcoulas et al. 2018; Eisenberg et al. 2023; Leung et al. 2023).

Regarding strengths and limitations, this study was based on the multicenter PRAC study's large prospective cohort of adults with severe obesity undergoing RYGB and SG as standard surgical procedures (Angrisani et al. 2021). For the first time, a multivariate instead of univariate trajectory analysis approach was used, with trajectory variables being selected based on their relevance for physical, psychopathological, and social aspects of health in persons with severe obesity mapping onto the World Health Organization's multidimensional health definition (World Health Organization 1946). We used data from baseline through the first 5 years after OS to derive a trajectory class solution with long-term prognostic significance. Of note, from baseline through 2–4-year follow-up, three-class solutions were found, but with different composition and lower stability (data not shown), underscoring the importance of examining multi-year postoperative trajectories to capture patterns of weight regain beyond initial weight loss (Pyykkö et al. 2025). Minimal selection, information, and confounding bias, and a high degree of generalizability to OS populations were ensured by minimal inclusion and exclusion criteria and well-established assessments.

Attrition was relatively low (17.29%) and within the range commonly reported in psychosocial follow-up studies with similar designs (10%–35% for RYGB; Devlin et al. 2018; Lent et al. 2018). Missing data were accounted for by longitudinal mixed-effects modeling, which is robust to missing data, thereby reducing a potential attrition bias. Notwithstanding, the decrease in the analyzed sample from 856 to 329 at 5 years and 255 at 6 years likely contributed to the graphical variations of the trajectories from T6 to T7 (Figure 1) and made split-sample LCMM analyses for SG and RYGB separately infeasible; therefore, the total sample LCMM was applied to the SG and RYGB subsamples

(Figure S5). Trajectory classes in the 6-year follow-up and complete-case subsamples closely resembled those in the total sample, supporting the robustness of our findings despite missing data. Regarding further limitations, ethnicity was not assessed due to cultural norms in Germany. Finally, it should be noted that instruments such as the PHQ-D, while widely used in obesity research, were not developed or normed specifically for bariatric populations. This may lead to misinterpretation of obesity-related symptoms, particularly those concerning sleep, eating behavior, or physical activity.

Future research should examine multivariate trajectories and their prognostic significance over a longer-term follow-up period with greater trajectory class sizes. Given the observed heterogeneity in individual courses within trajectory classes, especially the medium sustainability class, future studies may benefit from analytic approaches that capture within-class variability and intraindividual dynamics more explicitly. Other indicators with potential clinical relevance may be considered to further our understanding of postsurgical change across multiple symptoms (e.g., self-regulation, eating behavior, physical activity, hormonal changes; Schäfer et al. 2019). Although LCMM has limited value for identification of individuals at risk in clinical practice, our results support that patients who achieve nadir improvements in weight and mental health early may require clinical attention to prevent long-term relapse and aggravation. A more fine-grained analysis of change during the honeymoon phase of initial weight loss following OS may add specificity to this study's results. Clinically, monitoring change during the first years after OS appears essential to identify those in need of additional intervention, ideally using multiple established indicators, including those of mental health, as consideration of weight alone revealed inferior prognostic significance.

Author Contributions

Anja Hilbert: conceptualization, funding acquisition, writing – original draft, writing – review and editing. **Annika Strömer:** conceptualization, formal analysis, writing – original draft, writing – review and editing. **Christian Staerk:** conceptualization, formal analysis, writing – original draft, writing – review and editing. **Ben Schreglmann:** investigation, writing – review and editing. **Thomas Mansfeld:** investigation, writing – review and editing. **Johannes Sander:** investigation, writing – review and editing. **Florian Seyfried:** investigation, writing – review and editing. **Stefan Kaiser:** investigation, writing – review and editing. **Christine Stroh:** investigation, writing – review and editing. **Arne Dietrich:** investigation, writing – review and editing. **Ricarda Schmidt:** investigation, writing – original draft, writing – review and editing. **Andreas Mayr:** conceptualization, formal analysis, writing – original draft, writing – review and editing.

Acknowledgments

We thank all coworkers for their help with the conduct of this study and all patients for participation. Jamie L. Manwaring, PhD, ACUTE Center for Eating Disorders and Severe Malnutrition, Denver, Colorado, and Stephanie Günther assisted in editing the manuscript and were compensated for this work. Generative Artificial Intelligence tools based on large language models were used for English language correction and editing. Open Access funding enabled and organized by Projekt DEAL.

Ethics Statement

Ethical approval was granted by the Ethics Committee of Leipzig University Medical Center (Ref. No. 356-11), based on which approval was granted by site-specific Institutional Review Boards. Informed

written consent was obtained at the outset of the study. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Conflicts of Interest

Dr. Hilbert reports receiving research grants from the German Federal Ministry of Education and Research, German Research Foundation, Innovation Fund, and Roland Ernst Foundation for Health Care; royalties for books on the treatment of eating disorders and obesity with Hogrefe and Kohlhammer; honoraria for workshops and lectures on eating disorders and obesity and their treatment, including from Lilly and Novo Nordisk; honoraria as editor of the *International Journal of Eating Disorders*; and honoraria as a consultant for Takeda.

Data Availability Statement

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

References

- Adams, T. D., L. E. Davidson, S. E. Litwin, et al. 2017. "Weight and Metabolic Outcomes 12 Years After Gastric Bypass." *New England Journal of Medicine* 377: 1143–1155. <https://doi.org/10.1056/NEJMoa1700459>.
- Andersen, J. R., A. Aasprang, T. I. Karlsen, G. K. Natvig, V. Våge, and R. L. Kolotkin. 2015. "Health-Related Quality of Life After Bariatric Surgery: A Systematic Review of Prospective Long-Term Studies." *Surgery for Obesity and Related Diseases* 11: 466–473. <https://doi.org/10.1016/j.soard.2014.10.027>.
- Angrisani, L., A. Santonicola, P. Iovino, A. Ramos, S. Shikora, and L. Kow. 2021. "Bariatric Surgery Survey 2018: Similarities and Disparities Among the 5 IFSO Chapters." *Obesity Surgery* 31: 1937–1948. <https://doi.org/10.1007/s11695-020-05207-7>.
- Arterburn, D. E., D. A. Telem, R. F. Kushner, and A. P. Courcoulas. 2020. "Benefits and Risks of Bariatric Surgery in Adults: A Review." *JAMA* 324: 879–887. <https://doi.org/10.1001/jama.2020.12567>.
- Chen, H., P. Cohen, and S. Chen. 2010. "How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies." *Communications in Statistics: Simulation and Computation* 39: 860–864. <https://doi.org/10.1080/03610911003650383>.
- Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Routledge.
- Courcoulas, A. P., W. C. King, S. H. Belle, et al. 2018. "Seven-Year Weight Trajectories and Health Outcomes in the Longitudinal Assessment of Bariatric Surgery (LABS) Study." *JAMA Surgery* 153: 427–434. <https://doi.org/10.1001/jamasurg.2017.5025>.
- Dawes, A. J., M. Maggard-Gibbons, A. R. Maher, et al. 2016. "Mental Health Conditions Among Patients Seeking and Undergoing Bariatric Surgery: A Meta-Analysis." *JAMA* 315: 150–163. <https://doi.org/10.1001/jama.2015.18118>.
- Deutsche Adipositas Gesellschaft. 2024. "S3-Leitlinie Prävention und Therapie der Adipositas (Version 5.0)". <https://register.awmf.org/de/leitlinien/detail/050-001>.
- Devlin, M. J., W. C. King, M. A. Kalarchian, et al. 2018. "Eating Pathology and Associations With Long-Term Changes in Weight and Quality of Life in the Longitudinal Assessment of Bariatric Surgery Study." *International Journal of Eating Disorders* 51: 1322–1330. <https://doi.org/10.1002/eat.22979>.
- Eisenberg, D., S. A. Shikora, E. Aarts, et al. 2023. "2022 American Society of Metabolic and Bariatric Surgery (ASMBS) and International Federation for the Surgery of Obesity and Metabolic Disorders (IFSO) Indications for Metabolic and Bariatric Surgery." *Obesity Surgery* 33: 3–14. <https://doi.org/10.1007/s11695-022-06332-1>.
- El Ansari, W., and W. Elhag. 2021. "Weight Regain and Insufficient Weight Loss After Bariatric Surgery: Definitions, Prevalence, Mechanisms, Predictors, Prevention and Management Strategies, and Knowledge Gaps—A Scoping Review." *Obesity Surgery* 31: 1755–1766. <https://doi.org/10.1007/s11695-020-05160-5>.
- Fairburn, C. G., and S. J. Beglin. 2008. "Eating Disorder Examination-Questionnaire (Edition 6.0), 2008". <https://insideoutinstitute.org.au/assets/ede-q-eating-disorder-examination-questionnaire-subsc ales.pdf>.
- Freire, C. C., M. T. Zanella, A. Segal, C. H. Arasaki, M. I. R. Matos, and G. Carneiro. 2021. "Associations Between Binge Eating, Depressive Symptoms and Anxiety and Weight Regain After Roux-en-Y Gastric Bypass Surgery." *Eating and Weight Disorders* 26: 191–199. <https://doi.org/10.1007/s40519-019-00839-w>.
- Gräfe, K., S. Zipfel, W. Herzog, and B. Löwe. 2004. "Screening psychischer Störungen mit dem 'Gesundheitsfragebogen für Patienten (PHQ-D)'." *Diagnostica* 50: 171–181. <https://doi.org/10.1026/0012-1924.50.4.171>.
- Hales, C. M., C. D. Fryar, M. D. Carroll, D. S. Freedman, and C. L. Ogden. 2018. "Trends in Obesity and Severe Obesity Prevalence in US Youth and Adults by Sex and Age, 2007–2008 to 2015–2016." *JAMA* 319: 1723–1725. <https://doi.org/10.1001/jama.2018.3060>.
- Hilbert, A., M. de Zwaan, and E. Braehler. 2012. "How Frequent Are Eating Disturbances in the Population? Norms of the Eating Disorder Examination-Questionnaire." *PLoS One* 7: e29125. <https://doi.org/10.1371/journal.pone.0029125>.
- Hilbert, A., C. Staerk, A. Strömer, et al. 2022. "Nonnormative Eating Behaviors and Eating Disorders and Their Associations With Weight Loss and Quality of Life During 6 Years Following Obesity Surgery." *JAMA Network Open* 5: e2226244. <https://doi.org/10.1001/jamanetworkopen.2022.26244>.
- Hilbert, A., and B. Tuschen-Caffier. 2016. "Eating Disorder Examination, 2. Auflage. Tübingen, dgvt-Verlag, 2016".
- Hindle, A., X. de la Piedad Garcia, and L. Brennan. 2017. "Early Post-Operative Psychosocial and Weight Predictors of Later Outcome in Bariatric Surgery: A Systematic Literature Review." *Obesity Reviews* 18: 317–334. <https://doi.org/10.1111/obr.12496>.
- Kalarchian, M. A., W. C. King, M. J. Devlin, et al. 2019. "Mental Disorders and Weight Change in a Prospective Study of Bariatric Surgery Patients: 7 Years of Follow-Up." *Surgery for Obesity and Related Diseases* 15: 739–748. <https://doi.org/10.1016/j.soard.2019.01.008>.
- Keith, C. J., Jr., A. A. Gullick, K. Feng, J. Richman, R. Stahl, and J. Grams. 2018. "Predictive Factors of Weight Regain Following Laparoscopic Roux-en-Y Gastric Bypass." *Surgical Endoscopy* 32: 2232–2238. <https://doi.org/10.1007/s00464-017-5913-2>.
- Kolotkin, R. L., and J. R. Andersen. 2017. "A Systematic Review of Reviews: Exploring the Relationship Between Obesity, Weight Loss and Health-Related Quality of Life." *Clinical Obesity* 7: 273–289. <https://doi.org/10.1111/cob.12203>.
- Kolotkin, R. L., R. D. Crosby, K. D. Kosloski, and G. R. Williams. 2001. "Development of a Brief Measure to Assess Quality of Life in Obesity." *Obesity Research* 9: 102–111. <https://doi.org/10.1038/oby.2001.13>.
- Lent, M. R., Y. Hu, P. N. Benotti, et al. 2018. "Demographic, Clinical, and Behavioral Determinants of 7-Year Weight Change Trajectories in Roux-en-Y Gastric Bypass Patients." *Surgery for Obesity and Related Diseases* 14: 1680–1685. <https://doi.org/10.1016/j.soard.2018.07.023>.
- Leung, S. E., V. Daliri, S. E. Cassin, R. Hawa, and S. Sockalingam. 2023. "Mental Health Outcomes in Revisional Versus Primary Bariatric Surgery Patients: A Matched Case Control Study." *Journal*

- of *Psychosomatic Research* 170: 111335. <https://doi.org/10.1016/j.jpsychores.2023.111335>.
- Mueller, A., C. Holzapfel, H. Hauner, et al. 2011. "Psychometric Evaluation of the German Version of the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) Questionnaire." *Experimental and Clinical Endocrinology & Diabetes* 119: 69–74. <https://doi.org/10.1055/s-0030-1261922>.
- National Institute for Health and Clinical Excellence (UK). 2006. *Obesity: The prevention, Identification, Assessment and Management of Overweight and Obesity in Adults and Children*. National Institute for Health and Clinical Excellence (UK).
- Nielsen, H. J., B. G. Nedrebø, A. Fosså, et al. 2022. "Seven-Year Trajectories of Body Weight, Quality of Life and Comorbidities Following Roux-en-Y Gastric Bypass and Sleeve Gastrectomy." *International Journal of Obesity* 46: 739–749. <https://doi.org/10.1038/s41366-021-01028-5>.
- O'Brien, P. E., A. Hindle, L. Brennan, et al. 2019. "Long-Term Outcomes After Bariatric Surgery: A Systematic Review and Meta-Analysis of Weight Loss at 10 or More Years for all Bariatric Procedures and a Single-Centre Review of 20-Year Outcomes After Adjustable Gastric Banding." *Obesity Surgery* 29: 3–14. <https://doi.org/10.1007/s11695-018-3525-0>.
- Oria, H. E., and M. K. Moorehead. 2009. "Updated Bariatric Analysis and Reporting Outcome System (BAROS)." *Surgery for Obesity and Related Diseases* 5: 60–66. <https://doi.org/10.1016/j.soard.2008.10.004>.
- Puzziferri, N., T. B. Roshek 3rd, H. G. Mayo, R. Gallagher, S. H. Belle, and E. H. Livingston. 2014. "Long-Term Follow-Up After Bariatric Surgery: A Systematic Review." *JAMA* 312: 934–942. <https://doi.org/10.1001/jama.2014.10706>.
- Pyykkö, J. E., N. van Olst, V. E. A. Gerdes, et al. 2025. "Relations Between Trajectories of Weight Loss and Changes in Psychological Health Over a Period of 2 Years Following Bariatric Metabolic Surgery." *Quality of Life Research* 34: 1345–1361. <https://doi.org/10.1007/s11136-025-03906-1>.
- Reis, M. G., L. F. G. G. Moreira, L. S. V. de Andrade Carvalho, C. T. de Castro, R. A. L. Vieira, and N. S. Guimarães. 2023. "Weight Regain After Bariatric Surgery: A Systematic Review and Meta-Analysis of Observational Studies." *Obesity Medicine* 100528: 100528. <https://doi.org/10.1016/j.obmed.2023.100528>.
- Schäfer, L., C. Hübner, T. Carus, et al. 2019. "Pre- and Postbariatric Subtypes and Their Predictive Value for Health-Related Outcomes Measured 3 Years After Surgery." *Obesity Surgery* 29: 230–238. <https://doi.org/10.1007/s11695-018-3524-1>.
- Sharma, A. M., and R. F. Kushner. 2009. "A Proposed Clinical Staging System for Obesity." *International Journal of Obesity* 33: 289–295. <https://doi.org/10.1038/ijo.2009.2>.
- Shen, E., A. Baecker, M. Ji, et al. 2024. "Pre-Surgical Factors Related to Latent Trajectories of 5-Year Weight Loss for a Diverse Bariatric Surgery Population." *Surgery for Obesity and Related Diseases* 20: 621–633. <https://doi.org/10.1016/j.soard.2024.01.016>.
- Slurink, I. A. L., I. Nyklíček, R. Kint, et al. 2024. "Longitudinal Trajectories and Psychological Predictors of Weight Loss and Quality of Life Until 3 Years After Metabolic and Bariatric Surgery." *Journal of Psychosomatic Research* 178: 111590. <https://doi.org/10.1016/j.jpsychores.2024.111590>.
- Spitzer, R. L., K. Kroenke, and J. B. Williams. 1999. "Validation and Utility of a Self-Report Version of PRIME-MD: The PHQ Primary Care Study." *JAMA* 282: 1737–1744. <https://doi.org/10.1001/jama.282.18.1737>.
- Twisk, J., M. de Boer, W. de Vente, and M. Heymans. 2013. "Multiple Imputation of Missing Values Was Not Necessary Before Performing a Longitudinal Mixed-Model Analysis." *Journal of Clinical Epidemiology* 66: 1022–1028. <https://doi.org/10.1016/j.jclinepi.2013.03.017>.
- van de Laar, A. W., A. S. van Rijswijk, H. Kakar, and S. C. Bruin. 2018. "Sensitivity and Specificity of 50% Excess Weight Loss (50%EWL) and Twelve Other Bariatric Criteria for Weight Loss Success." *Obesity Surgery* 28: 2297–2304. <https://doi.org/10.1007/s11695-018-3173-4>.
- van Rijswijk, A. S., N. van Olst, W. Schats, D. L. van der Peet, and A. W. van de Laar. 2021. "What Is Weight Loss After Bariatric Surgery Expressed in Percentage Total Weight Loss (%TWL)? A Systematic Review." *Obesity Surgery* 31: 3833–3847. <https://doi.org/10.1007/s11695-021-05394-x>.
- von Elm, E., D. G. Altman, M. Egger, et al. 2007. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *Lancet* 370: 1453–1457. [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X).
- Voorwinde, V., T. Hoekstra, V. M. Monpellier, I. H. M. Steenhuis, I. M. C. Janssen, and M. M. van Stralen. 2022. "Five-Year Weight Loss, Physical Activity, and Eating Style Trajectories After Bariatric Surgery." *Surgery for Obesity and Related Diseases* 18: 911–918. <https://doi.org/10.1016/j.soard.2022.03.020>.
- Ward, Z. J., S. N. Bleich, A. L. Craddock, et al. 2019. "Projected U.S. State-Level Prevalence of Adult Obesity and Severe Obesity." *New England Journal of Medicine* 381: 2440–2450. <https://doi.org/10.1056/NEJMs a1909301>.
- Williamson, K., A. Nimegeer, and M. Lean. 2020. "Rising Prevalence of BMI ≥ 40 kg/m²: A High-Demand Epidemic Needing Better Documentation." *Obesity Reviews* 21: e12986. <https://doi.org/10.1111/obr.12986>.
- World Health Organization (WHO). 1946. "Constitution of the World Health Organization. Geneva". <https://apps.who.int/gb/gov/assets/constitution-en.pdf>.
- Youssef, A., C. Keown-Stoneman, R. Maunder, et al. 2020. "Differences in Physical and Mental Health-Related Quality of Life Outcomes 3 Years After Bariatric Surgery: A Group-Based Trajectory Analysis." *Surgery for Obesity and Related Diseases* 16: 1837–1849. <https://doi.org/10.1016/j.soard.2020.06.014>.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.

3.2 Publication B: Boosting multivariate structured additive distributional regression models

Strömer A, Klein N, Staerk C, Klinkhammer H, Mayr A. Boosting multivariate structured additive distributional regression models. *Statistics in Medicine* 2023; 42(11): 1779-1801.

<https://doi.org/10.1002/sim.9699>

Supplementary information can be found at:

<https://doi.org/10.1002/sim.9699>

Implementations are available on GitHub:

<https://github.com/AnnikaStr/DistRegBoost>

RESEARCH ARTICLE

Statistics
in Medicine WILEY

Boosting multivariate structured additive distributional regression models

Annika Strömer¹ | Nadja Klein² | Christian Staerk¹ |Hannah Klinkhammer^{1,3} | Andreas Mayr¹¹Department of Medical Biometrics, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany²Chair of Uncertainty Quantification and Statistical Learning, Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund), Dortmund, Germany³Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn, Germany**Correspondence**Annika Strömer, Department of Medical Biometrics, Informatics and Epidemiology, University Hospital Bonn, 53127 Bonn, Germany.
Email: stroemer@imbie.uni-bonn.de**Funding information**

Deutsche Forschungsgemeinschaft, Grant/Award Numbers: Grant/Award Numbers: 428239776, KL3037/2-1, MA7304/1-1

We develop a model-based boosting approach for multivariate distributional regression within the framework of generalized additive models for location, scale, and shape. Our approach enables the simultaneous modeling of all distribution parameters of an arbitrary parametric distribution of a multivariate response conditional on explanatory variables, while being applicable to potentially high-dimensional data. Moreover, the boosting algorithm incorporates data-driven variable selection, taking various different types of effects into account. As a special merit of our approach, it allows for modeling the association between multiple continuous or discrete outcomes through the relevant covariates. After a detailed simulation study investigating estimation and prediction performance, we demonstrate the full flexibility of our approach in three diverse biomedical applications. The first is based on high-dimensional genomic cohort data from the UK Biobank, considering a bivariate binary response (chronic ischemic heart disease and high cholesterol). Here, we are able to identify genetic variants that are informative for the association between cholesterol and heart disease. The second application considers the demand for health care in Australia with the number of consultations and the number of prescribed medications as a bivariate count response. The third application analyses two dimensions of childhood undernutrition in Nigeria as a bivariate response and we find that the correlation between the two undernutrition scores is considerably different depending on the child's age and the region the child lives in.

KEYWORDS

generalized additive models for location, scale and shape, model-based boosting, multivariate Gaussian distribution, multivariate logit model, multivariate Poisson distribution, semiparametric regression

1 | INTRODUCTION

Many modern regression models relate certain characteristics of a univariate response distribution to explanatory variables. Examples include generalized additive models (GAMs)^{1,2} and quantile regression models,³ where with the former

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

the conditional expectation and with the latter conditional quantiles of a univariate response distribution are modeled by an additive decomposition of different covariate effects. In biomedical research there are often multiple endpoints that are typically analyzed separately by univariate regression models where each endpoint serves once as response variable.⁴ However, in practice, the components of a multivariate response are often not (conditionally) independent, so that separate models might induce a loss of information and could even lead to potentially misleading conclusions.

A well-known approach for the analysis of multivariate responses, particularly common in the economics literature, is called seemingly unrelated regression.⁵ This classical approach is restricted to linear predictors and constant covariance matrices not depending on the covariates; however, extensions to semiparametric predictors for the marginal means exist.⁶ Beyond that, multiple discrete responses (eg, count data) can be analyzed using seemingly unrelated Poisson regression⁷ and non-linear predictors.⁸ Similar to the approach of Zellner⁵ these models are limited in their flexibility, and only the expected value of the response is linked to the covariates.⁹ A more flexible framework is provided by generalized additive models for location, scale and shape (GAMLSS),¹⁰ in which each parameter of the conditional distribution is modeled by an additive predictor. The use of additive predictors for all distribution parameters, such as location, scale or skewness parameters allows to incorporate different effect types for the covariates in a very flexible way. Klein et al.¹¹ extended this framework for multivariate responses to model the joint distribution of two or more responses in the spirit of GAMLSS relying on a fully Bayesian approach.

In the medical literature, one common application of GAMLSS for univariate responses is the construction of reference growth charts,^{12,13} where also the World Health Organization recommends to use GAMLSS.¹⁴ As the complete conditional distribution is modeled based on covariates (eg, the age of a child), the corresponding quantiles can nicely adapt to a covariate-specific skewness. For multivariate responses, these growth charts could hence not only be constructed separately for different biometrical parameters,¹³ but also jointly for multiple characteristics.¹⁵

In high-dimensional data situations where the number of predictors exceeds the number of observations ($p > n$), classical estimation approaches are no longer directly feasible for our multivariate distributional regression settings. A few exceptions exist where Bayesian variable selection¹⁶ and penalized regression methods^{17,18} have been proposed. Nevertheless, in terms of software implementation, GAMLSS based on penalized likelihood estimation is currently only available for univariate response variables.

An alternative fitting approach is statistical boosting, which was originally developed in the field of machine learning and later extended to statistical modeling.^{19,20} Its main features are the great flexibility regarding the effect types (eg, spatial, smooth, or random effects) and the data-driven variable selection mechanism. The latter can be particularly useful when the focus is on obtaining sparse models for a possibly high-dimensional covariate space.²¹ The concept of boosting has already been extended to distributional regression leading to an algorithm that is able to estimate and select additive predictors for all distribution parameters in univariate GAMLSS.^{22,23}

In this work, we adapt the boosting algorithm for multivariate responses by combining the properties of GAMLSS and the main features of statistical boosting. Due to the structure of the algorithm, our approach is able to simultaneously model all distribution parameters and to select possible predictor effects in multivariate distributional regression models: The new multivariate boosting approach allows to model not only the marginals but also the associations between multiple outcomes through additive predictors without requiring the manual selection or comparison of different models. The application of our approach is particularly suitable for exploratory analyses (hypothesis generating) where data-driven variable selection is of special interest.

Motivated by three biomedical applications, we focus on modeling and investigating specific bivariate regression models with emphasis on the most common parametric distributions in biomedical research: the bivariate Bernoulli distribution for binary outcomes, the bivariate Poisson distribution for count data and the bivariate Gaussian distribution for continuous outcomes.^{24–26}

In the first biomedical application, the joint genetic predisposition for chronic ischemic heart disease and high cholesterol is analyzed based on a large cohort data from the UK Biobank²⁷ via the bivariate Bernoulli distribution. The main interest is to study the dependence of these phenotypes on the genetic variants and to discover possible joint associations of the two outcome variables, which is not feasible via classical approaches modeling the phenotypes separately.²⁸ In our case, we want to gain deeper insights into the relationship between chronic ischemic heart disease and high cholesterol, and the genetic variants affecting their association.

In the second application, we investigate effects for the demand on health care in Australia reported by Cameron and Trivedi.²⁹ The two considered outcomes are the number of consultations with a doctor and the number of prescribed medications, whose association is modeled using the bivariate Poisson distribution for the covariates gender, age and annual income. The research question is based on a previous analysis by Karlis and Ntzoufras³⁰ however we illustrate that our approach offers higher flexibility.

In the last epidemiological application, two indicators for acute and chronic undernutrition of children in Nigeria are jointly analyzed, which is motivated by a previous analysis by Klein et al.¹¹ The two scores are modeled with a bivariate Gaussian distribution, in which besides the marginal expectations also the scale parameter and the correlation parameter depend on covariates. In addition to several covariates describing the life situation of the children, the mother and the household they are living, spatial effects based on regional information are incorporated.

The structure of this article is as follows: Section 2 starts with a brief introduction to multivariate distributional regression models. Then we investigate the different bivariate regression models and give an insight into statistical boosting and the extended algorithm. In Section 3, we illustrate different data settings using a simulation study while Section 4 illustrates the application on biomedical research questions for the considered distributional regression models in Section 2.

2 | BOOSTING MULTIVARIATE DISTRIBUTIONAL REGRESSION

2.1 | The notion of multivariate distributional regression models

In multivariate structured additive distributional regression¹¹ it is assumed that the conditional distribution $\mathbb{P}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ of a D -dimensional vector of responses $\mathbf{Y} = (Y_1, \dots, Y_D)^\top$ given covariate information summarized in $\mathbf{X} = \mathbf{x}$ has a K -parametric density $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\boldsymbol{\theta}(\mathbf{x}))$ with covariate dependent distribution parameters $\boldsymbol{\theta}(\mathbf{x}) \equiv \boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^\top$.

Each distribution parameter θ_k is linked to a structured additive predictor η_k ³¹ via bijective parameter-specific link functions g_k , such that $g_k(\theta_k) = \eta_k$ and $g_k^{-1}(\eta_k) = \theta_k$, $k = 1, \dots, K$. The inverse link functions $g_k^{-1} \equiv h_k$ are called response functions and ensure potential restrictions of the parameter space of θ_k . The additive predictors η_k depend on (possibly different) subsets of \mathbf{x} and are of the form

$$g_k(\theta_k) = \eta_k = \beta_{0k} + \sum_{j=1}^{p_k} f_{jk}(\mathbf{x}), \text{ for } k = 1, \dots, K,$$

where β_{0k} are the intercepts and each f_{jk} , $j = 1, \dots, p_k$, represents the functional effect of covariates \mathbf{x} . The effects of the covariates can be specified in a very flexible manner and can correspond to linear, non-linear, random, interaction and further effects.^{2,32} Motivated by our applications in Section 4, in this work we focus on the following effect types:

1. Linear effects are represented by $f_{jk}(\mathbf{x}) = \mathbf{x}_{jk}^\top \boldsymbol{\beta}_{jk}$, where $\boldsymbol{\beta}_{jk}$ are the regression coefficients and \mathbf{x}_{jk} is a covariate subset of \mathbf{x} for parameter θ_k (\mathbf{x}_{jk} can be chosen individually for each parameter θ_k).
2. Non-linear effects can be included using smooth functions $f_{jk}(\mathbf{x})$. As basis functions we use B-Splines with second order difference penalties.³³
3. Spatial effects based on observations assigned to discrete regions are incorporated using Markov random fields for modeling neighborhood structures $f_{jk}(\mathbf{x}) = f_{jk}(s_i)$, where s_i denotes the region s_i observation i is located in Reference 34.

2.2 | Examples of relevant response distributions

In the following, we describe three common bivariate parametric distributions for binary, count and continuous responses, the bivariate Bernoulli, the bivariate Poisson and the bivariate Gaussian distribution. While there are of course other multivariate distributions for discrete and continuous data,^{26,35} these three bivariate distributions are arguably most commonly used and are also relevant for our applications.

2.2.1 | Bivariate Bernoulli distribution

For analyzing potentially correlated binary variables $\mathbf{Y} = (Y_1, Y_2)^\top$, we consider the bivariate Bernoulli distribution with joint probability mass function

$$p(y_1, y_2) = p_{00}^{(1-y_1)(1-y_2)} p_{10}^{y_1(1-y_2)} p_{01}^{(1-y_1)y_2} p_{11}^{y_1 y_2}, \quad y_1, y_2 \in \{0, 1\},$$

TABLE 1 Contingency table

		Y_2		
		0	1	
Y_1	0	p_{00}	p_{01}	$1 - p_1$
	1	p_{10}	p_{11}	p_1
		$1 - p_2$	p_2	1

where $p_{ij} = P(Y_1 = i, Y_2 = j)$, $i, j \in \{0, 1\}$ are the joint probabilities. Then, the contingency table with marginal probabilities $p_d = P(Y_d = 1)$, $d = 1, 2$ is given in Table 1. In a bivariate logistic regression model (logit model), the marginal probabilities $p_1 = P(Y_1 = 1)$ and $p_2 = P(Y_2 = 1)$, as well as the odds ratio $\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}}$, can be estimated considering several covariates.^{36,37} If Y_1 and Y_2 are independent, then the odds ratio $\psi = 1$. The different additive predictors in the bivariate logit model are

$$\text{logit}(p_i) = \eta_{p_i}, \text{ for } i = 1, 2 \text{ and } \log(\psi) = \eta_\psi.$$

The joint probability p_{11} can be determined from the marginal probabilities p_1, p_2 and the odds ratio ψ via

$$p_{11} = \begin{cases} \frac{1}{2}(\psi - 1)^{-1} \{a - \sqrt{a^2 + b}\} & , \psi \neq 1 \\ p_1 p_2 & , \psi = 1, \end{cases}$$

where $a = 1 + (p_1 + p_2)(\psi - 1)$ and $b = -4\psi(\psi - 1)p_1 p_2$.³⁸ The joint probabilities p_{10}, p_{01} and p_{00} can be derived from p_{11} and the marginal probabilities.

An alternative approach for modeling bivariate binary responses is the bivariate probit model. However, in this work we focus on the logit model for two reasons: First, one distribution parameter directly corresponds to the odds ratio, which is easier to interpret and much more common in Biostatistics and biomedical research than the correlation of a latent bivariate response $\mathbf{Y}^* \sim N(\mathbf{0}, \Sigma)$ for a probit model, where $Y_d = 1$ if $Y_d^* > 0$ and 0 otherwise, $d = 1, 2$ and Σ a correlation matrix. Second, the bivariate logit model is computationally favorable since it does not require the latent variables \mathbf{Y}^* .

2.2.2 | Bivariate Poisson distribution

An important bivariate model for analyzing bivariate count data can be constructed from combining three random variables. If $Z_k, k = 1, 2, 3$ follow independent Poisson distributions with parameters $\lambda_k > 0$, then the two random variables $Y_1 = Z_1 + Z_3$ and $Y_2 = Z_2 + Z_3$ follow a bivariate Poisson distribution with joint probability function given by

$$p(y_1, y_2) = \exp(-(\lambda_1 + \lambda_2 + \lambda_3)) \frac{\lambda_1^{y_1} \lambda_2^{y_2}}{y_1! y_2!} \sum_{k=0}^{\min(y_1, y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k, \quad y_1, y_2 \in \mathbb{N}_0.$$

The marginals also follow Poisson distributions with expectations $\mathbb{E}(Y_1) = \lambda_1 + \lambda_3$ and $\mathbb{E}(Y_2) = \lambda_2 + \lambda_3$. The parameter λ_3 controls the dependency between Y_1 and Y_2 and corresponds to the covariance $\text{Cov}(Y_1, Y_2) = \lambda_3$. If the variables Y_1 and Y_2 are independent, then $\lambda_3 = 0$ and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions. For further details on the bivariate Poisson distribution, we refer to Kocherlakota and Kocherlakota²⁵ and Johnson et al.³⁵

In a bivariate Poisson regression model, each distribution parameter $\lambda_k, k = 1, 2, 3$ can be modeled in terms of several explanatory variables via

$$\log(\lambda_k) = \eta_{\lambda_k}, \quad k = 1, 2, 3,$$

where η_k is the corresponding predictor for λ_k .

A drawback of this definition of the bivariate Poisson distribution is its property of modeling only data with positive correlations. An alternative was developed in Lakshminarayana et al.³⁹ by defining the bivariate Poisson distribution as the product of Poisson marginals with a multiplicative factor. This definition also allows for negative correlations, but results in more difficult interpretations. A further alternative allowing for overdispersion in the marginal distributions is the bivariate negative binomial distribution.^{25,30,40}

2.2.3 | Bivariate Gaussian distribution

The bivariate Gaussian distribution is one of the most commonly known distributions for considering two continuous responses. In this case, the random vector is written by $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the density of $\mathbf{Y} = (Y_1, Y_2)^T$ is given by

$$f(y_1, y_2) = \frac{1}{2\pi \sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right), \quad y_1, y_2 \in \mathbb{R},$$

and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and $\boldsymbol{\Sigma} = \text{Cov}(Y_1, Y_2)$ are its mean vector and covariance matrix, respectively. The latter is defined by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

with marginal variances $\sigma_1^2 = \text{Var}(Y_1)$ and $\sigma_2^2 = \text{Var}(Y_2)$ and correlation parameter $\rho = \text{Cor}(Y_1, Y_2)$. All parameters of the bivariate Gaussian distribution can be again modeled depending on covariates with parameter specific link-functions:

$$\mu_1 = \eta_{\mu_1}, \quad \mu_2 = \eta_{\mu_2}, \quad \log(\sigma_1) = \eta_{\sigma_1}, \quad \log(\sigma_2) = \eta_{\sigma_2} \quad \text{and} \quad \rho/\sqrt{(1-\rho^2)} = \eta_{\rho}.$$

For further practical and theoretical details of the bivariate Gaussian distribution, we refer to Kotz et al.²⁶ When the marginal distributions exhibit heavy tails, the bivariate t -distribution is an attractive alternative to the bivariate normal distribution (see Klein et al.¹¹ and references therein).

2.3 | Estimation via model-based boosting

Boosting originally arose from the field of supervised machine learning⁴¹ but gained increasing popularity in statistics after the concept was adapted to fit statistical regression models.^{19,20} Boosting algorithms are a flexible alternative to classical estimation approaches and have several practical advantages, such as the applicability to high-dimensional data problems and data-driven variable selection.^{21,42,43} In the context of regression, there exist different types of boosting algorithms.^{21,44} Here, we will focus on a component-wise gradient boosting algorithm with regression-type base-learners, which we refer to as *statistical boosting*.^{45,46}

This statistical boosting approach is based on minimizing a pre-specified loss function, which represents the regression problem and typically corresponds to the negative log-likelihood l of the response distribution. In every iteration of the boosting algorithm, so-called base-learners are separately fitted to the negative gradient of the loss function, and the best-performing one is updated to the current estimate. A base-learner in our context is a regression function, and usually corresponds to one specific covariate effect in the additive predictor (eg, a linear model as base-learner leads to a linear effect). An overview of possible base-learners can be found in Hofner et al.⁴⁷

For fitting multivariate distributional regression models, we extend the statistical boosting algorithm for generalized additive models for location, scale and shape²² to multivariate distributions. A schematic overview of the selection of base-learners in one iteration of the boosting algorithm for multivariate responses can be found in Figure 1.

First, for each additive predictor η_k , $k = 1, \dots, K$, a set of base-learners $h_1(x_1), \dots, h_{p_k}(x_{p_k})$ has to be specified in advance. Then, the partial derivative $u = \partial l / \partial \theta_k$ of the negative log-likelihood function l with respect to the different distribution parameters θ_k is calculated and each base-learner is fitted separately to the gradient of the corresponding parameter k . For each parameter, the best performing base-learner j_k^* is determined. After these best-fitting base-learners are selected for each dimension k , only the overall best update (with the highest loss reduction) of all distribution parameters is

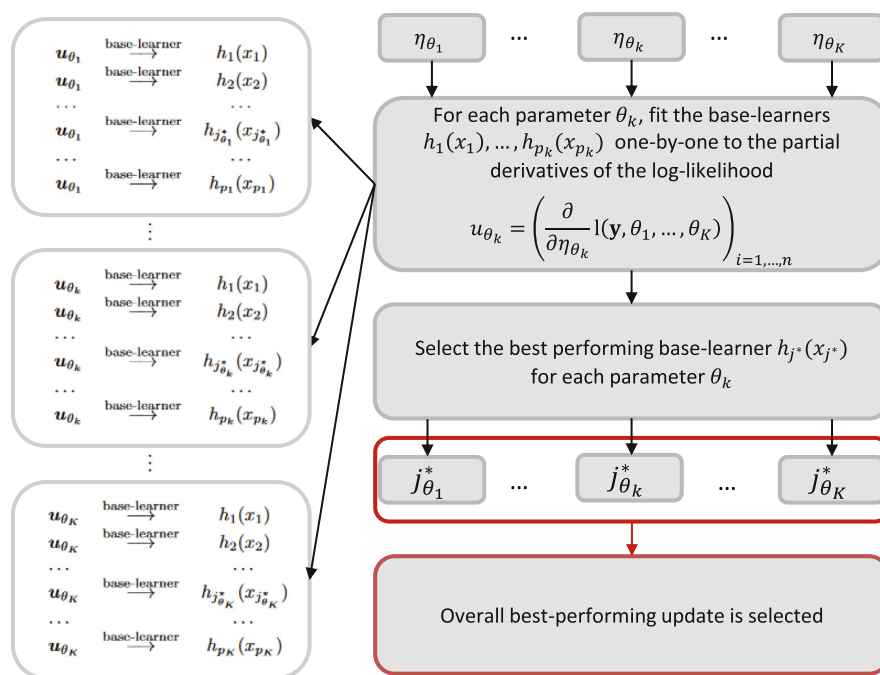


FIGURE 1 Graphical representation of boosting multivariate structured additive distributional regression (displaying one boosting iteration)

finally added to the corresponding additive predictor, with the estimated effect multiplied by a small fixed step-length, for example, $\nu = 0.1$. That means, in every iteration the best-fitting base-learner is determined for each distribution parameter and then compared across the different dimensions. This refers to a so-called non-cyclic version of boosting for distributional regression, leading to a single update of only one distribution parameter in each iteration.²³ Current best-practice in statistical boosting is to use fixed small step-lengths like $\nu = 0.1$ without optimization. Very recently there was work on adaptive step-lengths, particularly for more complex and multi-dimensional models such as GAMLSS. However, so far only the Gaussian location-scale model has been investigated empirically by Zhang et al.⁴⁸ The authors propose to use a different step-length for the two parameters, but solutions for more complex models require further research.

The main tuning parameter of the algorithm is the number of boosting iterations, which is typically chosen by cross-validation or resampling techniques. As the algorithm is usually stopped before convergence (*early stopping*), the optimization of the stopping iteration leads to the prevention of overfitting and encourages the sparsity of the resulting model by data-driven variable selection.⁴⁹ In particular, those variables, whose corresponding base-learners have never been selected in the update process, are effectively excluded from the final model. The variable selection is simultaneously based on all additive predictors of the corresponding multivariate distribution. The algorithm does not impose any hierarchy between distribution parameters, but only judges the potential predictor variables based on their performance in increasing the joint likelihood. In addition, early stopping typically leads to an improvement in the prediction accuracy and shrinkage of the effect estimates. We provide an implementation of statistical boosting for multivariate distributional regression, which is integrated in the R package **gamboostLSS**.⁵⁰

The boosting approach yields several advantages compared to existing Bayesian and likelihood-based approaches in the context of GAMLSS.^{10,11} First, boosting incorporates data-driven variable selection. The issue of variable selection is particularly important in complex model classes, for example, for multivariate distributional regression. The complexity can further increase in settings with many distribution parameters K or high-dimensional predictors with many covariates. In these cases the boosting approach could be favorable since it avoids manual selection of a large number of potential candidate models. Second, the effect estimates are shrunk towards zero due to early stopping of the boosting algorithm. This tends to result in more stable predictions as the variance of the estimates is reduced. Finally, the boosting algorithm can be also applied for high-dimensional data problems, where we have more covariates than observations ($p > n$). Other

approaches, such as more classical Bayesian approaches, are no longer applicable or computationally very demanding for these data situations.

3 | SIMULATIONS

To evaluate the performance of the proposed statistical boosting approach, we conducted a detailed simulation study for the three response distributions presented in Section 2.2. For each distribution, the particular settings are guided by the different applications in Section 4. With our simulations, we aim to answer the following questions:

- Does the boosting approach yield accurate estimates for the corresponding distribution parameters of the bivariate distributions?
- Can the boosting approach identify the truly informative variables and their effects?
- How do the bivariate models perform compared to univariate models that assume independence between the two response components?

In particular, we evaluate the estimation, variable selection and predictive performance. Note that for each considered simulation setting, different variables are informative for the distribution parameters and some of them partially overlap. Therefore, we refer to informative and non-informative variables and do not mention all of them individually for the different settings.

For all simulations, the step-length (learning rate) of the boosting algorithm is set to a fixed value of $v = 0.1$ for each parameter of the bivariate models, as well as for the univariate boosted models. The stopping iteration m_{stop} is optimized by minimizing the empirical risk on an additional validation data set with $n_{\text{val}} = 1500$ observations, following the same distribution as the training data. In addition, test data with 1000 observations were generated for the evaluation of the predictive performance (from the same distribution as the training data). As evaluation criteria, multivariate proper scoring rules, namely the negative log-likelihood and the energy score, were used. The energy score generalizes the continuous ranked probability score for multivariate quantities.⁵¹ In addition, univariate distribution-specific evaluation criteria were used, that is, the mean squared error of prediction (MSEP), the area under the curve (AUC) and the Brier score. The Brier score can be used to assess the accuracy of binary classifications and prediction models and is similar to the mean squared error of prediction by considering the mean squared difference between the actual binary outcome and its predicted probability.⁵² In contrast to the AUC, which basically only measures the discriminatory power, the Brier score additionally also considers the calibration of the prediction model. Note that the AUC and Brier score do not account for the dependence between the two outcomes and are calculated separately for both outcomes, while the negative log-likelihood and energy scores are probabilistic measures considering the entire joint outcome distribution. A total of 100 simulation runs were performed for each simulation setting.

The corresponding R code to reproduce the results is available on GitHub <https://github.com/AnnikaStr/DistRegBoost>. Further simulation results, such as comparisons with seemingly unrelated regression and Bayesian approaches, can be found in the Appendix.

3.1 | Bivariate Bernoulli distribution

3.1.1 | Simulation design

For the simulation of the bivariate logit model, we considered a situation with $n = 1000$ observations and $p = 1000$ covariates for each of the three distribution parameters, which corresponds to a high-dimensional situation as the number of possible regression coefficients has tripled due to the three distribution parameters ($3p > n$). For data generation, the R package **VGAM**⁵³ was used, whereby the parameters p_1, p_2 and ψ were simulated with the following linear predictors

$$\begin{aligned}\text{logit}(p_1) = \eta_{\mu_1} &= X_1 + 1.5X_2 - X_3 + 1.5X_4, & \text{logit}(p_2) = \eta_{\mu_2} &= 2X_1 - X_2 + 1.5X_3, \\ \text{log}(\psi) = \eta_{\psi} &= -1.5 + 1X_3 + 1.5X_4.\end{aligned}$$

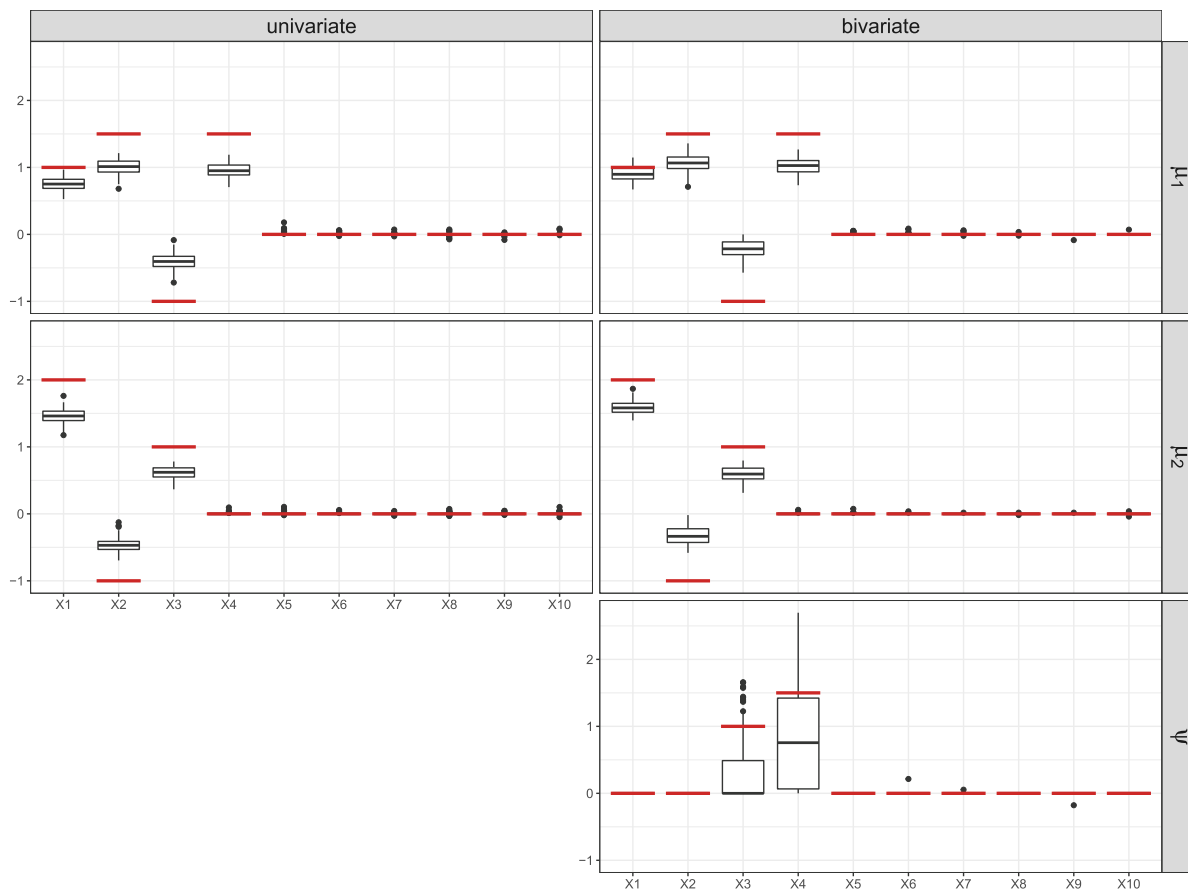


FIGURE 2 Results for the estimated linear effects of the univariate (left) and bivariate Bernoulli (right) model of the first ten covariates X_1, \dots, X_{10} from 100 simulation runs. The red horizontal lines correspond to the true values

Overall, only the first six covariates out of the $p = 1000$ had a relevant effect on any of the distribution parameters (four for p_1 , three for p_2 and two for ψ). The covariates were simulated from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with a Toeplitz covariance structure $\Sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq p$, where $\rho = 0.5$ is the correlation between consecutive variables X_j and X_{j+1} . The covariates were incorporated in the boosting approach by using simple linear models as base-learners. As measures for the predictive performance, AUC, the Brier score, the negative log-likelihood and energy score were considered.

3.1.2 | Results

Figure 2 presents the coefficient estimates of the first ten covariates X_1, \dots, X_{10} in form of boxplots for the univariate (left) and bivariate (right) model with the red horizontal lines corresponding to the true values. The univariate and bivariate models reflect the true structure for η_{μ_1} and η_{μ_2} , as well as η_{ψ} for the bivariate model, with both models leading to very similar results. The informative variables for μ_1 and μ_2 were correctly selected in almost every simulation run. Specifically, we obtained overall selection rates (averaged over all informative variables) of 100% for the univariate models and of 100% for μ_1 and 98.75% for μ_2 in the bivariate model. The selection rate for ψ is slightly lower than for the other parameters with a selection rate of 59.5% (see Appendix Table A3). The non-informative variables were selected very rarely overall, resulting in sparse models and accurate model specifications that are able to recover the ground truth.

TABLE 2 Resulting predictive performance on independent test data for the linear setting of the bivariate Bernoulli distribution; mean (SD) values from 100 simulation runs are reported for the univariate and bivariate models

	Univariate	Bivariate
AUC (Y_1)	0.88 (0.01)	0.88 (0.01)
AUC (Y_2)	0.84 (0.01)	0.84 (0.01)
Brier score (Y_1)	0.14 (0.01)	0.14 (0.01)
Brier score (Y_2)	0.16 (0.01)	0.16 (0.01)
Energy score	0.28 (0.21)	0.27 (0.01)
Negative log-likelihood	930.51 (24.24)	906.64 (29.24)

A comparison of the predictive performance is provided in Table 2, showing that the univariate and bivariate models were very similar in terms of AUC, Brier score, and energy score, with the bivariate model having slightly better negative log-likelihood. In addition, the energy score for the univariate models showed a larger standard deviation. Further simulation results of this linear setting for a low-dimensional data situation ($p = 10$ and $n = 1000$) can be found in Appendix A.1.

3.2 | Bivariate Poisson distribution

3.2.1 | Simulation design

For the bivariate Poisson regression model, we investigated both linear and non-linear settings with $p = 10$ covariates and $n = 1000$ observations for each distribution parameter. For the linear setting, the underlying true predictors were specified as

$$\begin{aligned}\log(\lambda_1) = \eta_{\lambda_1} &= -X_1 + 0.5X_2 + 1.5X_3, & \log(\lambda_2) = \eta_{\lambda_2} &= 2X_1 - X_3 + 1.5X_4 + X_5, \\ \log(\lambda_3) = \eta_{\lambda_3} &= 0.5X_5 + X_6 - 0.5X_7,\end{aligned}\quad (1)$$

where the covariates followed a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with Toeplitz covariance structure and correlation coefficient $\rho = 0.5$. Thus, the first seven covariates were informative for any of the distribution parameter (three for λ_1 and λ_3 , four for λ_2). For this setting, simple linear models were incorporated as base-learners. For the non-linear setting, the true additive predictors were given by

$$\begin{aligned}\log(\lambda_1) = \eta_{\lambda_1} &= \sqrt{X_1}X_1, & \log(\lambda_2) = \eta_{\lambda_2} &= \cos(2X_2), \\ \log(\lambda_3) = \eta_{\lambda_3} &= \sin X_3,\end{aligned}\quad (2)$$

where the covariates were independently simulated from the uniform distribution $U(0, 1)$ and only one covariate was informative for each of the distribution parameters. As base-learners, we chose P-splines (20 equidistant knots with a second-order difference penalty and four degrees of freedom). The R **extraDistr**⁵⁴ package was used to simulate data from the bivariate Poisson regression model.

3.2.2 | Results

Figure 3 displays the coefficient estimates for the linear Poisson regression models (1). The boxplots present the estimated coefficients for the univariate (left) and bivariate models (right). Overall, boosting the bivariate regression model was able to identify the informative variables and to accurately estimate the true effects represented by the red horizontal lines. In comparison, the univariate models for λ_1 and λ_2 resulted in much smaller estimated coefficients. For both models, the informative variables were selected in almost every simulation run: considering λ_1 and λ_2 , the

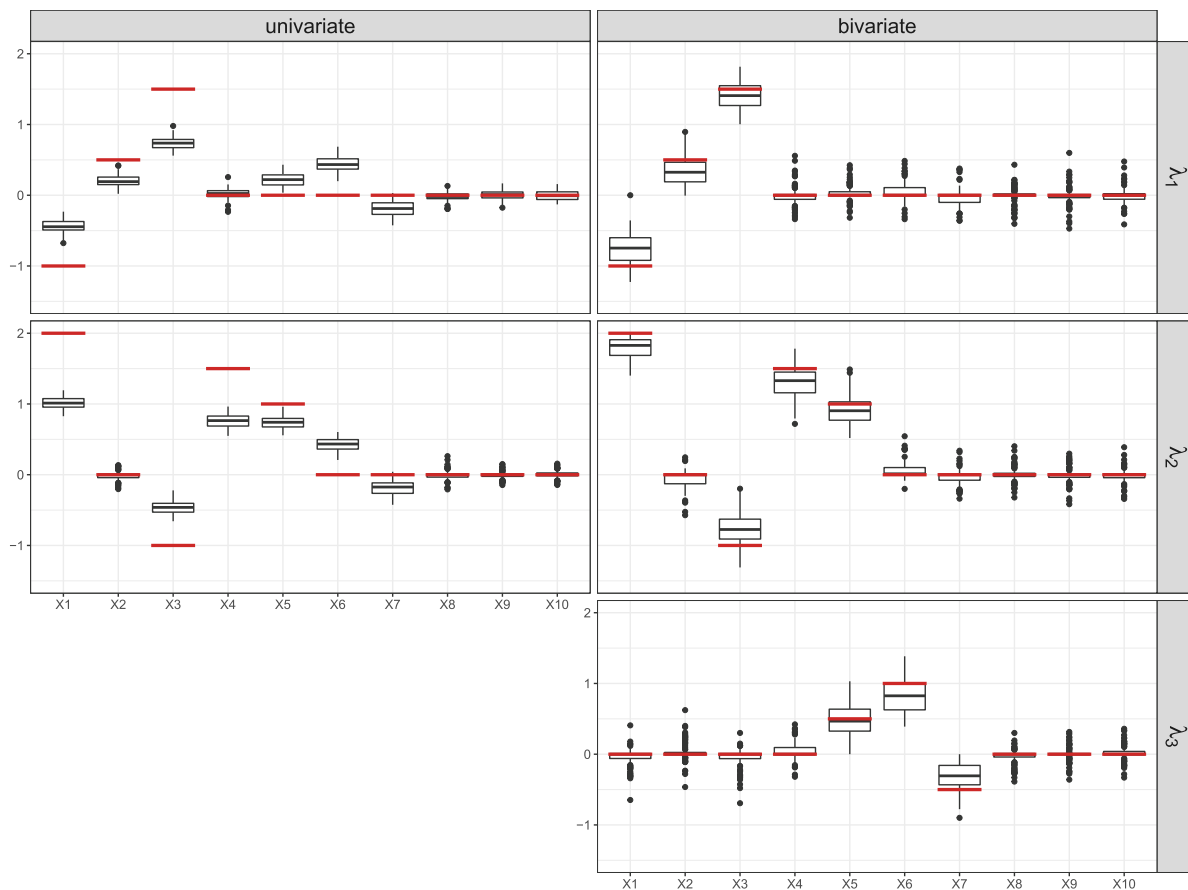


FIGURE 3 Results for the estimated linear effects of the univariate (left) and bivariate Poisson model (right) from 100 simulation runs. The horizontal lines correspond to the true values

univariate models and the bivariate model had a selection rate of almost 100% for the informative variables, whereby also for λ_3 a high selection rate of 95.67% for the informative variables was achieved. On the other hand, the univariate models as well as the bivariate model selected also several non-informative variables with a small coefficient size. A more detailed overview on the selection rates for the specific parameters can be found in Table A8 of the Appendix.

Furthermore, we considered the MSEP, the negative log-likelihood, and the energy score for the evaluation of the predictive performance on test data (see Tables 3 and 4). The MSEP only accounts for the marginal distributions and displays here a slightly better performance for the univariate models. The negative log-likelihood and the energy score, which also take the association into account, showed a better performance for the bivariate model.

Figure 4 displays the effect estimates of the informative variables X_1 , X_2 and X_3 for the non-linear setting (2). Overall, the estimated splines approximate the true effects well for each parameter of the bivariate model and clearly outperform the univariate models for λ_1 and λ_2 . The informative variables were selected in each simulation run. However, as in the linear model, we observed also high selection rates for the non-informative variables in both models (see Appendix A.2, Table A8).

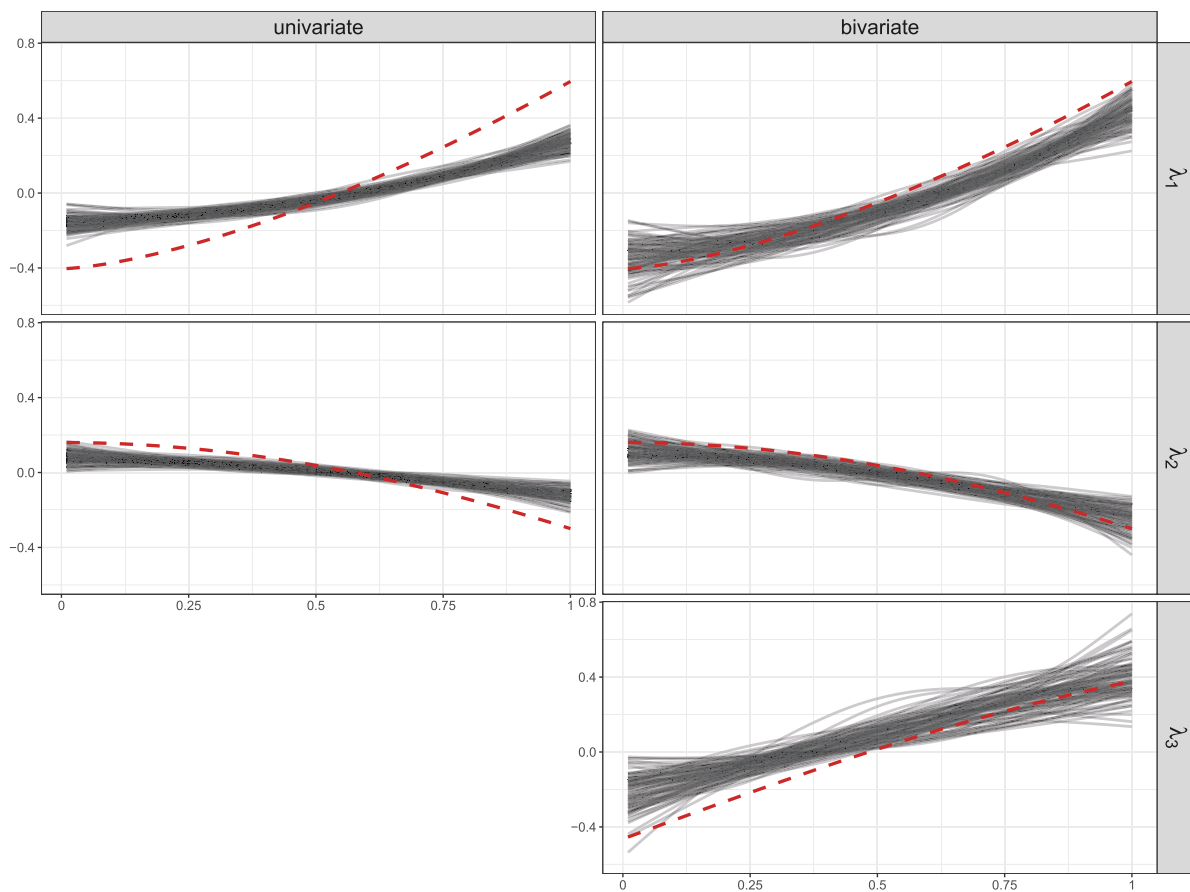
In terms of predictive performance, similar to the linear setting, the MSEP indicated a better performance of the univariate models, while the bivariate models, as expected, yielded better results in terms of the negative log-likelihood. The energy score was very similar for both models but overall slightly better for the bivariate model. Further simulation results for these settings in case of high-dimensional data with $p = 1000$ covariates and $n = 1000$ observations can be found in Appendix A.2.

TABLE 3 Resulting predictive performance on independent test data for the linear and non-linear settings of the bivariate Poisson regression; mean (SD) values from 100 simulation runs are reported for the univariate and bivariate models

	Linear model		Non-linear model	
	Univariate	Bivariate	Univariate	Bivariate
MSEP (Y_1)	2.66 (0.18)	3.96 (0.29)	4.64 (0.25)	8.18 (0.65)
MSEP (Y_2)	2.86 (0.23)	4.11 (0.34)	5.49 (0.29)	9.06 (0.68)
Energy score	1.48 (1.11)	1.36 (0.03)	1.95 (0.04)	1.95 (0.04)
Negative log-likelihood	3598.42 (54.31)	3413.68 (40.91)	4433.06 (52.06)	4246.96 (42.58)

TABLE 4 Resulting predictive performance on independent test data of the bivariate Gaussian regression; mean (SD) values from 100 simulation runs are reported for the univariate and bivariate models

	Univariate	Bivariate
MSEP (Y_1)	1.59 (0.11)	1.59 (0.11)
MSEP (Y_2)	1.38 (0.07)	1.38 (0.07)
Energy score	1.03 (0.02)	1.01 (0.02)
Negative log-likelihood	3370.41 (89.59)	3098.11 (109.97)

**FIGURE 4** Results for the estimated non-linear effects for the univariate (left) and the bivariate Poisson model (right) of the informative variables from 100 simulation runs. The red dotted lines correspond to the true effects

3.3 | Bivariate Gaussian distribution

3.3.1 | Simulation design

For the simulation of a bivariate Gaussian distributed outcome, we considered a setting with linear, non-linear and spatial effects with $p = 10$ covariates and $n = 1000$ observations with the following true predictors

$$\begin{aligned}\mu_1 &= \eta_{\mu_1} = \sin(2X_1)/0.5 + X_6 + 0.5X_7 + f_{\text{spat}} & \mu_2 &= \eta_{\mu_2} = 2 + 3 \cos(2X_2) + 0.5X_7 + X_8 + f_{\text{spat}} \\ \log(\sigma_1) &= \eta_{\sigma_1} = \sqrt{X_3}X_3 - 0.5X_8 + f_{\text{spat}} & \log(\sigma_2) &= \eta_{\sigma_2} = \cos(X_4)X_4 + 0.25X_9 + f_{\text{spat}} \\ \rho/\sqrt{1-\rho^2} &= \eta_{\rho} = \log(X_5^2) + X_{10} + f_{\text{spat}},\end{aligned}$$

where the covariates were independently simulated from the uniform distribution $U(0, 1)$. Each included covariate was informative for one of the distribution parameters; more precisely, for each parameter three covariates, one linear and one non-linear, and additionally the spatial effect. For the linear effects (X_6, \dots, X_{10}) we used simple linear models as base-learners and P-splines for the non-linear effects (X_1, \dots, X_5). The spatial effects were simulated with $f_{\text{spat}}(s) = \sin(x_s^c) \cos(0.5y_s^c)$, $s \in 1, \dots, S$, based on the centroids of the standardized coordinates of the discrete regions in Western Germany with overall $S = 327$ regions. The neighborhood structure was modeled by the spatial base-learner using a Markov random field based on the R package **BayesX**.⁵⁵

3.3.2 | Results

Considering the linear effects (X_6, \dots, X_{10}), the effect estimates for both models reflect the true structures of the linear part of the predictors, whereby the bivariate model better approximates the true values (see Figure A6 in Appendix A.3). The bivariate model was also able to capture the true non-linear functions well (Figure 5); only small deviations are observed for the variance and for the correlation ρ at the left border. The results for the univariate models appear to be very similar regarding the univariate effects and can be found in the Appendix (Figure A7). For the spatial effects, the true structure for the regions in West Germany was identified by each distribution parameter (a graphical representation of the true structure and the estimated spatial effects are in Appendix A.3).

The informative variables for the univariate and bivariate models were selected in nearly all 100 simulation runs, where the bivariate model also correctly selected the informative variables for the correlation between the outcomes. Whereas, we can not examine the correlation with the univariate models. The selection rates for the non-informative variables were slightly higher for the bivariate model (see Appendix Table A11).

Regarding predictive performance, the MSE, the energy score, and the negative log-likelihood were considered. For the MSE and the energy score, similar results were observed for the univariate and the bivariate models. The negative log-likelihood on the test set showed an improvement in predictive performance considering the bivariate model. Further simulation results for this setting in case of high-dimensional data with $p = 1000$ covariates and $n = 1000$ observations can be found in Appendix A.3.

3.4 | Summary

Overall, we obtained promising results for all three considered distributional regression families (logistic, Poisson, and Gaussian regression), highlighting that the boosting algorithm yields appropriate estimates for the different parameters and is capable of identifying the most informative variables from a potentially much larger set of candidate variables. The comparison with the univariate models showed that the estimated effects for the bivariate model were able to provide better approximations to the true structure of the predictors than the univariate models (particularly for the Poisson and Gaussian regression models). We noticed that for the logistic regression model, the selection rates for the association parameter, the odds ratio, tended to be lower than for the association parameter of the bivariate Poisson and Gaussian distribution. Furthermore, the number of selected non-informative variables was higher for the univariate as well as the bivariate models for the Poisson distribution. However, the linear low-dimensional setting for the bivariate logistic regression model (see Appendix A.1) and the low-dimensional Gaussian regression model showed higher selection rates in this

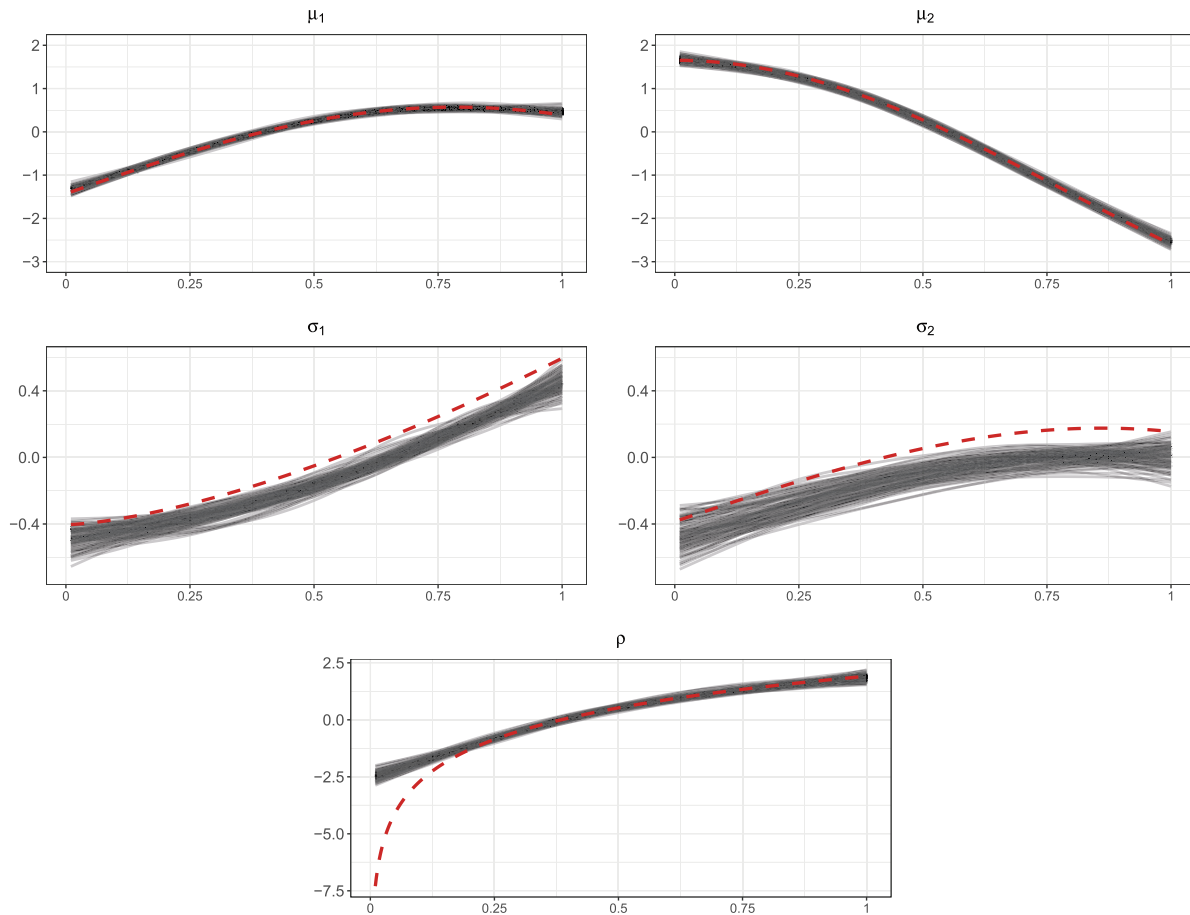


FIGURE 5 Results for the estimated non-linear effects (X_1, \dots, X_5) of the bivariate Gaussian regression model from 100 simulation runs. The red dotted lines correspond to the true effects

situation as well (Appendix A.3). Conclusively, this highlights a tendency of the algorithm to select more non-informative variables in low-dimensional settings.

Regarding prediction accuracy, as expected, the univariate and bivariate models performed similarly for evaluation criteria that consider only the marginals (AUC, Brier score and MSE). Only for the Poisson distribution, the univariate model performed slightly better regarding the MSE. This can be explained by the particular design of this bivariate distribution, that is, the summation of the means for both outcomes ($\mathbb{E}(Y_1) = \lambda_1 + \lambda_3$ and $\mathbb{E}(Y_2) = \lambda_2 + \lambda_3$). In Figure 3, for example, we observe that the informative variables X_5, X_6 and X_7 for parameter λ_3 were selected quite frequently with a higher estimated coefficient in the univariate models. These wrongly selected variables for the marginals resulted in an improvement of the MSE. In the bivariate model, we account for the association between Y_1 and Y_2 by modeling the dependency in terms of the covariates. The MSE does not account for the association and the variables describing dependency are not reflected in the marginals as in the univariate models. Regarding the predictive scores which account for associations between the outcomes, the energy score tended to be very similar for the univariate and bivariate models, while the negative log-likelihood was consistently better for the bivariate models.

Overall, the bivariate models are obviously more complex because they also model the association between the two outcomes. Regarding the marginals, however, they yield similar models as when analyzing both responses separately. Therefore, we also get similar prediction performance for the marginals from the univariate and bivariate models. The distributional evaluation measures like the negative log-likelihood, on the other hand, are consistently better for the bivariate models as they consider the complete multivariate distribution.

4 | BIOMEDICAL APPLICATIONS

In this section we consider three diverse biomedical data sets to illustrate the applicability of our extended boosting approach for multivariate distributional regression models based on binary, count and continuous outcomes presented in Section 2.2.

4.1 | Genetic predisposition for chronic ischemic heart disease and high cholesterol

For analyzing the association between high cholesterol and chronic ischemic heart disease in dependency of different genetic variants, we used cohort data from the UK Biobank (under application number 81202). The UK Biobank is a large biomedical cohort study containing genetic and health information from over half a million British participants.²⁷

In classical approaches for analyzing a potential genetic liability to a specific phenotype such as high cholesterol or chronic heart disease, each considered genetic variant is fitted individually to the phenotype using a simple linear model.²⁸ In this context, previous works including genome-wide association studies^{56,57} have investigated to find genetic variants associated with high cholesterol and heart disease. Using our boosting algorithm for multivariate distributional regression, the main interest here is to investigate the association between chronic ischemic heart disease and high cholesterol, both considered as binary phenotypes (high cholesterol > 6.16 mmol/l). In particular, we aim to identify genetic variants affecting their association by estimating the two phenotypes jointly in a bivariate logistic model. That means we do not only want to model the individual distributions of the two phenotypes, but also estimate the dependency between these phenotypes as a function of genetic variants, which is not possible with conventional approaches.

The considered data set consists of 20,000 randomly sampled observations of individuals with white British ancestry, with additional 10,000 observations used to validate the optimal stopping iteration. The fixed step-length was set to $\nu = 0.1$. For each phenotype, 1000 variants were selected in a pre-screening step based on the largest marginal associations between the variants and the phenotype, which were computed with the PLINK2 function `-variant-score`.^{58,59} After pre-screening, the data set contains a total of 1865 variants (with 135 variants selected for both phenotypes). Variants with minor allele frequency not less than 1% were randomly sampled with the `-thin-count` function. Missing genotypes were imputed by the reference allele using the R package **bigsnpr**.⁶⁰ Note that the pre-screening of 1000 genetic variants for each phenotype and the usage of 20,000 randomly sampled observations from the much larger cohort was performed to avoid computational memory problems. While there exists recent approaches to use boosting to fit multivariable regression models for single phenotypes on the complete data set, classical methods to model the genetic liability are based on summing up univariate effects.^{61,62}

Figure 6 shows the resulting estimated coefficients (expressed in exponential absolute values of the estimated coefficients) for the three distribution parameters. When comparing these Manhattan plots with the classical univariate ones (based on the marginal association evaluated on the $-\log_{10}(P)$ scale) for high cholesterol and chronic ischemic heart disease, we find that the bivariate boosting model tended to identify variants with a higher coefficient value (stronger effect) from similar or the same genomic locations, where the univariate models also showed large univariate associations (see Appendix B.1).

For high cholesterol, for example, the variants with the smallest univariate p -values are located on chromosomes 18 and 19; on these chromosomes there were also the variants that had the highest estimated coefficients in the bivariate boosting model. These findings are consistent with the location of known cholesterol-associated genes.⁵⁷ Variants from these chromosomes were also selected with our approach for the odds ratio. Our model selected several variants for chronic ischemic heart disease that are in line with the findings of the meta-analysis of genome-wide association studies examining DNA sequence variants associated with ischemic heart disease of Elosua and Sayols-Baixeras⁶³ (eg, the variants rs11206510, rs2891168, and rs4420638).

Overall, several variants were selected by the boosting approach for each distribution parameter, that is, 75 for μ_1 , 154 for μ_2 , and 19 variants for ψ . For the marginal means, mainly those variants were selected that were primarily filtered due to the specific phenotype (the 1000 most highly associated from the univariate screening for both phenotypes). In particular, for μ_1 , 63 of the 75 selected variants had been chosen in the pre-screening for ischemic heart disease, so that 12 of the 75 selected variants for μ_1 were primarily selected for high cholesterol. Regarding high cholesterol, 110 variants were selected from the ones that had been pre-selected for this phenotype, while 44 of the selected variants for high cholesterol had been originally pre-selected for ischemic heart disease. The two marginal means μ_1 and μ_2 had six selected variants in common, and both had one variant that was also selected for the odds ratio ψ (namely for μ_1 : rs10455872 and for μ_2 :

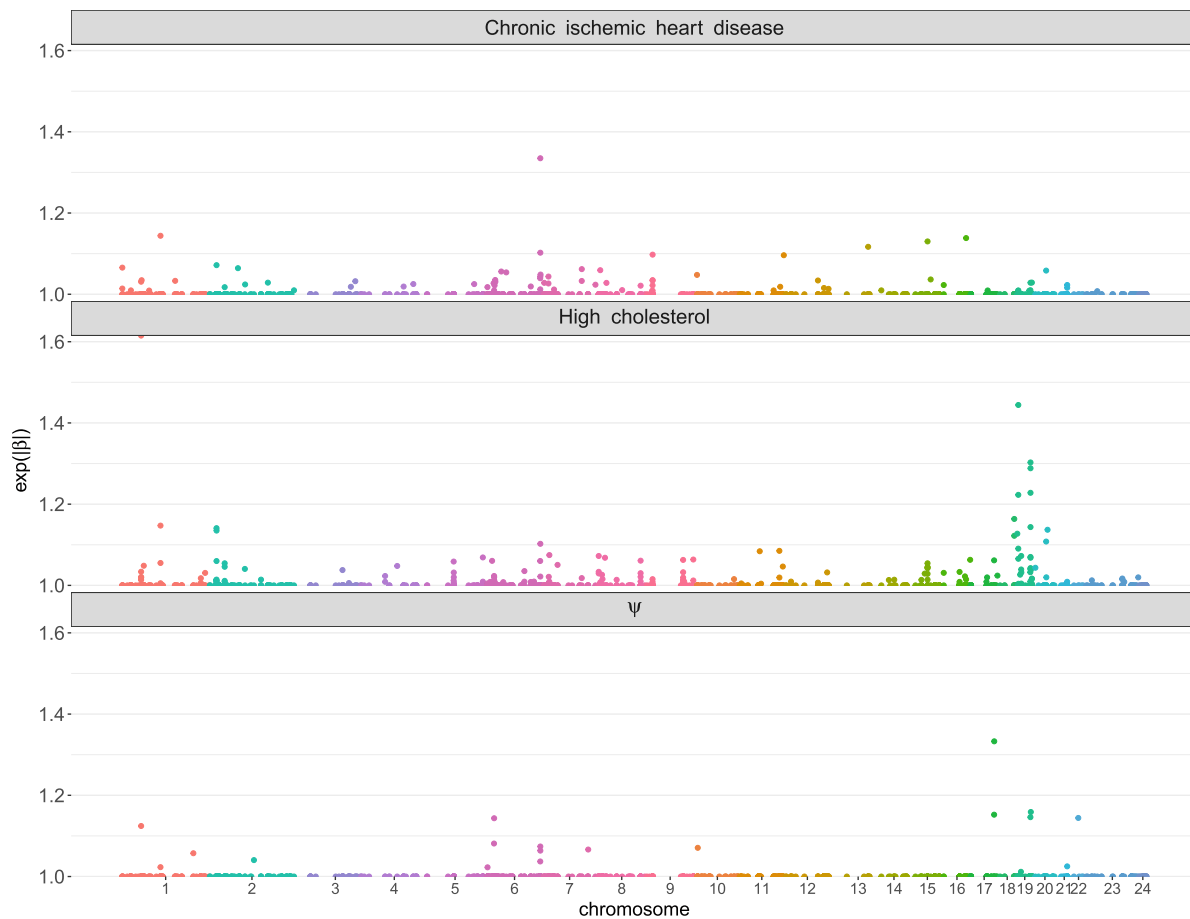


FIGURE 6 Manhattan plots for the coefficient estimates (expressed in exponential absolute values of the estimated coefficients) of the boosted bivariate logistic regression model of the joint analysis of high cholesterol and ischemic heart disease from the UK Biobank data. The x-axis represents the genomic location of the variants

rs77542162). The odds ratio included two variants that were among the 1000 most highly correlated pre-selected variants for both phenotypes, namely rs505151 and rs2229094. The other 17 variants selected for the odds ratio were divided as follows: 10 from μ_1 (ischemic heart disease) and 7 from μ_2 (high cholesterol). This means the algorithm identified several variants that affect the dependency between the two phenotypes. The odds ratio is the most common measure for examining the dependency between two binary outcomes in biomedical research and the interpretation in our context is very similar. Thus, the selected variants for the association parameter have an effect on both outcomes, with a positive effect increasing the association between heart disease and high cholesterol and conversely.

In summary, our algorithm provides the ability to study the joint genetic predisposition for chronic ischemic heart disease and high cholesterol. With our approach we can also model the dependence of the association between these two phenotypes on genetic variants, which is not possible with classical approaches. In addition, in line with the literature on cardiovascular genetics, our model selected several variants in genomic regions which had been previously identified to be relevant for the considered phenotypes.

4.2 | Demand for health care in Australia

The first analysis on the demand for health care in Australia, based on the Australian health survey from 1977 to 1978, was reported by Cameron and Trivedi.²⁹ The considered data set consists of $n = 5190$ observations (which is only a subset

TABLE 5 Results of the bivariate poisson model for the demand of health care for model A and model B (see Figure 7 for the non-linear effect estimates for age and income in model B)

	Covariate	$\lambda_{\text{consultations}}$	$\lambda_{\text{medications}}$	λ_3
Model A	Intercept	-2.10	-2.20	-2.62
	Gender (female)	0.05	0.59	0.61
	Age	1.40	3.29	—
	Income	-0.31	-0.10	—
Model B	Intercept	-2.29	-2.22	-0.35
	Gender (female)	0.13	0.60	0.19

of the overall collected survey). The bivariate count variables of interest are the number of consultations with a doctor (in the past 2 weeks) and the number of prescribed medications (used in the last 2 days), which we model using bivariate Poisson regression. The explanatory variables are *gender* (female coded as 1, male as 0), *age* (in years divided by 100) and annual *income* (in Australian dollars; AUD; divided by 1000, measured as midpoints of coded ranges). More details on the survey and its original analysis can be found in Cameron and Trivedi.²⁹ The data are provided in the R package **bivpois**,³⁰ which is available on GitHub (<https://github.com/cran/bivpois>).

In the following, we use the same representation of the bivariate Poisson distribution as introduced in Section 2.2.2. Each distribution parameter $\lambda_k, k = 1, 2, 3$ is modeled based on explanatory variables. We consider the two following models:

ModelA *Gender, age and income* are included as covariates for $\lambda_{\text{consultations}}$ (number of doctor consultations) and $\lambda_{\text{medications}}$ (number of medications prescribed), but only *gender* is considered as a covariate for the covariance parameter λ_3 (corresponding to Model (b) in Karlis and Ntzoufras³⁰).

ModelB For each model parameter, P-splines are used as base-learners for the continuous variables *age* and *income*, while for *gender* linear effects are used.

The stopping iteration of both models was tuned via 25-fold bootstrapping and the step-length was set to a fixed $\nu = 0.1$.

Considering the results of Model A presented in the upper part of Table 5, we observe that with increasing *age*, both the numbers of doctor consultations and prescribed medications are estimated to increase. *Income* has negative marginal effects on both responses, which means that higher *income* is associated with fewer prescribed medications and fewer doctor appointments. For the covariance parameter λ_3 , only *gender* was included as an explanatory variable in Model A. The joint effect of *gender* on the number of doctor consultations and prescribed medications indicates that males and females have different covariance terms. The estimated effect of 0.61 for *gender* suggests that the association between numbers of consultations and medications is higher for women than for men.

The lower part of Table 5 and Figure 7 present the results for Model B. With an increasing *age* up to 50 years, the number of doctor consultations is estimated to increase linearly, with a slight decrease starting around the age of 57 years. The estimated effect of *age* on the number of prescribed medications and the covariance parameter is linear and is negative throughout the covariance. The *income* is estimated to have a U-shape effect for the medical consultation, with a minimum between 800 AUD and 1150 AUD. The estimated effect of *gender* for Model A is slightly larger than for Model B.

Overall, the estimated effects of Model A are consistent with the results of Karlis and Ntzoufras.³⁰ In addition, we also considered a non-linear model. Both the linear and non-linear models indicated that the expected numbers of doctor consultations and prescribed medications increase with *age*. For *income*, the expected numbers of doctor visits and prescribed medications decreases with increasing *income* for Model A. The expected number of doctor visits also decreases in Model B as *income* increased, whereby a U-shaped effect for *income* can be observed.

Furthermore, because of the bivariate modeling, we also obtain information about the relationship between the outcomes. Here, both models showed a higher association between the number of doctor consultations and prescribed medications for women. Furthermore, Model B also included *age* and *income* as covariates for the covariance parameter and the model suggested that the association becomes greater with increasing *age*.

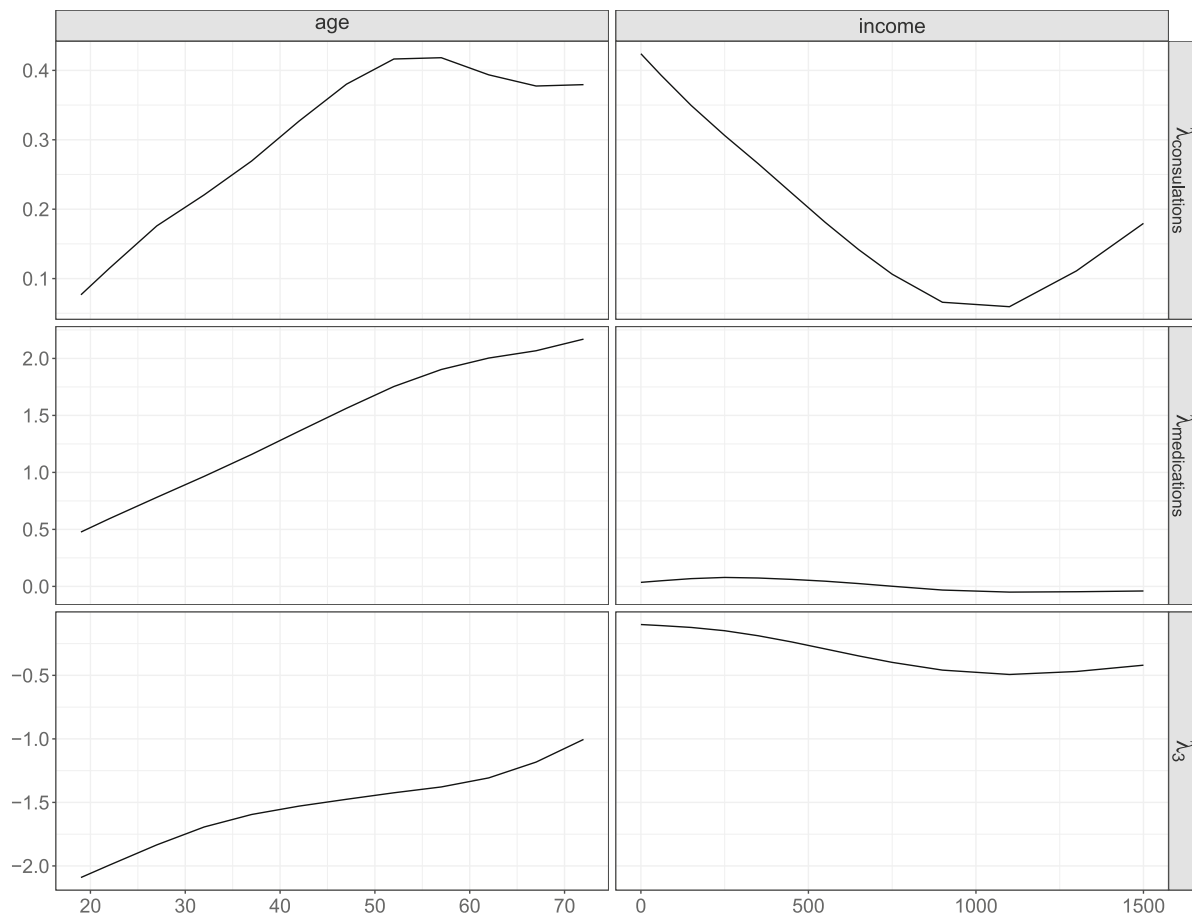


FIGURE 7 Partial effects of age and income on the demand for health care in Australia for model B

4.3 | Risk factors for undernutrition in Nigeria

To analyze childhood undernutrition, a large database is available from the Demographic and Health Survey (DHS, <https://dhsprogram.com/>), containing nationally representative information about the population's health and nutrition status in numerous developing and transition countries. Here, we consider a data set used in Klein et al.⁶⁴ which contains data from Nigeria collected in 2013 with overall 23,042 observations (after exclusion of outliers and inconsistent observations). The bivariate responses are *stunting*, which is defined as stunted growth measured as the insufficient height of the child concerning its age (chronic undernutrition), and *wasting*, which refers to insufficient weight for height (acute undernutrition). We analyze the joint distribution of these two responses using the bivariate Gaussian distribution with covariate-dependent marginal means and standard deviations as well as a covariate-dependent correlation parameter.

For continuous variables, P-splines were applied as base-learners, namely for *cage* (age of the child in months), *edu-partner* (years of partner's education), *mage* (age of the mother in years) as well as *mbmi* (body mass index of the mother). Several other categorical covariates (12 covariates in total, eg, *bicycle*, *car*, *cbirthorder*) were included using simple linear models as base-learners. Furthermore, the neighborhood structure of the districts in Nigeria was incorporated and modeled by the spatial base-learner using a Markov random field. For a full description of the explanatory variables, see Appendix B.2. The stopping iteration of both models was tuned by 25-fold bootstrap and the step-length was set to $\nu = 0.1$.

Figures 8 and 9 show the results for the non-linear and spatial effects for all parameters. The estimated linear effects are given in Appendix Table B2. *Stunting* is estimated to be more affected by variables describing children's living situation, particularly *ctwin* (child is a twin) and the birth order (*cbirthorder*). Following our model, with higher birth order, the *stunting* score decreases, with negative values indicating that the children's growth is below the expected growth of a child

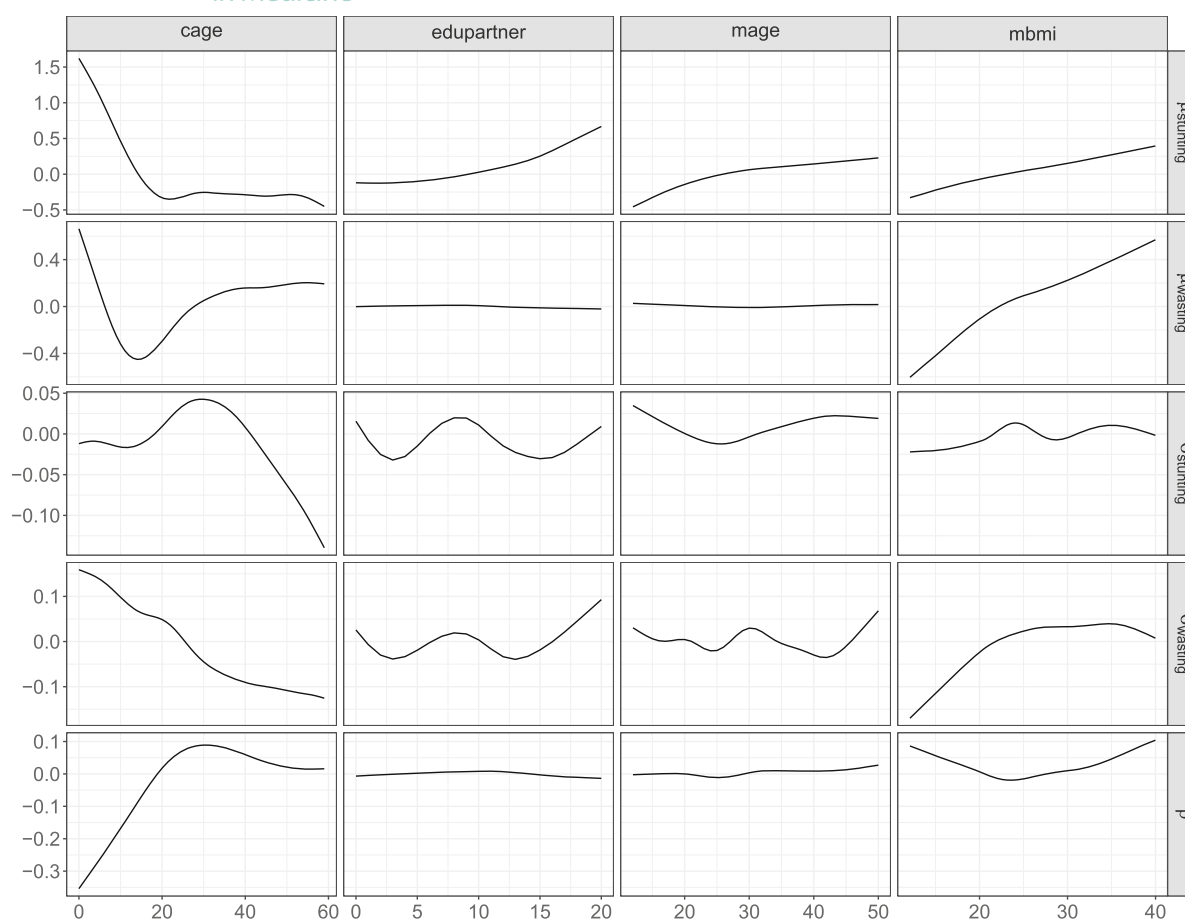


FIGURE 8 Non-linear effects of *cage*, *edupartner*, *mage* and *mbmi* for *stunting* and *wasting* of the bivariate Gaussian regression model for the Nigeria data

with normal nutrition. For wasting, *ctwin* had the largest effect, displaying also an increased risk for acute undernutrition. These results are in line with those of Klein et al.⁶⁴

Furthermore, *stunting* and *wasting* were both influenced by *cage* and *mbmi* as well. Following our model, for *mbmi*, a higher BMI of the mother indicates a higher acute and chronic undernutrition. For *cage*, *stunting* and *wasting* is estimated to decrease (ie, risk increases) up to around 20 months. After 20 months, the risk for *wasting* is estimated to decrease again while remaining similar for *stunting*.

The scale parameter for *wasting*, for example, indicates a higher variability for children up to around 25 months. For children older than 25 months, the variability decreases slightly, whereby we observed a greater variability for *stunting* between 20 and 40 months. The correlation is negative for children younger than 20 months and is approximately zero after a small positive correlation between 20 and 50 months. This finding indicates an interaction between *stunting* and *wasting* depending on the child's age, which is non-linear and stronger for younger children. Thus, children with a greater height in the first years of life have a lower weight for height and vice versa. The other covariates have only a minor estimated effect on the correlation parameter. These results are consistent with previous findings,^{11,64} which also holds for the spatial effects. The regional effect was selected to be informative for all distribution parameters. The effect of chronic undernutrition, for example, showed a lower risk of stunted growth in regions in southern Nigeria due to a positive effect. These regions also have a lower variability of chronic undernutrition compared to the average regions in the center of the county. This means that in this part of Nigeria the score for *stunting* is estimated to be on average lower and its variability is also smaller. By contrast, the regions in the north are estimated to have a higher risk for *stunting*. In terms of the correlation, some regions in the north are estimated to have a negative effect, while other regions in the south are

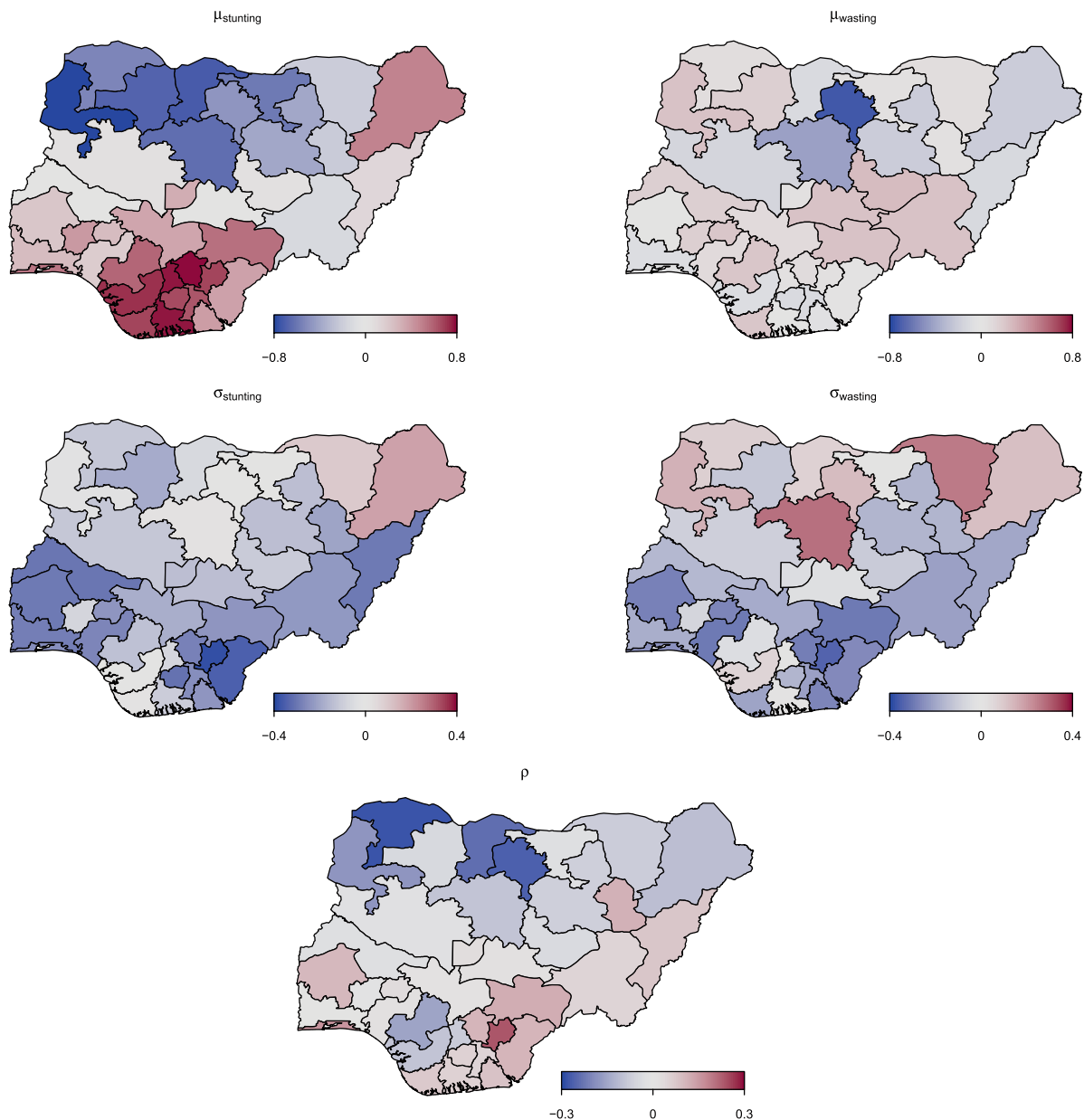


FIGURE 9 Spatial structure of *stunting* and *wasting* in Nigeria for the distribution parameters of the Gaussian distribution

estimated to show a slight positive effect on the correlation. A positive effect suggests that these regions have a problem of acute undernutrition as well as chronic undernutrition.

Overall, chronic undernutrition (*stunting*) is mostly affected by the living conditions of the children, for example, the birth order. Whereby, *stunting* and *wasting* were both influenced by the mother's BMI and particularly by the child's age. Additional effects of the covariates on the scale and correlation parameters also suggested greater uncertainty for younger children for acute undernutrition through a positive effect on the standard deviation, with variability decreasing with age. Furthermore, we observed a stronger negative correlation between *stunting* and *wasting* in younger children, that is, as *stunting* increases, *wasting* is expected to be lower. This means that children with a greater height tend to suffer from a lower weight for height at a younger age.

5 | DISCUSSION

We developed statistical boosting for modeling distributional regression with multivariate outcomes. Motivated by our biomedical applications, we considered three important multivariate parametric distributions: the bivariate Bernoulli, Poisson and Gaussian distributions. As special merits over classical maximum likelihood or Bayesian approaches to multivariate GAMLSS, our boosting framework can directly be used for high-dimensional data problems ($p > n$), while allowing for a data-driven variable selection mechanism that allows for sparse models for all parameters of a multivariate distribution.

In simulation studies, we have illustrated that the proposed boosting approach is able to identify the correct predictors in different data situations, including low- and high-dimensional settings and incorporating different effect types such as spatial effects. A comparison with the boosted univariate models showed that the bivariate models yielded more accurate estimates for the true structure of the effects. The wide applicability of our approach is illustrated on three different biomedical data sets, where we extend previous studies and also confirm findings from the literature. Applying our approach to examine jointly the genetic predisposition for chronic ischemic heart disease and high cholesterol not only provides information on the dependency of these phenotypes on the genetic variants, but also allows to identify the variants that affect the association between both phenotypes. This is in strong contrast to classical methods to estimate, for example, polygenic risk scores via accumulating effects from univariate linear models with single variants as predictor variables.⁶⁵ Our approach does not only incorporate multivariable predictor models, but also considers multivariate outcomes and hence allows to assess also the genetic predisposition for the association between several phenotypes, such as heart disease and high cholesterol. To the best of our knowledge, this is the first time multivariate distributional regression was adapted to model the joint genetic liability for multiple phenotypes.

In examining possible effects of patients characteristics on demand for health care, we found that age and income are relevant predictors, but also that gender affected the association between the number of doctor consultations and prescribed medications, with a stronger association found for women (cf. Karlis and Ntzoufras³⁰).

In the third application analyzing the risk of undernutrition in Nigeria, an association was found between chronic undernutrition and the child's living condition. In addition, the age of the child had a relevant influence on all distribution parameters related to chronic and acute undernutrition; furthermore, the regional effect was selected not only for the margins but also for the scale and correlation parameters.

To summarize, the application of our approach to boost multivariate distributional regression is particularly beneficial in settings where at least some of the following criteria are met: (i) multiple associated responses are of interest; (ii) the association or other characteristics of the joint distribution depend on covariates; (iii) there are multiple explanatory variables available without clear prior-knowledge; (iv) the aim of the analysis is exploratory, hypothesis-generating or prediction.

A limitation regarding the considered distributions in our approach is the restriction of the Poisson distribution to positive dependency between the two responses. A possible solution for this restriction in future research could be the use of alternative parameterization,³⁹ which also allow for modeling negative correlations; however, these have the disadvantage that the interpretation of effects on these parameters becomes much more difficult. A limitation of our algorithm is the relatively high selection rates for variables with only minor importance, which occurs particularly in low-dimensional settings. In this context, Strömer et al.⁶⁶ have recently proposed an approach to deselect predictors with negligible impact to obtain sparser models with statistical boosting. We want to investigate the incorporation of this proposal in the context of multivariate GAMLSS in the future. Moreover, as the number of distribution parameters and the complexity of the model increases (eg, due to many non-linear effects), the algorithm becomes computationally more intensive. To address this problem, also alternative approaches for early stopping could be considered. A promising approach in the future which has been developed for univariate location models is probing, where randomly shuffled versions of the original observed variables (probes) are added to the data set and the algorithm stopped when the first probe is selected.⁶⁷ Furthermore, a unique fixed step-length for all distribution parameters (as it is currently good practice in statistical boosting) could lead in some settings to an imbalance in the updates of predictors. In the most extreme case, the algorithm stops before some of the distribution parameters received an update at all. This could be tackled by scaling the gradient vectors or directly the outcome variable.⁵⁰ Zhang et al.⁴⁸ recently proposed an approach for adaptive step-lengths in Gaussian location and scale models to find the optimal step-length for the parameters. Further research is warranted on extending adaptive step-lengths approaches beyond Gaussian models to the more complex multivariate regression models.

Last, our focus has been on bivariate distributional regression models, but we will consider extending the models to higher dimensional responses in future research. From an algorithmic perspective, the extension should be

straight-forward as it only adds more distribution parameters in our proposed framework. However, not only the construction of appropriate response distributions but also the interpretation of the effect estimates becomes more challenging. For example, for the multivariate Gaussian distribution, the main challenge is the parameterization of the covariance matrix and a promising route here could be based on a (modified) Cholesky decomposition.⁶⁸ Similar, the extension of the bivariate Poisson distribution to higher dimensions has some difficulties due to the complicated form of the joint probability function. The most common extension would force all the pairs of variables to have the same covariance,⁶⁹ whereby Karlis and Meligkotsidou⁷⁰ already discussed a model with a two-way covariance term that allows for different covariances between the variables.

ACKNOWLEDGEMENTS

The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776, KL3037/2-1, MA7304/1-1). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The code used for the simulations and the biomedical applications is available at GitHub <https://github.com/AnnikaStr/DistRegBoost>. The genomic cohort data is available upon request from the UK Biobank at <https://www.ukbiobank.ac.uk/>. This research has been conducted using the UK Biobank resource under application number 81202. The data set for health care in Australia is openly available from Cameron and Trivedi at <http://cameron.econ.ucdavis.edu/racd/count.html>. The Nigeria data set may be accessed upon request from the Demographic and Health Survey (DHS, <https://dhsprogram.com/>).

ORCID

Annika Strömer  <https://orcid.org/0000-0002-1284-3318>

Nadja Klein  <https://orcid.org/0000-0001-5196-3374>

Christian Staerk  <https://orcid.org/0000-0003-0526-0189>

Hannah Klinkhammer  <https://orcid.org/0000-0003-3752-1275>

Andreas Mayr  <https://orcid.org/0000-0001-7106-9732>

REFERENCES

- Hastie T, Tibshirani R. *Generalized Additive Models*. 1st ed. London: Chapman & Hall; 1990.
- Wood SN. *Generalized Additive Models: An Introduction with R*. 2nd ed. London: Chapman & Hall/CRC; 2006.
- Koenker R. *Quantile Regression*. 1st ed. United Kingdom: Cambridge University Press; 2005.
- Offen W, Chuang-Stein C, Dmitrienko A, et al. Multiple co-primary endpoints: medical and statistical solutions: a report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of America. *Drug Inf J*. 2007;41(1):31-46.
- Zellner A. An efficient method of estimating seemingly unrelated regressions and test of aggregation bias. *J Am Stat Assoc*. 1962;57(298):348-368.
- Lang S, Adebayo SB, Fahrmeir L, Steiner WJ. Bayesian geoadditive seemingly unrelated regression. *Comput Stat*. 2003;18:263-292.
- King G. A seemingly unrelated Poisson regression model. *Sociol Methods Res*. 1989;17(3):235-255.
- Gallant A. Seemingly unrelated nonlinear regressions. *J Econom*. 1975;3(1):35-50.
- Fiebig DG. *Seemingly Unrelated Regression*. United Kingdom Oxford: Blackwell Publishers; 2001:101-121.
- Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc: Ser C (Appl Stat)*. 2005;54(3):507-554.
- Klein N, Kneib T, Klasen S, Lang S. Bayesian structured additive distributional regression for multivariate Responses. *J R Stat Soc. Ser C: Appl Stat*. 2015;64(4):569-591.
- Cole TJ. Commentary: Methods for calculating growth trajectories and constructing growth centiles. *Stat Med*. 2019;38(19):3571-3579.
- Papageorgiou AT, Ohuma EO, Altman DG, et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *Lancet*. 2014;384(9946):869-879.
- World Health Organization. *WHO Child Growth Standards: Length/Height for Age, Weight-for-Age, Weight-for-Height and Body Mass Index-for-Age, Methods and Development*. Geneva, Switzerland: World Health Organization; 2006.
- Hans N, Klein N, Faschingbauer F, Schneider M, Mayr A. Boosting distributional copula regression. *Biometrics*. 2022;1-13.
- Zhu B, Dunson DB, Ashley-Koch AE. Adverse subpopulation regression for multivariate outcomes with high-dimensional predictors. *Stat Med*. 2012;31(29):4102-4113.
- Wu C, Cui Y, Ma S. Integrative analysis of gene-environment interactions under a multi-response partially linear varying coefficient model. *Stat Med*. 2014;33(28):4988-4998.
- Liu H, Sunil RJ. Generalized finite mixture of multivariate regressions with applications to therapeutic biomarker identification. *Stat Med*. 2020;39(28):4301-4324.

19. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann Stat*. 2000;28(2):337-407.
20. Friedman J. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232.
21. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci*. 2007;22(4):477-505.
22. Mayr A, Fenske N, Hofner B, Kneib T, Schmid M. Generalized additive models for location, scale and shape for high dimensional data – a flexible approach based on boosting. *J R Stat Soc: Ser C (Appl Stat)*. 2012;61(3):403-427.
23. Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B. Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Stat Comput*. 2018;28:673-687.
24. Marshall AW, Olkin I. A family of bivariate distributions generated by the bivariate Bernoulli distribution. *J Am Stat Assoc*. 1985;80(390):332-338.
25. Kocherlakota S, Kocherlakota K. *Bivariate Discrete Distributions*. 1st ed. New York: Dekker; 1992.
26. Kotz S, Balakrishnan N, Johnson N. *Continuous Multivariate Distributions, Volume 1. Models and Applications*, 2nd ed. New York: John Wiley & sons; 2000.
27. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.
28. Burgess S, Thompson S. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. 1st ed. New York: Chapman and Hall/CRC; 2015.
29. Cameron A, Trivedi P. *Regression Analysis of Count Data*. 1st ed. Cambridge, UK: Cambridge University Press; 1998.
30. Karlis D, Ntzoufras I. Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. *J Stat Softw*. 2005;14(10):1-36.
31. Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data: a Bayesian perspective. *Stat Sin*. 2004;14(3):731-761.
32. Fahrmeir L, Kneib T, Lang S, Marx B. *Regression: Models, Methods and Applications*. 1st ed. Berlin: Springer-Verlag; 2013.
33. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci*. 1996;11(2):89-121.
34. Rue H, Held L. *Gaussian Markov Random Fields*. 1st ed. New York/Boca Raton: Chapman & Hall/CRC; 2005.
35. Johnson N, Kotz S, Balakrishnan N. *Discrete Multivariate Distributions*. 1st ed. New York: Wiley; 1997.
36. McCullagh P, Nelder J. *Generalized Linear Models*. Monographs on Statistics and Applied Probability Series. 2nd ed. London: Chapman and Hall/CRC; 1989.
37. Palmgren J. Regression models for bivariate binary responses. *UW Biostatistics Working Paper Series*. 1989 Working Paper 101.
38. Dale JR. Global cross-ratio models for bivariate, discrete, ordered Responses. *Biometrics*. 1986;42(4):909-917.
39. Lakshminarayana J, Pandit S, Rao KS. On a bivariate Poisson distribution. *Commun Stat Theory Methods*. 1999;28(2):267-276.
40. Ma Z, Hanson TE, Ho YY. Flexible bivariate correlated count data regression. *Stat Med*. 2020;39(25):3476-3490.
41. Freund Y. Boosting a weak learning algorithm by majority. *Inf Comput*. 1995;12(2):256-285.
42. Li Z, Luo Z, Sun Y. Robust nonparametric integrative analysis to decipher heterogeneity and commonality across subgroups using sparse boosting. *Stat Med*. 2022;41(9):1658-1687.
43. Wu M, Ma S. Robust semiparametric gene-environment interaction analysis using sparse boosting. *Stat Med*. 2019;38(23):4625-4641.
44. Tutz G, Binder H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*. 2006;62(4):961-971.
45. Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms. From machine learning to statistical modelling. *Methods Inf Med*. 2014;53(6):419-427.
46. Mayr A, Binder H, Gefeller O, Schmid M. Extending statistical boosting. An overview of recent methodological developments. *Methods Inf Med*. 2014;53(6):428-435.
47. Hofner B, Mayr A, Robinzonov N, Schmid M. Model-based boosting in R: a hands-on tutorial using the R package mboost. *Comput Stat*. 2014;29:3-35.
48. Zhang B, Hepp T, Greven S, Bergherr E. Adaptive step-length selection in gradient boosting for Gaussian location and scale models. *Comput Stat*. 2022;37:2295-2332.
49. Mayr A, Hofner B, Schmid M. The importance of knowing when to stop. A sequential stopping rule for component-wise gradient boosting. *Methods Inf Med*. 2012;51(2):178-186.
50. Hofner B, Mayr A, Schmid M. gamboostLSS: an R package for model building and variable selection in the GAMLSS framework. *J Stat Softw*. 2016;74(1):1-31.
51. Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*. 2008;17:211-235.
52. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1-3.
53. Yee TW. The VGAM package for categorical data analysis. *J Stat Softw*. 2010;32(10):1-34.
54. Wolodzko T. extraDistr: additional univariate and multivariate distributions. R package version 1.9.1. 2020.
55. Umlauf N, Klein N, Simon T, Zeileis A. bamls: a Lego toolbox for flexible Bayesian regression (and beyond). *J Stat Softw*. 2021;100:1-53. doi:10.18637/jss.v100.i04
56. Linsel-Nitschke P, Götz A, Erdmann J, et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease – A Mendelian randomisation study. *PLOS One*. 2008;3(8):1-9.
57. Richardson TG, Sanderson E, Palmer TM, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: a multivariable Mendelian randomisation analysis. *PLOS Med*. 2020;17(3):1-22.

58. Purcell S, Chang C. Plink 2.0. 2015 <https://www.cog-genomics.org/plink/2.0/>
59. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga Sci.* 2015;4(1):1-16.
60. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinform.* 2018;34(16):2781-2787.
61. Maj C, Staerk C, Borisov O, et al. Statistical learning for sparser fine-mapped polygenic models: the prediction of LDL-cholesterol. *Genet Epidemiol.* 2022;46(8):589-603.
62. Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Front Genet.* 2023;13:1-16. doi:10.3389/fgene.2022.1076440
63. Elosua R, Sayols-Baixeras S. The genetics of ischemic heart disease: from current knowledge to clinical implications. *Revista Española de Cardiología (English Edition).* 2017;70(9):754-762.
64. Klein N, Carlan M, Kneib T, Lang S, Wagner H. Bayesian effect selection in structured additive distributional regression models. *Bayesian Anal.* 2021;16(2):545-573.
65. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc.* 2020;15:2759-2772.
66. Strömer A, Staerk C, Klein N, Weinhold L, Titze S, Mayr A. Deselection of base-learners for statistical boosting – with an application to distributional regression. *Stat Methods Med Res.* 2022;31(2):207-224.
67. Thomas J, Hepp T, Mayr A, Bischl B. Probing for sparse and fast variable selection with model-based boosting. *Comput Math Methods Med.* 2017.
68. Pourahmadi M. Covariance estimation: the GLM and regularization perspectives. *Stat Sci.* 2011;26(3):369-387.
69. Karlis D. An EM algorithm for multivariate poisson distribution and related models. *J Appl Stat.* 2003;30(1):63-77.
70. Karlis D, Meligkotsidou L. Multivariate poisson regression with covariance structure. *Stat Comput.* 2005;15:255-265.

How to cite this article: Strömer A, Klein N, Staerk C, Klinkhammer H, Mayr A. Boosting multivariate structured additive distributional regression models. *Statistics in Medicine.* 2023;42(11):1779-1801. doi: 10.1002/sim.9699

3.3 Publication C: Modelling dependent censoring in time-to-event data by boosting copula regression

Strömer A, Klein N, Van Keilegom I, Mayr A. Modelling dependent censoring in time-to-event data by boosting copula regression. *Lifetime Data Analysis* 2025; 31: 994-1016.

<https://doi.org/10.1007/s10985-025-09674-x>

Supplementary information can be found at:

<https://doi.org/10.1007/s10985-025-09674-x>

Implementations are available on GitHub:

<https://github.com/AnnikaStr/CopBoostDepCens>



Modelling dependent censoring in time-to-event data using boosting copula regression

Annika Strömer¹ · Nadja Klein² · Ingrid Van Keilegom³ · Andreas Mayr¹

Received: 13 November 2024 / Accepted: 30 September 2025 / Published online: 21 October 2025
© The Author(s) 2025

Abstract

Survival analysis plays a pivotal role across disciplines, including engineering, economics, and social sciences—not just in biomedical research. In many of these applications, incomplete observations due to censoring are common, arising from limited follow-up periods, study dropouts, or administrative constraints. A standard assumption in such settings is that the censoring mechanism is independent of the survival process. This assumption primarily holds when censoring occurs at the end of the observation period. However, there may be dependence between event and censoring times. For example, if a patient’s health deteriorates and they withdraw due to poor prognosis, the time of censoring depends on their health status, leading to dependent censoring as sicker patients are censored earlier. To address such situations adequately in statistical analyses, we propose a model-based boosting approach using distributional copula regression. Our approach models the joint distribution of survival and censoring times by linking unknown marginal distributions through an unknown parametric copula. All distribution parameters of the resulting joint distribution are estimated simultaneously as functions of potentially different covariates. A key merit of the boosting approach is its data-driven variable selection, which is particularly important for such flexible models. Estimation remains feasible even for high-dimensional data with more covariates than observations, where classical estimation frameworks meet their limits. To investigate the performance of our method, we conduct a comprehensive simulation study, and demonstrate its practical application using a recent observational study analyzing the overall survival of patients with colon cancer. The data has a high proportion of right-censored observations without information on the cause of censoring.

Keywords Copula · Distributional regression · Gradient boosting · Survival analysis · Variable selection

Extended author information available on the last page of the article

1 Introduction

Time-to-event data analyses play a crucial role in various disciplines, including biostatistics (Collett 2003; Klein and Moeschberger 2005; Samoladas et al. 2010; Hassan et al. 2018; De Bin and Stikbakke 2023; Srujana et al. 2024; Stijven et al. 2024). Their focus is on studying the time to the occurrence of an event of interest, such as death, disease progression, treatment response or another specific endpoint. A common challenge is that patients in a study may not experience the event of interest within the study period or may be lost to follow-up before the event occurs due to study dropout. Consequently, censored observations are an inherent and natural characteristic of survival data. Most widely used approaches, such as the Kaplan-Meier estimator (Kaplan and Meier 1958) or the Cox proportional hazards model (Cox 1972), can handle time-to-event data with censored observations, including the most common form of right-censoring. In these approaches, a key assumption is that the survival time T and censoring time C are statistically independent conditional on the covariates. This is illustrated in the left part of Fig. 1, where C provides no information about T given the covariates \mathbf{X} and vice versa.

However, this assumption is questionable in many situations. For example, if patients withdraw from a medical trial because their health condition is deteriorating, approaches that assume independent censoring could lead to biased results (Huang and Zhang 2008). This is due to a direct link between the survival and censoring times as illustrated in case a) in the right of Fig. 1. Thus, if we assume that sicker patients drop out of the study due to poor health, censored patients are more likely to be sicker than the non-censored patients and the survival time of the patients may be overestimated. A further dependence can exist if it is caused by unobserved confounding variables, as indicated by case b) in the right of Fig. 1. Here, the unobserved variable U influences survival and censoring times. This could occur if a patient in a trial experiences side effects of a treatment that requires an alternative treatment, causing the patient to drop out of the study.

Several distinct approaches have been developed to account for dependent censoring. These include frailty models, which introduce random effects to capture unobserved heterogeneity, such as that caused by unmeasured covariates, and to induce dependence between survival and censoring times only indirectly through a shared frailty term (e.g., Huang and Wolfe 2002; Schneider et al. 2020). In contrast, copula-

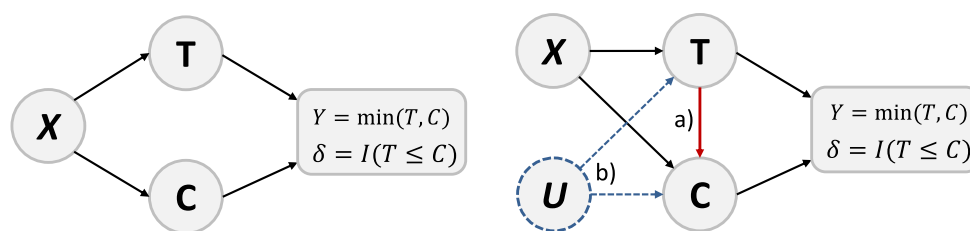


Fig. 1 Graphical illustration of two survival scenarios with right-censoring. In both scenarios, the event and censoring times T and C are conditional on the covariates \mathbf{X} . Furthermore, Y is the observed survival time, δ is the event indicator and U is an unobserved confounder. The left-hand graph displays independent censoring and the right-hand graph dependent censoring: a) direct dependence from T to C and b) indirect dependence through the unmeasured covariate U

based approaches have become increasingly popular for directly modeling the dependence between survival and censoring times (e.g., Emura and Chen 2016; Deresa and Van Keilegom 2020a, 2020b; Deresa et al. 2022; Midtjord et al. 2022; Czado and Van Keilegom 2023; Deresa and Van Keilegom 2024b). By specifying a joint distribution through a copula function and marginal models, these methods allow for more flexible and explicit dependence structures, going beyond the implicit modeling of frailty-based approaches. The identifiability of copula models under dependent censoring, however, is challenging because the right-censoring mechanism obviously prevents the simultaneous observation of (T, C) and thus obscures the direct relationship between T and C in the observed data (Tsiatis 1975; Crowder 1991). To ensure model identifiability, some authors assume that the copula is known. The first contribution in this direction was made by Zheng and Klein (1995), who proposed a generalization of the Kaplan-Meier estimator, the so-called copula-graphic estimator. This method was further explored by Rivest and Wells (2001) in the context of Archimedean copulas, assuming a known copula. Since then, numerous approaches have been developed in this vein (see among others, Braekers and Veraverbeke 2005; Huang and Zhang 2008; Chen 2010; Sujica and Van Keilegom 2018; Emura and Chen 2018; Deresa and Van Keilegom 2021). For instance, Chen (2010) extended the copula framework to semiparametric transformation models, using a known copula to characterize the association between survival and censoring times. While this method offers considerable flexibility for modeling marginal distributions, the assumption of a fully known copula function is often unrealistic in practical biostatistical applications. Czado and Van Keilegom (2023) recently proved that a known copula is not necessary for identifying the joint distribution of T and C . They proposed a model based on a parametric copula for the relationship between T and C along with parametric marginal distributions. However, a limitation is that covariates cannot be included into the model. Deresa et al. (2022) addressed this by exploring a copula-based method for bivariate data with left truncation and dependent right censoring, including multiple covariates but only for the margins.

To develop a framework that does not require the copula to be known, while allowing to incorporate covariates for all model parameters (in the spirit of generalized additive models for location, scale and shape, Rigby and Stasinopoulos 2005), we propose a model-based boosting approach for dependent censoring in survival analysis via distributional copula regression. Building on the work of Czado and Van Keilegom (2023) and Deresa et al. (2022), our method utilizes a parametric copula combined with arbitrary parametric marginal distributions. Estimation is carried out via statistical boosting, which facilitates the simultaneous modelling of all distribution parameters as functions of potentially different sets of covariates. Our boosting algorithm is based on the work of Hans et al. (2023), who proposed a general framework for boosting in distributional copula regression.

Our approach comes with three particular merits. First, the dependence between survival and censoring times can be modelled and explained through covariates. This may provide deeper insights and a better understanding of the underlying relationship between survival and censoring times. Second, our approach can handle high-dimensional cases where the number of covariates exceeds the number of observations ($p > n$). For such settings, most classical approaches are no longer feasible. Third,

the boosting approach has the advantage of including a data-driven variable selection mechanism. This feature is particularly beneficial when dealing with many potential predictors, ensuring that only the most relevant variables are included without compromising predictive power.

Using a recent observational study from oncology that investigates the overall survival of patients with colon cancer we demonstrate the merits of our approach in practical research. The routine data were gathered from a specialized cancer center and include potential predictors such as common clinical variables related to the tumor, patient demographics, and treatment (Seipp et al. 2021). This dataset represents a typical scenario encountered in biostatistical applications, characterized by a relatively high proportion of right-censored observations. Since the censoring mechanism is inherently unknown, it is impossible to determine whether the censoring is dependent or independent. However, in observational studies based on routine data, dependent censoring is more likely to occur due to various factors; for example, both patients with advanced disease and better health are more likely to miss follow-up visits or withdraw from the study (Howe et al. 2010). Nevertheless, the Cox proportional hazards model or a parametric accelerated failure time (AFT) model (Wei 1992; Heller 2024) are often still the methods of choice. We examine how the results of the analysis differ when accounting for associations between survival and censoring times, and how these associations vary with the covariates.

The structure of this article is as follows: Sect. 2 presents distributional copula regression for dependent censoring in time-to-event data and how to perform estimation based on model-based boosting. In Sect. 3, we investigate the performance of the new boosting approach in a simulation study with different scenarios—including conditional independent censoring. We apply our approach to colon cancer data and present the results in Sect. 4 before we finally discuss strengths and limitations of our approach in Sect. 5. Supplementary Materials (SM) contain further simulations and additional results for the application.

2 Model and Methods

2.1 Model specification

Let T and C be the survival and censoring times, respectively. Based on the assumption of random right-censoring, we observe $Y = \min(T, C)$ and $\delta = \mathcal{I}(T \leq C)$, where $\mathcal{I}(A) = 1$ if A is true and zero otherwise. Let furthermore $\mathbf{X} = \mathbf{x} \in \mathbb{R}^p$ denote the covariate vector. In the following, we allow for dependence between T and C and model this dependence utilizing a one-parameter copula $\mathcal{C}(\cdot, \cdot \mid \theta)$ with parameter $\theta \in \mathbb{R}$. We assume that the marginal distributions F_T and F_C are non-negative, continuous with parametric densities $f_T(\cdot \mid \theta_T)$ and $f_C(\cdot \mid \theta_C)$, respectively; and $\theta_T \in \mathbb{R}^{K_T}$, $\theta_C \in \mathbb{R}^{K_C}$ are the respective distribution parameters.

To account for the covariate information \mathbf{x} in a regression setting, we relate the vector of model parameters $\alpha = (\theta_T^\top, \theta_C^\top, \theta)^\top$ to covariates, that is, $\alpha \equiv \alpha(\mathbf{x})$. Then, the conditional version of Sklar's theorem (Patton 2006) allows to write the joint cumulative conditional distribution function (CDF) of T and C given \mathbf{x} as

$$F_{T,C}(t, c \mid \boldsymbol{\alpha}(\mathbf{x})) = \mathcal{C}\{F_T(t \mid \boldsymbol{\theta}_T(\mathbf{x})), F_C(c \mid \boldsymbol{\theta}_C(\mathbf{x})) \mid \boldsymbol{\theta}(\mathbf{x})\}.$$

In practice however, we do observe (Y, δ) and the following proposition is an extension of Theorem 3.3 in Deresa et al. (2022), where we also allow the copula parameter to depend on the covariates.

Proposition 1 *Let Y , F_T , F_C , f_T , f_C and $\boldsymbol{\alpha}$ be defined as above. Then, $F_Y(y \mid \boldsymbol{\alpha}(\mathbf{x})) = F_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})) + F_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})) - \mathcal{C}\{F_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})), F_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})) \mid \boldsymbol{\theta}(\mathbf{x})\}$ and*

$$f_Y(y \mid \boldsymbol{\alpha}(\mathbf{x})) = f_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})) [1 - h_{C|T}\{F_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})) \mid F_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})); \boldsymbol{\theta}(\mathbf{x})\}] \\ + f_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})) [1 - h_{T|C}\{F_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})) \mid F_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})); \boldsymbol{\theta}(\mathbf{x})\}] \quad (1)$$

holds for the distribution and density of Y , respectively, where $h_{C|T}\{F_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})) \mid F_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})); \boldsymbol{\theta}(\mathbf{x})\}$ and $h_{T|C}\{F_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})) \mid F_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})); \boldsymbol{\theta}(\mathbf{x})\}$ are the conditional distribution functions of $T \mid C$ and $C \mid T$, respectively. These can be expressed in terms of their associated copula as

$$h_{C|T}\{F_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})) \mid F_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})); \boldsymbol{\theta}(\mathbf{x})\} = \frac{\partial}{\partial u} \mathcal{C}\{u, v \mid \boldsymbol{\theta}(\mathbf{x})\} \big|_{u=F_T(t|\boldsymbol{\theta}_T(\mathbf{x})), v=F_C(c|\boldsymbol{\theta}_C(\mathbf{x}))} \\ h_{T|C}\{F_T(y \mid \boldsymbol{\theta}_T(\mathbf{x})) \mid F_C(y \mid \boldsymbol{\theta}_C(\mathbf{x})); \boldsymbol{\theta}(\mathbf{x})\} = \frac{\partial}{\partial v} \mathcal{C}\{u, v \mid \boldsymbol{\theta}(\mathbf{x})\} \big|_{u=F_T(t|\boldsymbol{\theta}_T(\mathbf{x})), v=F_C(c|\boldsymbol{\theta}_C(\mathbf{x}))}.$$

To ensure model identifiability, we focus on parametric margins and one-parameter copula functions, more precisely on log-normal and Weibull-distributed margins that are identifiable with the Clayton, Gaussian and Gumbel copulas, as was shown in Czado and Van Keilegom (2023).

2.1.1 Predictor specification

To allow for covariate dependent distribution parameters, we associate each element $\boldsymbol{\alpha}(\mathbf{x}) = (\theta_{T,1}(\mathbf{x}), \dots, \theta_{T,K_T}(\mathbf{x}), \theta_{C,1}(\mathbf{x}), \dots, \theta_{C,K_C}(\mathbf{x}), \boldsymbol{\theta}(\mathbf{x}))^\top \equiv (\alpha_1, \dots, \alpha_K)^\top$, $K = K_T + K_C + 1$ to structured additive predictors η_k via parameter-specific link functions g_k , such that $g_k(\alpha_k) = \eta_k$, $k = 1, \dots, K$. The general idea of these structured additive predictors is e.g., described in Wood (2017) and assumes that each η_k is of the form $\eta_k = \beta_{0k} + \sum_{j=1}^{p_k} s_{jk}(\mathbf{x}_{jk})$, where β_{0k} are the intercepts and the s_{jk} , $j = 1, \dots, p_k$ represent the p_k functional effects of parameter-specific covariate subvectors $\mathbf{x}_{jk} \subset \mathbf{x}$ in distribution parameter k . In this paper, we focus on linear effects that can be represented by $s_{jk}(\mathbf{x}_{jk}) = \mathbf{x}_{jk}^T \beta_{jk}$, where \mathbf{x}_{jk} is a covariate subset of \mathbf{x} for the parameter α_k and β_{jk} are the regression coefficients, but other effects, such as nonlinear effects of univariate continuous covariates or spatial effects can be cast into this framework (see again, e.g., Wood 2017).

2.2 Estimation via model-based boosting

2.2.1 Background on boosting

For estimation, we resort to model-based boosting. The concept of boosting originated from machine learning (Freund 1995; Freund and Schapire 1997), and was later adapted to statistical modelling (Friedman et al. 2000, 2001). This statistical view on boosting formed the foundation for component-wise gradient boosting with regression-type base-learners with all its extensions for various objectives (Bühlmann and Hothorn 2007), which was later also referred to as *statistical or model-based boosting* (Mayr et al. 2014, 2017). The basic idea is to iteratively fit regression-type base-learners one-by-one to the negative gradient of a pre-specified loss function, which in our case corresponds to the negative log-likelihood, and the base-learners are the different linear effects. In every iteration only the best-fitting base-learner is updated and added to the respective current regression predictor (see Hofner et al. (2014), for a detailed overview). This procedure is repeated until the final stopping iteration m_{stop} is reached and the result can be somewhat compared to that of L_1 penalized lasso regression (Tibshirani 1996; Hepp et al. 2016).

2.2.2 Estimation procedure

Assume that we have an independent and identical distributed sample $\mathcal{D} = \{(y_i, \delta_i, \mathbf{x}_i), i = 1, \dots, n\}$. Then, using (1), the joint log-likelihood is given by

$$\begin{aligned} \ell(\boldsymbol{\alpha}; \mathcal{D}) = & \sum_{\delta_i=1} \log(f_T(y_i | \boldsymbol{\theta}_T(\mathbf{x}_i)) [1 - h_{C|T}\{F_C(y_i | \boldsymbol{\theta}_C(\mathbf{x}_i)) | F_T(y_i | \boldsymbol{\theta}_T(\mathbf{x}_i)); \boldsymbol{\theta}(\mathbf{x}_i)\}]) \\ & + \sum_{\delta_i=0} \log(f_C(y_i | \boldsymbol{\theta}_C(\mathbf{x}_i)) [1 - h_{T|C}\{F_T(y_i | \boldsymbol{\theta}_T(\mathbf{x}_i)) | F_C(y_i | \boldsymbol{\theta}_C(\mathbf{x}_i)); \boldsymbol{\theta}(\mathbf{x}_i)\}]). \end{aligned}$$

Our model is integrated into the boosting distributional copula regression framework of Hans et al. (2023). Here, all distribution parameters are modelled simultaneously, i.e., in every iteration all partial derivatives $u_k = -\partial\ell/\partial\alpha_k$ of the negative log-likelihood $-\ell$ with respect to the different distribution parameters α_k are calculated and each base-learner $b_{jk} \equiv \mathbf{x}_{jk}^\top \beta_{jk}$ is separately fitted to the respective gradients. For each distribution parameter, the best-fitting base-learner b_{jk}^* is identified, i.e., the one with the highest loss reduction, and compared among all distribution parameters. Only a small proportion (step-length ν) of the overall best-performing base-learner $\nu \times b_{jk}^*$ is added to the respective predictor η_k in every iteration, using a non-cyclic version of the basic boosting algorithm (see Thomas et al. 2018, for details). The step-length ν is set to a small fixed value within the range of $0 < \nu < 1$ (Schmid and Hothorn 2008). For boosting copula regression, Hans et al. (2023) suggest a value of $\nu = 0.01$ and we follow this default choice.

2.2.3 Benefits of boosting

Boosting can handle high-dimensional data (that is, $p > n$) and is thus an attractive alternative to classical inference methods for statistical models in such situations. The algorithm selects only one base-learner in each iteration, building up predictors step-by-step. At each step, the base-learner that achieves the largest empirical risk reduction is updated, allowing its effect to accumulate over iterations. Base-learners contributing most to risk minimization are selected repeatedly, while those never selected are effectively excluded from the final model. This data-driven variable selection, conducted simultaneously across all additive predictors, is controlled by the stopping iteration m_{stop} . The number of boosting iterations m_{stop} determines the complexity of the final model; a higher number of iterations may include more variables, some with negligible effects that may reflect noise rather than signal. To balance sparsity and prediction accuracy, m_{stop} is typically optimized by cross-validation, resampling techniques, or using an additional data set (if available). The latter usually yields comparable stability in the selected stopping iteration and resulting model at lower computational cost than cross validation or resampling. In addition to encouraging the sparsity of the resulting models, the optimization of m_{stop} also leads to the prevention of overfitting, since the algorithm usually stops before convergence (also referred to as early stopping, Mayr et al. (2012)). For enhanced sparsity, post-hoc refinement via stability selection (Meinshausen and Bühlmann 2010) or deselection (Strömer et al. 2022, 2025) can be applied.

We denote our approach *CopBoostDepCens* in the remainder of the paper.

3 Simulations

We conducted a detailed simulation study to evaluate the performance of CopBoostDepCens with regard to the following questions:

- Can CopBoostDepCens correctly estimate the corresponding distribution parameters and identify the respective active variables?
- How do different censoring rates influence the results?
- How does CopBoostDepCens compare to a model that assumes independent censoring?
- How does CopBoostDepCens perform in situations where the survival and censoring times are actually independent?

3.1 Simulation design

To represent lower, no and upper tail dependence scenarios, we generate data using the Clayton, Gaussian and Gumbel copulas. To ensure identifiability, we combine these with either log-normal or Weibull-distributed margins for both survival and censoring times. We denote the copula parameter with ρ , whereby we use ρ^* and ρ^\diamond to distinguish between different copula parameters for specific scenarios. Additionally, $\mu_T, \sigma_T, \mu_C, \sigma_C$ represent the location and scale parameters of the marginal

distributions of T and C , respectively. The corresponding true predictor specifications $\eta, g, \cdot \in \{\mu_T, \sigma_T, \mu_C, \sigma_C, \rho\}$ are

$$\begin{aligned} g_{\mu_T}(\mu_T) &= \eta_{\mu_T} = \beta_{0\mu_T} + 2x_1 + x_3 & g_{\sigma_T}(\sigma_T) &= \eta_{\sigma_T} = 0.7 + 0.7x_3 \\ g_{\mu_C}(\mu_C) &= \eta_{\mu_C} = \beta_{0\mu_C} - x_2 + 1.5x_4 & g_{\sigma_C}(\sigma_C) &= \eta_{\sigma_C} = 0.5x_2 \\ g_{\rho^*}(\rho^*) &= \eta_{\rho^*} = 2 + 1.5x_5 & g_{\rho^\diamond}(\rho^\diamond) &= \eta_{\rho^\diamond} = 0.25 + 0.4x_1 - 0.6x_5, \end{aligned}$$

where $(\beta_{0\mu_T}; \beta_{0\mu_C}) \in \{(-1; 0.8), (0.7; 0.8), (1; -0.4)\}$ for Weibull-distributed margins and $(\beta_{0\mu_T}; \beta_{0\mu_C}) \in \{(-0.9; 0.8), (1; 0.8), (1.5; -0.4)\}$ for log-normal distributed margins. These values are chosen to mimic different average proportions of censoring of 20%, 50% and 80%, respectively. The covariates x_1, \dots, x_p are independently drawn from uniform distributions on $(-1, 1)$. We fix the number of observations to $n = 1000$ but vary the number of available covariates across all model components. Specifically, we define p^* as the total number of available covariates included across all additive predictors (i.e., across all model parameters), whereas p_k denotes the number of covariates used in the k -th predictor (e.g. for one marginal distribution or the copula parameter). Since each of the $K = 5$ distribution parameters in our model has its own additive predictor, the total number of covariates is $p^* = \sum_{k=1}^K p_k$. In our simulations, the same set of covariates is used for all distribution parameters. We consider $p^* \in \{50, 250, 500, 1000, 2500\}$, representing different levels of noise variables without any influence. The high-dimensional case, $p^* = 2500$ is evaluated only in Setting 1. The following settings are examined:

Setting 1 (dependent censoring, positive association)

This setting evaluates all combinations of margins and copulas, with the copula parameter ρ modelled by $g_{\rho^*}(\rho^*) = \eta_{\rho^*}$. The average dependence between the margins, as measured by Kendall's τ , ranges from $[0.45; 0.94]$ for the Clayton copula, $[0.31; 0.96]$ for the Gaussian copula, and $[0.39; 0.97]$ for the Gumbel copula (depending on the realization of covariate x_5).

Setting 2 (dependent censoring, weaker and negative association)

This setting specifically explores the Gaussian copula, with the copula parameter ρ modelled by $g_{\rho^\diamond}(\rho^\diamond) = \eta_{\rho^\diamond}$. This setting allows for negative dependence and exhibits lower dependence values compared to Setting 1. On average, Kendall's τ is within $[-0.43; 0.63]$.

Setting 3 (no association between censoring and survival times)

In the independent setting, where $g_{\rho}(\rho) = \eta_{\rho} = 0$ (no dependent censoring) and thus the copula parameter ρ does not depend on any covariates, we focus on the case with $p = 10$ covariates. Further results can be found in SM A and B.

In all settings, only the additive predictor for ρ changes between the settings, while the other additive predictors for the marginal distributions remain the same. For each setting we generate 100 replicated data sets using the R package `copula`. The corresponding link functions for the marginal distributions and copulas are listed in

Table 1. The runtime for all settings can be found in SM A and B.

3.1.1 Benchmark methods

We benchmark CopBoostDepCens with a similar model for survival analysis, namely the distributional AFT model (referred to in the following as AFT model), which does not take dependent censoring into account. The AFT model is fully parametric and allows for direct modelling of both the location and scale parameters as functions of covariates (in the spirit of generalized additive models for location, scale and shape, Rigby and Stasinopoulos 2005). We also estimate the AFT models with gradient boosting and the same settings and predictor specifications as CopBoostDepCens. As CopBoostDepCens also considers AFT-type distributions as margins, the distributional AFT model hence represents a very similar approach to ours, only that it focuses on the survival times T assuming independent censoring. The most popular model for survival analysis would be the Cox model. Due to its semi-parametric nature with unspecified baseline hazards, we do not consider it to be a fair competitor. However, additional results for the Cox models are available in SM A and B.

3.1.2 Performance metrics

To evaluate the predictive performance, we generate an additional test data set of equal sample size to assess the following metrics for each setting and method: Brier score and integrated absolute error. Let $S(t|x_i) = P(T > t|x_i)$ represent the true survival function at time t , indicating the probability that an individual with covariate vector x_i survives beyond time t . The predicted survival function, derived from the model, is denoted with $\hat{S}(t|x_i)$.

The Brier score (BS) measures the accuracy of the predicted survival function $\hat{S}(t|x_i)$ and thus the calibration of the model by calculating the mean squared difference between predicted probabilities and actual outcomes. For right-censored data, the Brier score at time $t = Y_1, \dots, Y_n$ is computed as:

$$\text{BS}(t) = \frac{1}{n} \sum_{i=1}^n \left[\frac{\hat{S}(t|x_i)^2 \cdot \mathcal{I}(Y_i \leq t, \delta_i = 1)}{\hat{G}(Y_i)} + \frac{(1 - \hat{S}(t|x_i))^2 \cdot \mathcal{I}(Y_i > t)}{\hat{G}(t)} \right],$$

Table 1 Link functions for the considered marginal distributions and copulas

	Parameter	Link function	
<i>Margins</i>			
Log-normal	Location	Identity	$g_{\mu_T}(u) = u, g_{\mu_C}(v) = v$
	Scale	Log	$g_{\sigma_T}(u) = \log(u), g_{\sigma_C}(v) = \log(v)$
Weibull	Scale	Log	$g_{\mu_T}(u) = \log(u), g_{\mu_C}(v) = \log(v)$
	Shape	Log	$g_{\sigma_T}(u) = \log(u), g_{\sigma_C}(v) = \log(v)$
<i>Copulas</i>			
Clayton	Dependence	Log	$g_{\rho}(\rho) = \log(\rho)$
Gaussian	Dependence	Inverse hyperbolic tangent	$g_{\rho}(\rho) = \tanh^{-1}(\rho)$
Gumbel	Dependence	Log shifted by 1	$g_{\rho}(\rho) = \log(\rho - 1)$

where $\hat{G}(t)$ is an estimate of the survival function $P(C > t)$ for the censoring time, defined as the probability that the censoring time C exceeds time t . The latter is typically estimated using the Kaplan-Meier estimator (Kaplan and Meier 1958; Graf et al. 1999), which is applied to the AFT model. However, with CopBoostDepCens, censoring times can directly be estimated from the model, which accounts for covariates, unlike the Kaplan-Meier estimator. Thus, for CopBoostDepCens, we replace \hat{G} by $\hat{G}(t|\mathbf{x}_i)$, which estimates $P(C > t|\mathbf{x}_i)$. This is similar to the approach of Lillelund et al. (2025), which replaces the Kaplan-Meier estimator with a copula-graphic estimator for estimating $G(t)$. Lower values for the Brier score indicate greater accuracy.

The integrated absolute error (IAE) measures the absolute error over time, weighting all errors equally and is defined by

$$\text{IAE} = n^{-1} \sum_{i=1}^n \int_0^{t_{\max}} |S(t|\mathbf{x}_i) - \hat{S}(t|\mathbf{x}_i)| dt$$

where $S(t|\mathbf{x}_i)$ denotes as before the true survival function and $\hat{S}(t|\mathbf{x}_i)$ represents the predicted survival function at time t (Moradian et al. 2017). The largest value among Y_1, \dots, Y_n is denoted as t_{\max} . Results for the integrated Brier score and integrated squared error can be found in SM A and B.

Note that most standard evaluation metrics were originally developed under the assumption of independent censoring. When applied to settings with dependent censoring, they should be interpreted with caution. We therefore focus on the Brier score, adjusted to incorporate censoring, and the IAE, which remains valid regardless of the censoring mechanism.

3.1.3 Tuning and implementation details

The CopBoostDepCens models have been implemented as an add-on to the R package gamboostLSS, where also the AFT model is implemented. The stopping iteration m_{stop} is optimized by minimizing the empirical risk on an additional validation data set with $n_{\text{val}} = 1000$ observations. The step-length is set to $\nu = 0.01$ for CopBoostDepCens and the AFT model for all distribution parameters (cf., Hans et al. 2023). All code required to replicate our results is available on GitHub: <https://github.com/AnnikaStr/CopBoostDepCens>.

3.2 Results

In the following, we present the results for Weibull margins and the Gaussian copula for all settings. Results for the Clayton and Gumbel copulas for Setting 1 and Setting 2 with Weibull margins can be found in SM, Section A, whereas results with log-normal margins and all three copulas are given in Section B.

3.2.1 Setting 1 (dependent censoring, positive association)

Figure 2 presents boxplots of coefficient estimates of the informative and non-informative (non-inf.) variables of CopBoostDepCens on the 100 simulation replicates. Results are shown for each distribution parameter (columns) of the Gaussian copula with Weibull-distributed margins for different numbers of noninformative variables (rows 1–5). The box colors represent the average proportions of censoring. The red horizontal lines show the true values of each corresponding coefficient. Missing boxplots indicate that the algorithm did not select the corresponding variables.

CopBoostDepCens effectively captures the true structure of informative variables for each parameter of the margins reasonably well, even with a large number of noise variables and high censoring rates. For instance, in the high-dimensional case of $p^* = 2500$, where $p^* > n$, the algorithm performs comparably to lower-dimensional settings in identifying informative variables. Only a slightly stronger shrinkage of effect estimates is observed. Estimates for non-informative variables are consistently shrunk to zero, with fewer false positives as p^* increases. The true effects for the dependence parameter ρ are also correctly identified but show stronger shrinkage as p^* increases. In addition, x_1 is occasionally falsely selected for 20% and 50% censoring, with lower rates at 80% (see SM: Table A3).

Figure A1 (SM A.1.1) presents the coefficient estimates for the benchmark method: The truly informative variables for the survival time are correctly identified for both distribution parameters of the AFT models, but slightly overestimate their coefficients and include some noise variables, particularly those informative for censoring. Table 2 (rows 1–4) shows the means and standard deviations (SD) of the Brier score and IAE for performance evaluation. CopBoostDepCens consistently outperforms the AFT model in Brier scores, indicating better calibration and prediction accuracy. For the IAE, the AFT model performs best at 20% censoring, but CopBoostDepCens maintains stable results across varying covariates and censoring rates and performs best at 50% and 80% censoring. This can be explained by the fact that unlike the AFT model, CopBoostDepCens needs to estimate the censoring time, which is difficult when only 20% of the data is censored. Additionally, IAE is also calculated for the censoring time.

3.2.2 Setting 2 (dependent censoring, weaker and negative association)

Figure A2 in SM A.1.2 shows the coefficient estimates for CopBoostDepCens of Setting 2. The parameter estimates of the margins closely resemble those from Setting 1, despite lower and partly negative dependence, indicating correct identification of the informative variables. For the dependence parameter, the informative variables are also correctly detected with a slight overestimation of the intercept at 50% censoring and for the coefficient of x_1 at 20% censoring. This overestimation is expected because x_1 has a strong effect on μ_T and with 20% censoring, more events are observed, leading to a slightly overestimated effect on the dependence parameter. For the AFT model, only minor differences are observed: The AFT model correctly identifies informative variables for both distribution parameters, slightly overestimating the effect for x_1 , which improves with more covariates (see SM A.1.2, Fig. A3). Cop-

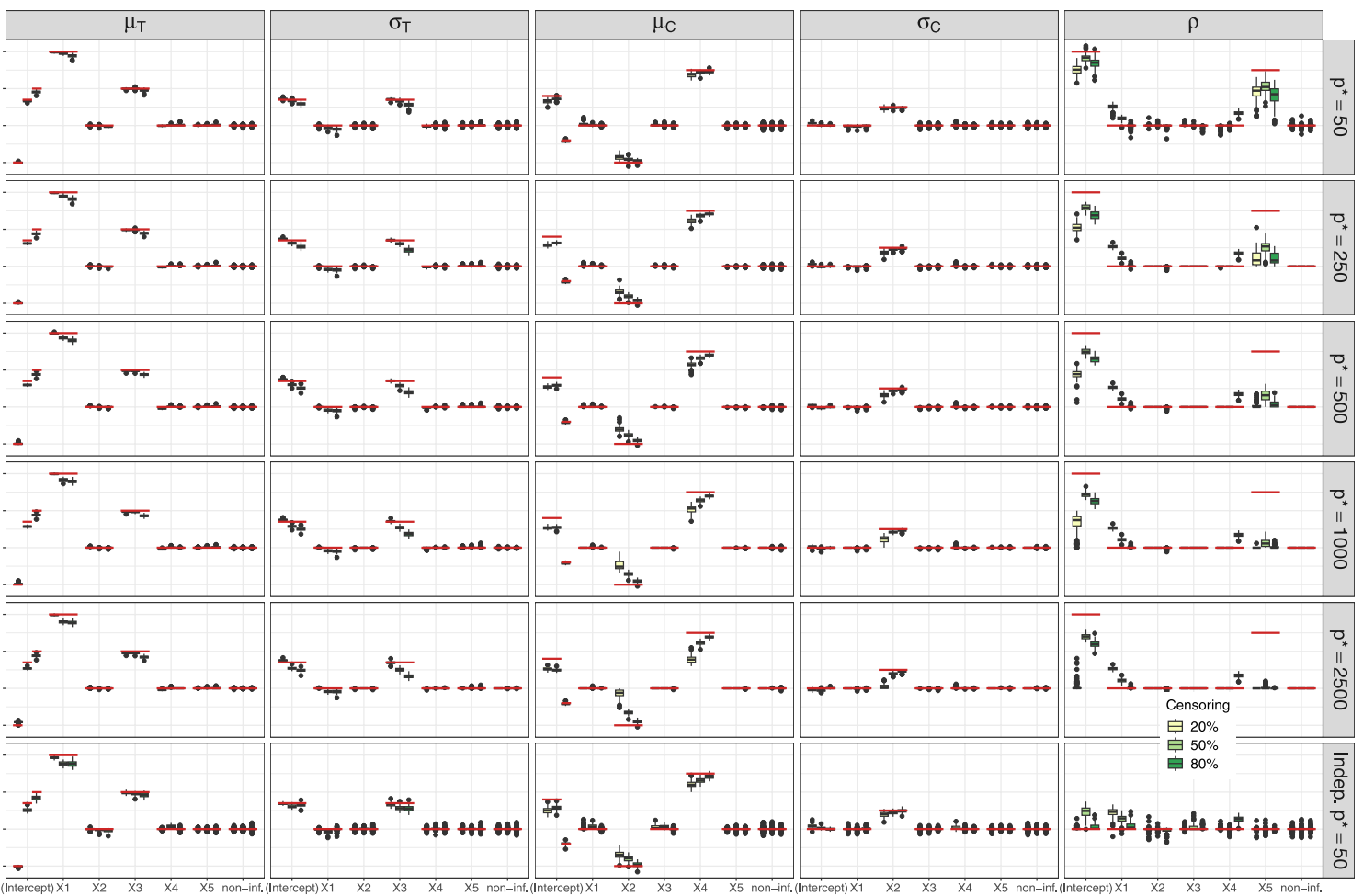


Fig. 2 Simulation study. Boxplots of estimated coefficients of CopBoostDepCens on the 100 replicates. Results are shown for each distribution parameter (columns) of the Gaussian copula with Weibull-distributed margins for different numbers of noise variables (rows 1–5, Setting 1) as well as for Setting 3 with $p = 10$ (row 6). The box colors represent the average proportions of censoring. The red horizontal lines show the true values for each

Table 2 Simulation study

		Brier score		Integrated absolute error		
				Survival time		Censoring time
		Censoring (%)		Copula	AFT	Copula
$p^* = 50$	20	0.07 (0.00)	0.10 (0.13)	0.94 (0.09)	0.13 (0.02)	0.88 (0.12)
	50	0.08 (0.01)	0.13 (0.18)	1.13 (0.05)	1.84 (0.26)	2.05 (0.13)
	80	0.06 (0.02)	0.10 (0.16)	1.23 (0.16)	3.87 (1.00)	1.41 (0.19)
$p^* = 250$	20	0.07 (0.00)	0.11 (0.16)	0.94 (0.08)	0.13 (0.02)	0.84 (0.10)
	50	0.09 (0.02)	0.12 (0.16)	1.15 (0.05)	1.59 (0.23)	1.99 (0.11)
	80	0.06 (0.01)	0.11 (0.18)	1.29 (0.19)	2.89 (0.81)	1.37 (0.19)
$p^* = 500$	20	0.07 (0.00)	0.10 (0.14)	0.93 (0.08)	0.12 (0.02)	0.83 (0.11)
	50	0.10 (0.02)	0.13 (0.17)	1.19 (0.06)	1.42 (0.21)	2.02 (0.12)
	80	0.06 (0.02)	0.08 (0.09)	1.30 (0.18)	2.30 (0.55)	1.34 (0.14)
$p^* = 1000$	20	0.07 (0.00)	0.10 (0.14)	0.93 (0.07)	0.13 (0.02)	0.87 (0.10)
	50	0.10 (0.01)	0.13 (0.18)	1.24 (0.08)	1.27 (0.20)	2.03 (0.13)
	80	0.06 (0.01)	0.10 (0.15)	1.34 (0.18)	1.90 (0.51)	1.37 (0.17)
$p^* = 2500$	20	0.07 (0.00)	0.11 (0.15)	0.91 (0.08)	0.13 (0.02)	1.00 (0.13)
	50	0.07 (0.00)	0.11 (0.15)	1.32 (0.08)	1.16 (0.16)	2.07 (0.13)
	80	0.06 (0.01)	0.10 (0.12)	1.35 (0.20)	1.57 (0.41)	1.35 (0.19)
Indep.	20	0.07 (0.00)	0.12 (0.15)	0.88 (0.09)	0.04 (0.01)	0.81 (0.13)
$p^* = 50$	50	0.08 (0.01)	0.17 (0.21)	1.42 (0.19)	0.25 (0.06)	1.99 (0.21)
	80	0.05 (0.02)	0.11 (0.10)	1.18 (0.32)	0.39 (0.16)	1.26 (0.23)

Brier score (SD) and integrated absolute error (SD) for CopBoostDepCens and AFT models on the 100 replicates of the Gaussian copula with Weibull-distributed margins for different numbers of noise variables (rows 1–5, Setting 1) as well as for Setting 3 with $p^* = 50$ (row 6). The best-performing model for each metric is highlighted in bold

BoostDepCens achieves the lowest Brier scores, while the AFT model consistently performs better in IAE, which, as for Setting 1, can be attributed to the fact that CopBoostDepCens needs to estimate the censoring time (Tables A4 and A5, SM A.1.2).

3.2.3 Setting 3 (no association between censoring and survival times)

Setting 3 investigates how the model performs under conditional independent censoring, where the censoring times depend only on known covariates but not on the survival times. Similar to the observations in the dependent settings, CopBoostDepCens performs well in identifying the informative variables across different censoring rates for both survival and censoring time (see Fig. 2; row 6). However, the model includes some noise variables in the dependence parameter. The AFT model provides rather accurate estimates without overestimation (Fig. A4, SM A.1.3). This is also reflected in the IAE in Table 2, which are consistently lowest for the AFT model and do not increase as much with higher censoring as in Setting 1. The values for CopBoostDepCens are similar to the dependent settings, including for the censoring time. CopBoostDepCens remains superior for the Brier score, showing no noticeable change from the dependent settings.

3.3 Overall summary of simulation results

CopBoostDepCens demonstrates a favorable performance in both identifying and estimating informative variables across various scenarios. It effectively identifies the relevant variables for each distribution parameter, even under more challenging conditions characterized by weaker and negative dependence. CopBoostDepCens consistently captures the true structure of the data across different levels of censoring and thus provides robust results for the considered censoring rates. In addition, performance remains reliable even when survival and censoring times are truly independent.

In terms of predictive performance, CopBoostDepCens also performs well. When compared to a model that assumes independent censoring, such as the AFT model, the results are partially comparable, particularly for the integrated absolute error which focuses on survival time alone. However, the AFT model is limited by its exclusive focus on survival time, whereas CopBoostDepCens accounts for both survival and censoring times, as well as their potential dependence. In scenarios with only low censoring rates and/or weak dependencies between T and C , the AFT model seems to be suitable if the focus is solely on survival time. However, the strength of CopBoostDepCens lies in its ability to capture complex relationships between survival and censoring time that simpler models may miss, providing a broader perspective that is particularly valuable in scenarios that require a deeper understanding of their interplay. Note that, interestingly, our approach exhibits a decreased runtime in high-dimensional settings, as overfitting tends to occur earlier, forcing the algorithm to stop early.

4 Observational study on survival of colon cancer patients

We illustrate the application of our method via an observational study investigating the overall survival of colon cancer patients after surgery based on routine data.

4.1 Data and model building

We analyse data that contain information on $n = 546$ patients listed in a registry of a local German acute care hospital. All enrolled patients underwent the surgical resection of the affected part of the intestine with radical regional excision of adjacent lymph node stations, following the corresponding guidelines (Seipp et al. 2021). The data are publicly available in the R package *dirtree* (Seipp and Otto-Sobotka 2022).

In what follows, we focus on the outcome of overall survival since surgery. The event was observed for 201 patients, while 345 patients were right-censored. Median follow-up time was 26.6 months, for a histogram of the follow-up times for patients with and without event, see Fig. 3. The high censoring rate (63.1%) is a most likely result of the relatively good prognosis for colon cancer patients and the short follow-up times. There is no information, however, on the individual causes of censoring, but as this is an observational study, it is very likely that patients drop out of the study for a reason related to their survival time. The following clinical variables are con-

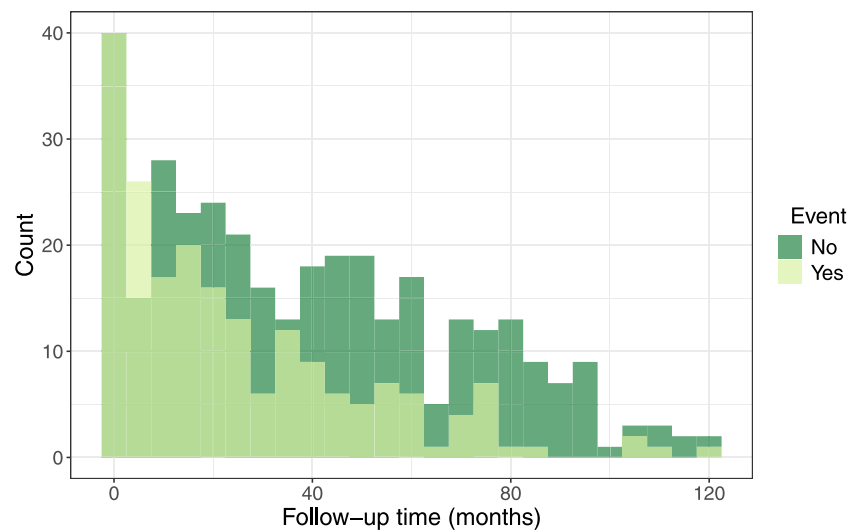


Fig. 3 Observational study. Histogram of the follow-up time in months for patients with event (light green) and without event (dark green)

sidered: chemotherapy (yes or no), ASA score (general health status, mild or severe), UICC cancer stage (I–IV, higher stage means a further progressed tumor), age of the patient, LNE (number of pathologically examined lymph nodes during surgery), LNR (lymph node ratio, number of cancerous lymph nodes divided by the number of examined lymph nodes), sex, R status (residual tumor after surgery, yes or no) and preexisting cancer (yes or no). A more detailed description of the data can be found in Seipp et al. (2021).

4.1.1 Model selection

Our aim is to model the distribution of observed events and censoring times of the patients as functions of all available clinical variables as potential predictors using CopBoostDepCens. We tested Weibull and log-normal marginals in combination with the Clayton, Gaussian and Gumbel copulas. To select the best model, we evaluated the predictive log-likelihood score based on 10-fold cross-validation (cf. Hans et al. 2023) (see SM C). The best stopping iteration m_{stop} for each model was also determined based on 10-fold cross-validation.

4.1.2 Benchmarks

In line with the approach used in the simulation study, we compare CopBoostDepCens to a distributional AFT model, which does not account for potential dependent censoring but is also estimated via boosting (as in the simulations). Additionally, we include a boosted Cox model in the comparison, as it is the most traditional and widely utilized method for analyzing survival data. This evaluation serves as a sanity check rather than a competitive comparison. The R-code to reproduce all analyses can be found on GitHub <https://github.com/AnnikaStr/CopBoostDepCens>.

4.2 Results

The best-performing combination of marginal distributions and copula with respect to predictive performance was the Weibull distribution with a Clayton copula. We hence also used the Weibull distribution for the AFT model.

Table 3 displays the estimated coefficients for CopBoostDepCens, the Cox and AFT models. One can observe at first glance that many of the clinical variables provided in the dataset are identified as relevant predictors for the survival of colon cancer patients by the boosting approaches. The resulting models are not particularly sparse. For example, for our copula model all variables are selected for the location parameter of the survival time μ_T . This is also confirmed by the Cox and AFT models, where almost all variables have been selected for the hazards or the location parameter μ , respectively. When interpreting the coefficients, one has to keep in mind that the copula and the AFT model rely on modelling directly the event time, while the Cox model focuses on the hazards as quantity of interest. As a result, for example the age of a patient has a negative effect on μ_T in the copula model and μ in the AFT model but a positive (multiplicative) effect for the Cox model. Both refer hence to a negative impact of age on the survival of patients with colon cancer. When just looking at the survival part of the copula model, one can again (cf., Section 3) observe the similarities of CopBoostDepCens and the AFT model when utilizing the same marginal distribution: the resulting estimated coefficients for μ_T and μ as well as σ_T and σ are overall very similar regarding magnitude and direction of effects.

The strength of CopBoostDepCens is to take the association (ρ) between the survival and the censoring time into account. Compared to the distribution of event times, less variables are relevant for modelling the distribution of censoring times and the association between event and censoring times. In particular for the dependence parameter, only two variables are selected: chemotherapy and tumor stage IV. This provides evidence for dependent censoring, influenced also by covariates. While chemotherapy is in general associated with longer survival times (positive effect on

Table 3 Observational study

	Copula model					AFT model		Cox model
	μ_T	σ_T	μ_C	σ_C	ρ	μ	σ	exp(coef)
Intercept	6.070	-0.270	3.061	-0.428	0.546	6.734	-0.329	0.094
Age	-0.016	-0.001	0.007	0.002	-	-0.019	-	1.021
Sex, male	-0.033	0.071	0.006	-0.081	-	-0.017	0.118	-
Chemotherapy	0.540	0.592	0.117	0.186	-1.170	0.334	0.752	0.500
ASA score, severe	-0.693	-0.195	-	-0.143	-	-0.728	-0.242	3.259
UICC cancer stage II	-0.092	-	-	0.029	-	-	-	1.373
stage III	-0.393	0.052	0.111	0.231	-	-0.513	-	2.100
stage IV	-1.081	-0.147	-	-	-0.250	-1.349	-0.228	3.175
LNE	0.009	0.008	0.007	0.011	-	0.006	0.007	0.981
LNR	-1.616	0.434	0.560	-	-	-1.679	0.387	3.702
R status	-0.153	-0.014	0.220	0.519	-	-0.063	-0.033	2.424
Preexisting cancer	-0.398	-0.020	-0.060	-0.014	-	-0.462	-	1.524

Estimated coefficients for all clinical variables with CopBoostDepCens, the AFT model and the Cox model

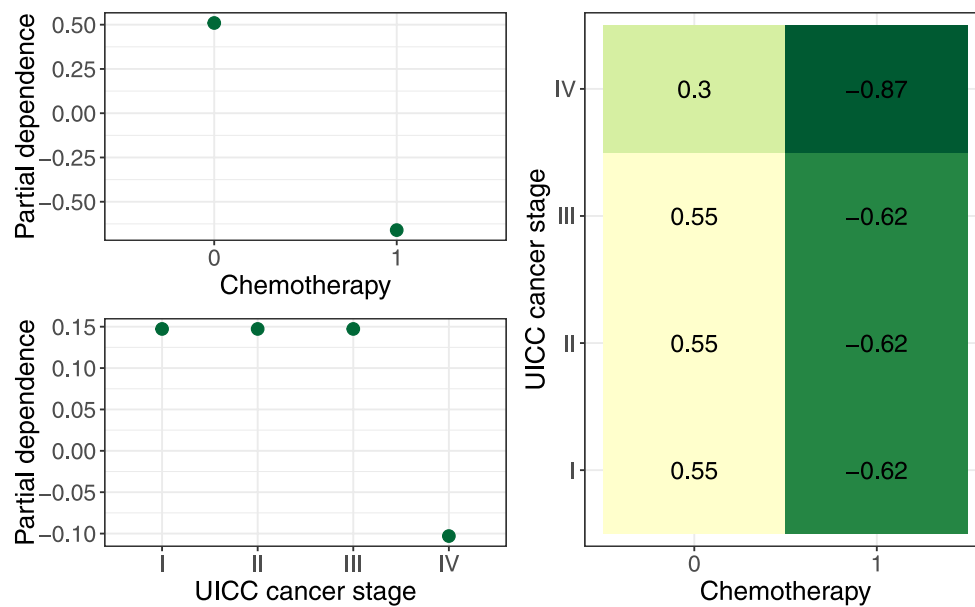


Fig. 4 Partial effects on the copula dependence parameter (ρ), illustrating the effect of chemotherapy and UICC cancer stage on ρ . The top left plot shows the marginal effect of chemotherapy, while the bottom left plot shows the marginal effect of cancer stage. The right plot displays the joint partial dependence, illustrating the interaction between chemotherapy and cancer stage on ρ

μ_T) it has a strong negative effect on the association parameter, leading to a smaller association between survival time and censoring time. This can be interpreted as a broader potential range of survival time given that a patient is censored. A similar effect can be observed for the tumor staging, where stage IV is unfortunately associated with much shorter survival of the patient, but also has a negative effect on the association parameter.

Figure 4 displays the partial effects on the dependence parameter (ρ), showing how chemotherapy and UICC cancer stage influence ρ . The top left plot shows the marginal effect of chemotherapy, while the bottom left plot shows the marginal effect of cancer stage. When chemotherapy is not administered, the dependence parameter is around 0.5, indicating a positive association between survival and censoring time. When patients receive chemotherapy, ρ decreases to approximately -0.7, closely aligning with the estimated coefficient of -1.17. Similarly, I - III of UICC cancer are associated with a dependence of about 0.15, while stage IV shows a decrease to approximately -0.1, consistent with the coefficient estimate of -0.25. The right plot displays the joint partial dependence, showing that the combination of chemotherapy and stage IV results in the lowest dependence values. This confirms that both factors, particularly chemotherapy, strongly reduce the association between survival and censoring time, in line with the coefficients in Table 3.

5 Discussion

We developed CopBoostDepCens, a model-based boosting approach that relies on parametric marginal distributions and copulas, ensuring model identification. CopBoostDepCens does not assume a known copula and allows to include large numbers of potential predictor variables.

Modelling both survival and censoring times together through a copula in combination with a component-wise boosting approach bridges the gap from a compelling theoretical idea to a practical and versatile modelling option in many disciplines.

We demonstrated through simulations that CopBoostDepCens is able to identify the correct predictors in various data situations, including varying numbers of potential covariates and different censoring rates. While boosting the distributional AFT model (assuming independence) also may yield satisfying results, particularly in situations with a low censoring rate or a weak dependence between survival and censoring times, CopBoostDepCens provides additional insights into the dependence between survival and censoring times, not captured by classical approaches. For instance, in our analysis of the overall survival of colon cancer patients—a scenario commonly encountered by biostatisticians—we identified, among other findings, a negative effect of chemotherapy on the relationship between survival and censoring. This insight could not be obtained using previous models.

Evaluating the predictive performance of our model and comparing it with standard approaches such as the AFT model, is challenging. Many conventional metrics, such as the mean absolute error or concordance index, either exclude censored observations or assume non-informative censoring, which leads to biased results or misleading conclusions (Qi et al. 2023). Even metrics that attempt to account for censoring, such as the Brier score, pose challenges. Traditional implementations estimate the censoring distribution using Kaplan–Meier estimator, which assumes independence between survival and censoring times. In contrast, CopBoostDepCens explicitly models the censoring distribution as function of covariates. This fundamental difference complicates the comparison between models and raises concerns about the appropriateness of using standard evaluation metrics that assume independent censoring.

These limitations underscore the need for developing metrics that adequately capture the model performance under dependent censoring (Foomani et al. 2023). Recent work has started to close this gap: Prince et al. (2025) demonstrates that the Brier score is sensitive to the choice of weighting and provides guidance on inverse probability weighting (IPCW), while Lillelund et al. (2025) demonstrates the bias of traditional metrics and propose a copula-based alternative that uses a known Archi-

median copula to estimate $\hat{G}(t)$. However, this approach requires a known copula structure and ignores covariates. Our method avoids these constraints: rather than relying on a copula-graphic estimator, our model directly estimates the censoring distribution conditional on covariates and allows the dependency parameter to vary with covariate effects.

An approach related to ours is that of Midtjord et al. (2022). The authors employ the Clayton copula to account for dependent censoring. However, their model is

restricted to the Clayton copula and limited to a fixed dependence parameter. Furthermore, it assumes that censoring is independent of covariates. In contrast, our boosting approach flexibly models all distributional parameters as functions of covariates.

This flexibility and the ability to model each parameter of the margins and the copula based on covariates also comes with the challenge of interpreting the effects of covariates involved in multiple parameters. Further research is warranted to reduce model complexity, for example, by deselecting covariates that have only a minor impact on overall model performance (Strömer et al. 2022, 2025). An inherent limitation of CopBoostDepCens is the difficulty in providing standard errors for the resulting coefficients, which is true for all boosting approaches. Due to early stopping and the resulting shrinkage of effect estimates, there are no closed formulas for standard errors. To address this, permutation tests could be utilized for significance testing and to provide p -values (Mayr et al. 2017; Hepp et al. 2019), though this would further increase computational costs. Another limitation of our boosting approach is that we currently only have empirical evidence from the simulation study and the application on its performance, but no formal theoretical proof regarding the consistency of effect estimates or asymptotics. Together with the missing standard errors, this makes statistical boosting in general most suitable for exploratory data analyses or prediction modelling—not for confirmatory data analyses (Mayr and Hofner 2018; Strömer et al. 2023).

Lastly, our conditional censoring approach is naturally limited to settings with a sufficient number of censored and uncensored individuals. In our simulations, we considered different censoring rates ranging from 20% to 80%. Also it is clear, that for a complex model as ours where we relate all distribution parameters from two marginal distributions as well as a copula parameter to covariates, one needs a reasonable large sample size for reliable estimation and to detect the most informative predictors. In our simulations, we used $n = 1000$ for most settings, a setting with $n = 100$ can be found in the Supplementary Material. This limitation is already true for most distributional regression approaches, but for CopBoostDepCens becomes even more pressing, as only parts of the observations can be used for estimating the corresponding marginal distributions.

While our current implementation uses one-parameter copulas, the modular structure of the boosting algorithm would allow the integration of more advanced copula families. These could include, for example, Vine (Czado and Nagler 2022) and mixture copulas (Pan et al. 2025) to better capture asymmetric or time-varying dependencies. However, careful consideration of pair-copula selection, identifiability constraints and sample size requirements is necessary for these extensions, particularly in smaller studies (Barthel et al. 2019). Another potential topic for future research involves extending the approach to accommodate left censored or truncated data (Deresa et al. 2022). Doing so could broaden the applicability of the model to a wider range of survival analysis scenarios. Building on recent advances by Deresa and Van Keilegom (2024a) and Deresa and Van Keilegom (2024b), future work could incorporate semi-parametric margins (e.g., a Cox proportional hazards model for survival times with nonparametric baseline hazards) to relax the parametric assumptions while preserving identifiability under dependent censoring.

6 Supplementary Materials

Tables and figures referenced in Sects. 3 and 4 are available with this paper online. The code is available at <https://github.com/AnnikaStr/CopBoostDepCens>.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10985-025-09674-x>.

Acknowledgements Open Access funding enabled and organized by Projekt DEAL. The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776, KL3037/2-1, MA7304/1-1). Ingrid Van Keilegom gratefully acknowledges funding from the FWO and F.R.S. - FNRS (Excellence of Science programme, project ASTeRISK, grant no. 40007517), and from the FWO (senior research projects fundamental research, grant no. G047524N)

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Barthel N, Geerdens C, Czado C, Janssen P (2019) Dependence modeling for recurrent event times subject to right-censoring with d-vine copulas. *Biometrics* 75(2):439–451
- Braekers R, Veraverbeke N (2005) A copula-graphic estimator for the conditional survival function under dependent censoring. *Can J Stat* 33(3):429–447
- Bühlmann P, Hothorn T (2007) Boosting algorithms: regularization, prediction and model fitting. *Stat Sci* 22(4):477–505
- Chen Y-H (2010) Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *J R Stat Soc Ser B (Stat Methodol)* 72(2):235–251
- Collett D (2003) *Modelling survival data in medical research*, 2nd edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis
- Cox DR (1972) Regression models and life-tables. *J R Stat Soc: Ser B (Methodol)* 34(2):187–202
- Crowder M (1991) On the identifiability crisis in competing risks analysis. *Scand J Stat* 18(3):223–233
- Czado C, Nagler T (2022) Vine copula based modeling. *Annu Rev Stat Appl* 9(1):453–477
- Czado C, Van Keilegom I (2023) Dependent censoring based on parametric copulas. *Biometrika* 110(3):721–738
- De Bin R, Stikbakke VG (2023) A boosting first-hitting-time model for survival analysis in high-dimensional settings. *Lifetime Data Anal* 29(2):420–440
- Deresá N, Van Keilegom I, Antonio K (2022) Copula-based inference for bivariate survival data with left truncation and dependent censoring. *Insurance: Math Econ* 107:1–21
- Deresá NW, Van Keilegom I (2020a) Flexible parametric model for survival data subject to dependent censoring. *Biom J* 62(1):136–156

- Deresan NW, Van Keilegom I (2020b) A multivariate normal regression model for survival data subject to different types of dependent censoring. *Comput Stat Data Anal* 144:106879
- Deresan NW, Van Keilegom I (2021) On semiparametric modelling, estimation and inference for survival data subject to dependent censoring. *Biometrika* 108(4):965–979
- Deresan NW, Van Keilegom I (2024) Copula based Cox proportional hazards models for dependent censoring. *J Am Stat Assoc* 119(546):1044–1054
- Deresan NW, Van Keilegom I (2024b) Semiparametric transformation models for survival data with dependent censoring. *Ann Inst Stat Math* 77(3):425–457
- Emura T, Chen Y-H (2016) Gene selection for survival data under dependent censoring: a copula-based approach. *Stat Methods Med Res* 25(6):2840–2857
- Emura T, Chen Y-H (2018) Analysis of survival data with dependent censoring: copula-based approaches. Springer
- Foomani AHG, Cooper M, Greiner R, Krishnan RG (2023) Copula-based deep survival models for dependent censoring. [arXiv:2306.11912](https://arxiv.org/abs/2306.11912) [cs.LG]
- Freund Y (1995) Boosting a weak learning algorithm by majority. *Inf Comput* 12(2):256–285
- Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 55(1):119–139
- Friedman J, Hastie T, Tibshirani R (2000) Special invited paper: additive logistic regression: a statistical view of boosting. *Ann Stat*, pp 337–374
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat*, pp 1189–1232
- Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 18:2529–2545
- Hans N, Klein N, Faschingbauer F, Schneider M, Mayr A (2023) Boosting distributional copula regression. *Biometrics* 79(3):2298–2310
- Hassan MK, Brodmann J, Rayfield B, Huda M (2018) Modeling credit risk in credit unions using survival analysis. *Int J Bank Market* 36(3):482–495
- Heller GZ (2024) Simple or complex statistical models: non-traditional regression models with intuitive interpretations. *Stat Model* 24(6):503–519
- Hepp T, Schmid M, Gefeller O, Waldmann E, Mayr A (2016) Approaches to regularized regression: a comparison between gradient boosting and the lasso. *Methods Inf Med* 55(5):422–430
- Hepp T, Schmid M, Mayr A (2019) Significance tests for boosted location and scale models with linear base-learners. *Int J Biostat* 15(1):20180110
- Hofner B, Mayr A, Robinsonov N, Schmid M (2014) Model-based boosting in R: a hands-on tutorial using the R package mboost. *Stat Comput* 29:3–35
- Howe CJ, Cole SR, Napravnik S, Eron JJ Jr (2010) Enrollment, retention, and visit attendance in the university of north carolina center for aids research hiv clinical cohort, 2001–2007. *AIDS Res Hum Retroviruses* 26(8):875–881
- Huang X, Wolfe RA (2002) A frailty model for informative censoring. *Biometrics* 58(3):510–520
- Huang X, Zhang N (2008) Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics* 64(4):1090–1099
- Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53(282):457–481
- Klein J, Moeschberger M (2005) Survival analysis: techniques for censored and truncated data. Statistics for Biology and Health, Springer, New York
- Lillelund CM, Qi S, Greiner R (2025) Practical evaluation of copula-based survival metrics: beyond the independent censoring assumption. [arXiv:2502.19460](https://arxiv.org/abs/2502.19460)
- Mayr A, Binder H, Gefeller O, Schmid M (2014) The evolution of boosting algorithms: from machine learning to statistical modelling. *Methods Inf Med* 53(6):419–427
- Mayr A, Hofner B (2018) Boosting for statistical modelling: a non-technical introduction. *Stat Model* 18(3–4):365–384
- Mayr A, Hofner B, Schmid M (2012) The importance of knowing when to stop. A sequential stopping rule for component-wise gradient boosting. *Methods Inf Med* 51(2):178–186
- Mayr A, Hofner B, Waldmann E, Hepp T, Meyer S, Gefeller O (2017) An update on statistical boosting in biomedicine. *Comput Math Methods Med* 2017(1):6083072
- Mayr A, Schmid M, Pfahlberg A, Uter W, Gefeller O (2017) A permutation test to analyse systematic bias and random measurement errors of medical devices via boosting location and scale models. *Stat Methods Med Res* 26(3):1443–1460
- Meinshausen N, Bühlmann P (2010) Stability selection. *J R Stat Soc B: Stat Methodol* 72(4):417–473

- Midtftjrd AD, De Bin R, Huseby AB (2022) A copula-based boosting model for time-to-event prediction with dependent censoring. arXiv preprint [arXiv:2210.04869](https://arxiv.org/abs/2210.04869) [stat.ME]
- Moradian H, Larocque D, Bellavance F (2017) L_1 splitting rules in survival forests. *Lifetime Data Anal* 23:671–691
- Pan R, Nieto-Barajas LE, Craiu R (2025) Multivariate temporal dependence via mixtures of rotated copulas
- Patton AJ (2006) Modelling asymmetric exchange rate dependence. *Int Econ Rev* 47:527–556
- Prince T, Bommert A, Rahnenführer J, Schmid M (2025) On the estimation of inverse-probability-of-censoring weights for the evaluation of survival prediction error. *PLoS ONE* 20(1):1–22
- Qi S, Kumar N, Farrokh M, Sun W, Kuan L, Ranganath R, Henao R, Greiner R (2023) An effective meaningful way to evaluate survival models. *Proc Mach Learn Res* 202:28244–28276
- Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. *J R Stat Soc: Ser C: Appl Stat* 54(3):507–554
- Rivest L-P, Wells MT (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *J Multivar Anal* 79(1):138–155
- Samoladas I, Angelis L, Stamelos I (2010) Survival analysis on the duration of open source projects. *Inf Softw Technol* 52(9):902–922
- Schmid M, Hothorn T (2008) Boosting additive models using component-wise P-splines. *Comput Stat Data Anal* 53(2):298–311
- Schneider S, Demarqui FN, Colosimo EA, Mayrink VD (2020) An approach to model clustered survival data with dependent censoring. *Biom J* 62(1):157–174
- Seipp A, Otto-Sobotka F (2022) dirttee: Distributional regression for time to event data. R package version 1:1
- Seipp A, Uslar V, Weyhe D, Timmer A, Otto-Sobotka F (2021) Weighted expectile regression for right-censored data. *Stat Med* 40(25):5501–5520
- Srujana B, Verma D, Naqvi S (2024) Machine learning vs. survival analysis models: a study on right censored heart failure data. *Commun Stat - Simul Comput* 53(4):1899–1916
- Stijven F, Molenberghs G, Van Keilegom I, Van der Elst W, Alonso A (2024) Evaluating time-to-event surrogates for time-to-event true endpoints: an information-theoretic approach based on causal inference. *Lifetime Data Analysis*, pp 1–23
- Strömer A, Klein N, Staerk C, Faschingbauer F, Klinkhammer H, Mayr A, (2025) Enhanced variable selection for boosting sparser and less complex models in distributional copula regression. *Stat Biosci*
- Strömer A, Klein N, Staerk C, Klinkhammer H, Mayr A (2023) Boosting multivariate structured additive distributional regression models. *Stat Med* 42(11):1779–1801
- Strömer A, Staerk C, Klein N, Weinhold L, Titze S, Mayr A (2022) Deselection of base-learners for statistical boosting—with an application to distributional regression. *Stat Methods Med Res* 31(2):207–224
- Sujica A, Van Keilegom I (2018) The copula-graphic estimator in censored nonparametric location-scale regression models. *Econom Stat* 7:89–114
- Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B (2018) Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Stat Comput* 28:673–687
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol* 58(1):267–288
- Tsiatis A (1975) A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci* 72(1):20–22
- Wei LJ (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med* 11:1871–1879
- Wood S (2017) Generalized additive models: an introduction with R, 2nd edition. Chapman & Hall/CRC
- Zheng M, Klein JP (1995) Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* 82(1):127–138

Authors and Affiliations

Annika Strömer¹  · Nadja Klein²  · Ingrid Van Keilegom³  · Andreas Mayr¹ 

✉ Annika Strömer
annika.stroemer@uni-marburg.de

¹ Department of Medical Biometry and Statistics, University of Marburg, Marburg, Germany

² Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany

³ ORSTAT, KU Leuven, Leuven, Belgium

3.4 Publication D: Deselection of base-learners for statistical boosting — with an application to distributional regression

Strömer A, Staerk C, Klein N, Weinhold L, Titze S, Mayr A. Deselection of base-learners for statistical boosting — with an application to distributional regression. *Statistical Methods in Medical Research* 2022; 31(2): 207–224.

<https://doi.org/10.1177/09622802211051088>

Supplementary information can be found at:

<https://doi.org/10.1177/09622802211051088>

Implementations are available on GitHub:

<https://github.com/AnnikaStr/DeselectBoost>

Deselection of base-learners for statistical boosting—with an application to distributional regression

Statistical Methods in Medical Research
1–18

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802211051088

journals.sagepub.com/home/smm



Annika Strömer¹ , Christian Staerk¹, Nadja Klein²,
Leonie Weinhold¹, Stephanie Titze³ and Andreas Mayr¹

Abstract

We present a new procedure for enhanced variable selection for component-wise gradient boosting. Statistical boosting is a computational approach that emerged from machine learning, which allows to fit regression models in the presence of high-dimensional data. Furthermore, the algorithm can lead to data-driven variable selection. In practice, however, the final models typically tend to include too many variables in some situations. This occurs particularly for low-dimensional data ($p < n$), where we observe a slow overfitting behavior of boosting. As a result, more variables get included into the final model without altering the prediction accuracy. Many of these false positives are incorporated with a small coefficient and therefore have a small impact, but lead to a larger model. We try to overcome this issue by giving the algorithm the chance to deselect base-learners with minor importance. We analyze the impact of the new approach on variable selection and prediction performance in comparison to alternative methods including boosting with earlier stopping as well as twin boosting. We illustrate our approach with data of an ongoing cohort study for chronic kidney disease patients, where the most influential predictors for the health-related quality of life measure are selected in a distributional regression approach based on beta regression.

Keywords

Beta regression, generalized additive models for location, scale, and shape, model-based boosting, variable selection, earlier stopping

Introduction

In modern biostatistics, model building and variable selection have become increasingly important, particularly in the context of applications in high-dimensional data settings, where the number of potential predictors p is larger compared to the sample size ($p \gg n$).¹ Important examples include genetic or molecular data (e.g. Chen et al.²; Choi et al.³), but also in more classical clinical studies one often aims to obtain a relatively sparse model with good prediction accuracy including only the most relevant variables (e.g. Steyerberg and Vergouwe⁴; Sauerbrei et al.⁵).

Component-wise gradient boosting⁶ provides a framework to handle this, with the key features of variable selection and the possibility to manage high-dimensional data problems. In combination with regression-type base-learners,⁷ it is able to maintain the usual interpretability of statistical regression models—equivalent to the ones that were estimated using classical penalized likelihood or Bayesian inference. Statistical boosting provides a large flexibility due to the modular nature

¹Department of Medical Biometrics, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Germany

²Emmy Noether Research Group in Statistics and Data Science, Humboldt-Universität zu Berlin, Germany

³Department of Nephrology and Hypertension, FAU Erlangen-Nuremberg, Germany

Corresponding author:

Annika Strömer, Department of Medical Biometrics, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Venusberg-Campus I, 53127 Bonn, Germany.

Email: stroemer@imbie.uni-bonn.de

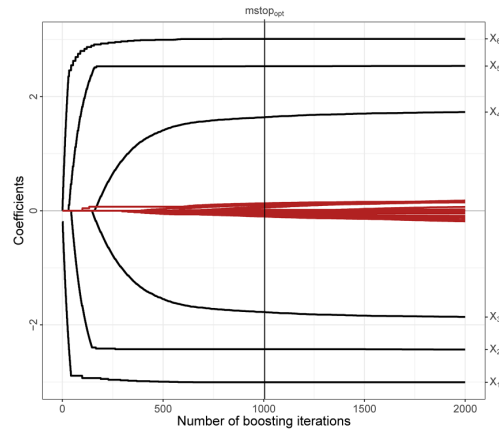


Figure 1. Coefficient paths along the number of boosting iterations for a simulated data set with $n = 500$ observations and $p = 50$ variables which were simulated from a multivariate normal distribution with Toeplitz covariance structure and a correlation of 0.9. Only variables X_1, \dots, X_6 are *informative* with true coefficients $\beta_{\text{inf}} = (-3, -2.5, -2, 2, 2.5, 3)^T$. The coefficient paths for the *non-informative* variables are colored red. The optimal stopping iteration $m_{\text{stop_opt}} = 1004$ was determined by 10-fold cross-validation.

of the approach: any type of base-learner (linear models, splines, spatial models) can be combined with any type of convex loss function.⁸

Despite these advantages, in some applications the algorithm tends to select too many variables. This often occurs for rather low-dimensional settings with relatively large sample sizes ($p < n$), where irrelevant base-learners often get selected with a very small effect size. This is associated with slow overfitting and thus with a higher number of boosting iterations m_{stop} which lead to a larger set of selected variables. For example, in a recent beta-regression analysis on the health-related quality of life (QoL) in $n = 3522$ chronic kidney disease patients, statistical boosting selected 33 out of $p = 54$ potential variables.⁹

As an illustration, Figure 1 displays the coefficient paths of component-wise boosting with the squared error loss in the context of linear regression for a simulated data set in which only the first six variables X_1, \dots, X_6 are informative. One can observe that the estimated coefficients of the six informative variables are the largest in absolute values, while several non-informative variables are incorporated with small coefficient sizes around zero. Therefore, the selected non-relevant variables have only a minor impact on the predictive performance but lead to a larger model with difficult interpretation.

Bühlmann and Hothorn¹⁰ tried to overcome this issue and extended the classical boosting approach to a two-stage design called twin boosting, which was inspired by the adaptive lasso (see Zou¹¹). The first stage consists of a classical boosting algorithm. The second stage is similar to the first, with the difference that variables that have not been selected are excluded; furthermore, variables that have been selected in the first stage receive weights based on the size of their coefficients, making it more likely that the important variables will be selected again in the second stage. Other approaches aiming to increase the sparsity of statistical boosting focus on reducing the number of iterations m_{stop} : for example, the one standard-error rule was originally considered by Breiman et al.¹² in the context of random forests and does not select the optimal tuning parameter regarding prediction accuracy, but in case of boosting the smallest m_{stop} that is still in the margin of one standard error from the minimum risk. Ellenbach et al.¹³ further extended this approach (*RobustC*) to obtain a less complex prediction rule that is less affected by the characteristics of the resampling scheme compared to the one standard-error rule. A potential disadvantage of approaches that lead to earlier stopping is that they suffer from the side-effect of inducing also a higher amount of shrinkage. This additional shrinkage of selected effect estimates might not necessarily lead to a better performance (cf. Van Calster et al.¹⁴).

Here we propose a general procedure to enhance the sparsity of statistical boosting models, where the final selection of variables is based on the risk reduction resulting from the individual updates of the corresponding base-learners. With this approach, we exclude those base-learners (and their corresponding variables) from the prediction model which attributed only slightly to the total risk reduction. As an alternative to earlier stopping of the boosting algorithm—that is moving “horizontally” on the regularization paths—we consider the individual contributions of different variables after a particular number of boosting iterations. The benefits of this “vertical” view on regularization paths have also recently been discussed in the context of other regularization methods such as the thresholded Lasso^{15; 16} including the possibility of deselecting noise variables which are included “early” on the regularization paths. Furthermore, we directly enforce the sparsity of the

final models without unnecessarily increasing the amount of shrinkage on effect estimates. We illustrate the proposed method with the selection of predictors for the health-related QoL data of the German Chronic Kidney Disease Study (GCKD). We compare our results to a previous analysis of these data⁹ which partly motivated the new methodological development. With the new deselection approach, we are able to select much sparser models while still yielding a similar prediction performance.

The remainder of the paper is structured as follows. In Section ‘Methods’, we introduce the new approach for an improved variable selection and consider alternative methods for achieving sparser models. In Section ‘Simulation study’, we compare these methods via simulated data under various conditions for different models. Finally, we apply our new approach to the QoL data and present the results in Section ‘Quality of life of chronic kidney disease patients’. Conclusively, Section ‘Discussion and conclusion’ summarizes our findings and discusses future research questions.

Methods

Model-based boosting

Boosting was first established in the context of machine learning^{17; 18} and was later extended to fit statistical models.^{19; 20} Statistical boosting algorithms^{21; 22} can be used to analyze high-dimensional data problems, in which classical inferential methods are no longer applicable (e.g. least squares method for linear regression models). Furthermore, boosting yields data-driven variable selection and shrinkage of effect estimates.⁶

The model fitting is carried out by iteratively minimizing the empirical risk of an appropriate loss function. This loss defines the regression problem and needs to be specified in advance. In generalized linear models (GLMs) and generalized additive models (GAMs), the loss function corresponds to the negative log-likelihood of the outcome distribution. For classical linear regression models, for example, we minimize the squared error (L_2 loss), which corresponds to maximizing the likelihood of a Gaussian distribution. Different effect types can be determined for each covariate (e.g. linear or smooth effects), which reflect the type of influence the variable has in the model. These underlying functions are called base-learners; in the simplest case they are univariate linear models representing linear effects. In each iteration, the negative gradient of the loss function is determined and every base-learner is separately fitted to the negative gradient. Afterwards, only the best performing base-learner is selected (i.e. the base-learner that best fits the negative gradient) and the corresponding estimated effect is multiplied by a small fixed step size (default is $\nu = 0.1$) before it is included in the model. Due to the selection of single base-learners in each iteration, the algorithm carries out variable selection. This process is repeated until the number of boosting iterations m_{stop} is reached, whereby every base-learner can be selected several times. In the classical boosting algorithm, every base-learner that was once included in the model cannot be deselected.²³

The number of boosting iterations is the main tuning parameter and can be selected, for example, by cross-validation or other resampling techniques. The optimization of the stopping iteration—also referred to as *early stopping*—is crucial to prevent overfitting and to favor the sparsity of the resulting model. The smaller m_{stop} , the fewer variables are included in the final model as only one base-learner is updated in each iteration. Additionally, early stopping typically improves the prediction accuracy and leads to shrinkage of effect estimates.²⁴

Earlier stopping strategies

Due to the influence of the number of boosting iterations m_{stop} on the variables finally selected by the algorithm, one approach to achieve sparser models is to enforce *earlier* stopping of the algorithm, that is, selecting a smaller m_{stop} . With this approach it is assumed that variables that are updated in early iterations of the algorithm have a greater influence on the prediction of the model than variables that are added later to the model. Typically, classical early stopping selects the stopping iteration $m_{\text{stop_opt}}$ that leads to the smallest (optimal) cross-validated prediction risk (CV).

The one standard error rule (oSE) is one approach to enforce earlier stopping and has already been used in context of penalized regression and regression trees.^{12; 25} With this approach, the tuning parameter m_{stop} is chosen as the smallest iteration for which the CV is within one standard error of the minimal CV (cf. Friedman et al.²⁵; Hastie et al.²⁶)

$$\text{CV}(m_{\text{stop}}) \leq \text{CV}(m_{\text{stop_opt}}) + \text{se}(\text{CV}(m_{\text{stop_opt}})).$$

The minimal cross-validated predictive risk $\text{CV}(m_{\text{stop_opt}})$ corresponds to the CV of the optimal stopping iteration. Furthermore, $\text{se}(\text{CV}(m_{\text{stop_opt}}))$ represents the standard error of the minimum over the CV folds. Consequently, this method has dependencies on the number of CV folds and the sample size.

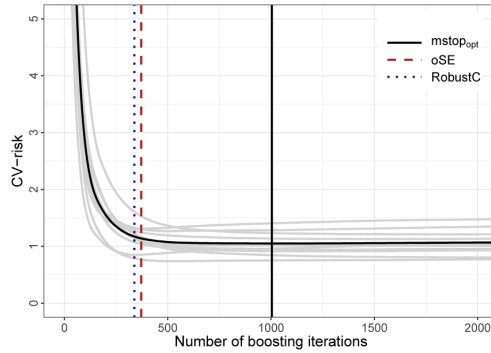


Figure 2. Application of the oSE and RobustC on the cross-validated estimation of the empirical risk with 10-fold cross-validation. The vertical solid black line reflects the optimal stopping iteration via cross-validation ($m_{\text{stop}} = 1004$), the red dashed one displays the oSE ($m_{\text{stop}} = 372$) and the blue dotted one RobustC ($m_{\text{stop}} = 339$) with $c_{rC} = 1.1$.

Based on the idea of the oSE approach, Ellenbach et al.¹³ proposed an alternative more robust approach, called RobustC. Here the smallest m_{stop} is chosen, whose CV is still within a range of a fixed additional tuning parameter c_{rC} multiplied with the minimum CV

$$CV(m_{\text{stop}}) \leq c_{rC} \times CV(m_{\text{stop_opt}}).$$

Ellenbach et al.¹³ suggested the values $c_{rC} \in \{1, 1.1, 1.3, 1.5, 2\}$ for the case of a binary outcome. The authors aimed for a less complex predictive rule and for choosing a robust tuning parameter, which is essential in cross-study predictions.¹³

Considering the example from the introduction, Figure 2 shows the CV-risk for 10-fold cross-validation with 2000 boosting iterations. The vertical solid black line shows the optimal stopping iteration and corresponds to the minimum average risk over the 10-fold cross-validation samples. The vertical dashed red line is the stopping iteration which yields the oSE, while the blue dotted line corresponds to the optimal iteration according to RobustC. One can observe that the stopping iterations of the earlier stopping strategies are less than half as large as the original m_{stop} .

A further alternative approach to obtain sparser models is probing. The idea is to extend the data set by random noise variables, the so-called probes, which are randomly shuffled versions of the originally observed variables. The algorithm stops when the first probe is selected. For more details on this approach see Thomas et al.²⁷

Deselection of base-learners with a small risk reduction

Several other approaches have been developed to enhance the sparsity of boosting models (e.g. Hofner et al.²⁸; Thomas et al.²⁷). Most of them focus on the selection step in the algorithm, or on the tuning of the stopping iteration m_{stop} (section ‘Earlier stopping strategies’). Our new procedure is based on *actively* deselecting variables that have been selected by the algorithm, but result in only minor importance regarding the predictions of the model.

We address this issue with an approach that aims at eliminating variables with a small impact and directly enforce the sparsity of the model. The general idea is to first apply a standard boosting algorithm with early stopping via cross-validation or resampling techniques; then, we determine the variables selected by boosting with a minor importance for the model and deselect those components. Afterwards, we boost again incorporating only the selected variables that survived as candidate variables. In this context, our procedure shows analogies to the twin boosting approach.¹⁰ In our deselection procedure, we consider the risk reduction as a measure for variable importance and deselect those variables that only represent a small percentage of the total risk reduction.

The risk reduction by base-learner j after m_{stop} boosting iterations can be defined as the *attributable* risk reduction R_j

$$R_j = \sum_{m=1}^{m_{\text{stop}}} I(j = j^{*[m]}) (r^{[m-1]} - r^{[m]}), \quad j = 1, \dots, p, \quad (1)$$

where I denotes the indicator function and $j^{*[m]}$ is the selected base-learner in iteration m . Furthermore, $r^{[m-1]} - r^{[m]}$ represents the risk reduction in iteration m , for risks $r^{[m]}$ and $r^{[m-1]}$ at iterations m and $m - 1$. For a given threshold $\tau \in (0, 1)$, we

deselect base-learner j if

$$R_j < \tau \cdot (r^{[0]} - r^{[m_{\text{stop}}]}), \quad (2)$$

where $r^{[0]} - r^{[m_{\text{stop}}]}$ represents the total risk reduction and R_j denotes the attributable risk reduction of base-learner j .

A schematic overview of the proposed procedure can be found in Box 1. Step 1 of the procedure consists of the initial boosting for which the coefficient paths are shown in Figure 3 (left), corresponding to the simulation example discussed earlier (see Figures 1 and 2). Overall, 23 variables (of the 50 variables) were selected (shown as horizontal red and black paths) after $m_{\text{stop}} = 1004$ boosting iterations which were tuned by 10-fold cross-validation (indicated by the vertical black line). For the deselection in Step 2, the attributable risk reduction along the iterations is shown for each individual base-learner in the central plot of Figure 3). To illustrate the effect of the deselection step of the proposed method, consider the thresholds $\tau = 0.01$ (horizontal dashed line) and $\tau = 0.1$ (horizontal dotted line). Here, it can be observed that our deselection procedure is fundamentally different to earlier stopping approaches discussed in section ‘Earlier stopping strategies’, as the choice of the threshold for the deselection corresponds to a vertical view on the individual risk reductions after a given number of boosting iterations (see central plot of Figure 3; on the other hand, earlier stopping simply corresponds to a horizontal shift on the usual regularization paths of boosting (see Figure 2).

In the following, we consider a threshold value of $\tau = 0.01$ and accordingly deselect those variables which contribute less than 1% to the total risk reduction. The black paths correspond to the variables included in the model after applying the deselection approach, while the red paths do not cross the 1% line and the corresponding variables are deselected from the model. We can observe that these variables contribute only slightly to the risk reduction and are incorporated with a coefficient size around zero in the initial boosting model (as shown in Figure 3, left). In this example, the deselection approach with threshold $\tau = 0.01$ deselects all noise variables from the model, but not the signal variables X_1, \dots, X_6 . Variables X_1, X_2, X_5 , and X_6 have by far the greatest individual contributions to the total risk reduction; however, variables X_3 and X_4 also exceed the 1% threshold (but not the 10% threshold). After the deselection step, we boost again (Step 3) with only the remaining variables and receive the final model (see Figure 3, right) which contains here exclusively the six informative variables.

Box 1: Deselection procedure

1. Initial boosting:
 - Early stopping: Tuning of m_{stop} based on cross-validation or resampling.
2. Deselection:
 - Identify the base-learners with the smallest impact on the risk reduction according to $R_j < \tau \cdot (r^{[0]} - r^{[m_{\text{stop}}]})$ (2) and remove them from the model.
3. Final boosting:
 - Boost again with the remaining variables and the m_{stop} of step 1.

Deselection of base-learners for distributional regression

In classical statistical models, the relationship between a response variable and covariates is most often modeled only based on the expected value. For example, a generalized additive model (GAM)²⁹ in which the conditional mean $\mu = \mathbb{E}(y|x)$ relates to an additive predictor η via a link-function g , is given by

$$g(\mu) = \eta(x) = \beta_0 + \sum_{j=1}^p f_j(x_j)$$

with the intercept β_0 and the additive effects f_j for the covariates X_j with $j = 1, \dots, p$ (including linear, smooth or random effects). Consider, for example, a Gaussian distribution, which has two parameters: the expected value μ and the scale parameter σ . In a classical GAM, we assume that σ is fixed and only model the mean parameter μ in terms of the covariates.

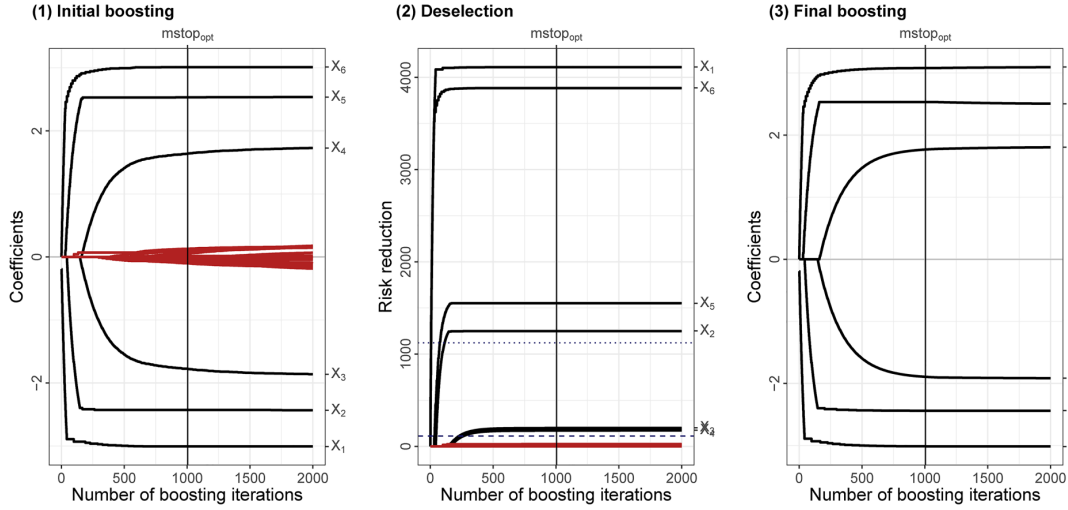


Figure 3. Simulation example for the deselection procedure (see Box 1). The left plot shows the coefficient paths of the initial boosting. The central plot displays the attributable risk reduction for the individual variables, together with the 1% threshold (dashed line) and 10% threshold (dotted line) of the total risk reduction. The coefficient paths of the right plot correspond to the final boosting.

In some cases, this may lead to an overly restrictive point of view, for example, in the presence of heteroscedasticity. In addition, skewness and kurtosis may be large so that more complex non-symmetric distributions are required where potentially skewness or higher order moments could be modeled through covariates to obtain a more accurate model. Following this idea, GAMs have been extended to generalized additive models for location, scale and shape (GAMLSS) by Rigby and Stasinopoulos,³⁰ where a general parametric density $P(y|\theta_1, \dots, \theta_K)$ with distributional parameters θ_k can be employed. Here, each distribution parameter θ_k , with $k = 1, \dots, K$, can be modeled by an additive predictor η_k depending on covariates. Furthermore, for each parameter θ_k , we have parameter-specific link-functions $g_k(\cdot)$ as well as parameter-specific covariates x_{k1}, \dots, x_{kp_k} . In general, the linear predictors in a GAMLSS for K distributional parameters can be written as follows:

$$g_k(\theta_k) = \eta_k = \beta_{0k} + \sum_{j=1}^{p_k} f_{jk}(x_{kj}), \quad k = 1, \dots, K,$$

where β_{0k} are the intercepts for the distributional parameters θ_k and f_{jk} denote the functions of the effect of variable X_j on the parameter θ_k .

GAMLSS can also be fitted via statistical boosting with the package **gamboostLSS**.³¹ As in the classical setting of boosting GAMs, the main tuning parameter is the stopping iteration m_{stop} which controls shrinkage of effect estimates and variable selection. Here, we focus on a non-cyclical boosting approach,³² which performs in every iteration only the overall best-performing update among the available candidate variables (base-learners) and distribution parameters. So the term *component-wise* boosting in this context does not only refer to the components of X , but also to the components of the parameter space $\theta_1, \dots, \theta_K$ of the corresponding likelihood. To receive the overall best performing base-learner, the empirical risk (the negative log-likelihood) of the best fitted base-learner is determined for each distribution parameter and then compared across the different dimensions.

The updates are independent for the parameters and each additive predictor may depend on different variables with the guarantee of data-driven variable selection in every submodel: Figure 4 displays the estimated coefficient paths in a linear Gaussian location-scale model with distributional parameters μ (left) and σ (center). The data set consists of $n = 500$ observations and $p = 20$ variables, where the first three variables X_1, X_2, X_3 are informative for the mean parameter μ with $\beta_{\mu_{\text{inf}}} = (-2, 1.25, 1)^T$, while variables X_4, X_5, X_6 are informative for the scale parameter σ with $\beta_{\sigma_{\text{inf}}} = (0.5, -0.5, 0.5)^T$. All other variables are non-informative with $\beta_{\mu} = 0$ and $\beta_{\sigma} = 0$. The explanatory variables were simulated from a multivariate normal distribution with Toeplitz covariance structure and a correlation of $\rho = 0.5$. The optimal number of boosting iterations is $m_{\text{stop}} = 6479$, optimized via 10-fold cross-validation. Note that in every iteration only a single component is updated for one parameter.

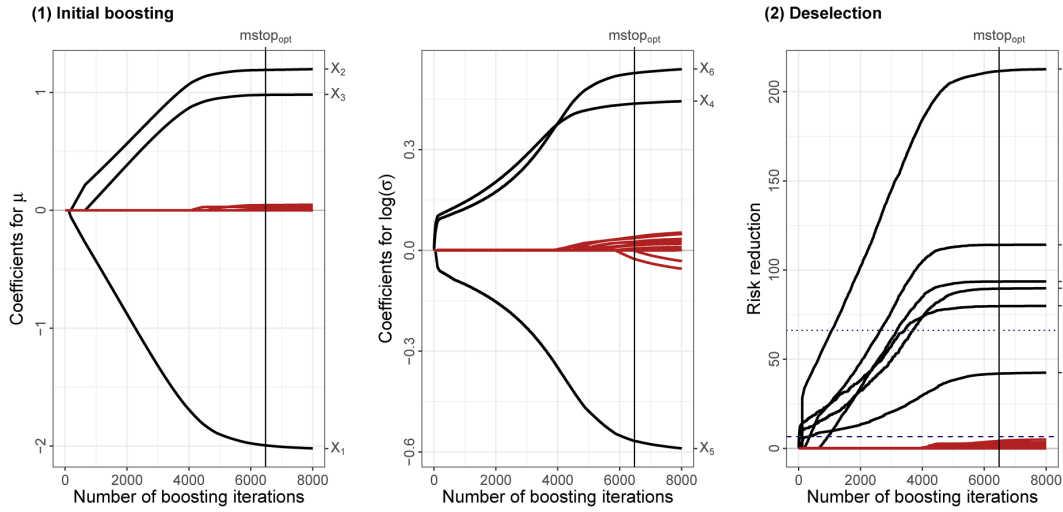


Figure 4. Simulation example for gamboostLSS with three informative variables for μ (X_1, X_2, X_3) and three for σ (X_4, X_5, X_6). The first two plots display the coefficient paths (for μ (left) and σ (center)) and the third plot shows the attributable risk reductions for the individual variables for both distributional parameters together with the 1% (dashed) and 10% (dotted) line of the total risk reduction. The variables corresponding to the black coefficient path are still in the model after deselection with $\tau = 0.01$.

We notice that in the first iterations, components of parameter σ were more often updated, which can be observed by the increase of the coefficient sizes for the variables X_4, X_5, X_6 in the first iterations. In total, the boosting model contains 18 of the 50 variables, with 11 variables selected for μ and 14 variables selected for σ (where seven variables were selected for both μ and σ). Hence, additional variable selection can be advantageous to obtain sparser and thus more interpretable models, which only include the informative variables.

Using equation (1 and considering $j = 1, \dots, \sum p_k$, the risk reduction in a GAMLSS for component j can be defined similar as before. For the deselection of variables with a low impact on the risk reduction for distributional regression, we consider the distributional parameters together, where each parameter can depend on different variables. Analogous to equation (2, we deselect component j if

$$R_j < \tau \cdot (r^{[0]} - r^{[m_{\text{stop}}]})$$

with fraction $\tau \in (0, 1)$ and total risk reduction $r^{[0]} - r^{[m_{\text{stop}}]}$. Note that the deselected components may arise from different distributional parameters and that with this definition, GAMs are included as a special case in the general formulation for a GAMLSS with $p_k \equiv p$ and $k = 1$.

For the simulation example, the risk reduction of the variables for μ and σ is shown in the right plot (Figure 4). As shown in Figure 3, the threshold value is chosen as $\tau = 0.01$ (horizontal dashed line) and $\tau = 0.1$ (horizontal dotted line). The black paths correspond to the variables remaining in the model after applying the deselection procedure (with $\tau = 0.01$) for distributional regression and have by far the highest impact on the risk reduction. The deselection results in a model including only the six informative variables (instead of the 18 initially selected variables). For the choice of an appropriate value for the threshold parameter τ , we examined different potential values observing the attributable risk reduction of the base-learner as shown in Figure 3 (second plot) and Figure 4 (third plot).

Considering Figure 3, the variables X_1, X_2, X_5 , and X_6 have the largest impact on the risk reduction. All of those variables remain in the model with a deselection threshold of 1% as well as the other two informative variables X_3 and X_4 . For the 10% boundary, X_3 and X_4 would not enter the model because of a smaller risk reduction. Even for the data example in Figure 4, 10% is not an appropriate choice, since the variables X_1, \dots, X_6 have a noticeable impact on the risk reduction, but X_5 would fall out at this limit. A threshold of 1% appears to be reasonable in the considered situation. However, in non-sparse situations, when many base-learners contribute only with a small risk reduction to the model, multiple signal variables may be deselected with a threshold of $\tau = 0.01$. This extreme scenario should be rare, and in such non-sparse data situations, enforcing variable selection might not be favorable in general.

An implementation of the enhanced variable selection approach is available at GitHub (<https://github.com/AnnikaStr/DeselectBoost>).

Simulation study

To evaluate the performance of our new approach for different data settings, we conduct a simulation study focusing on the variable selection properties as well as the prediction accuracy in comparison with the methods for earlier stopping, described in Section ‘Earlier stopping strategies’.

Specifically, the questions to be investigated in the simulation study are as follows:

1. Is the direct deselection approach able to identify the truly informative variables (decreasing the number of false positive variables selected by classical boosting)?
2. How does the reduction in selected variables affect the prediction accuracy?
3. How does the new procedure perform in comparison to the earlier stopping strategies, e.g. oSE and RobustC?
4. What is an appropriate value for τ in the proposed deselection approach?

Settings

To examine those questions, different settings are considered: First, we start with classical mean regression models (linear, non-linear, and logistic regression) and afterwards, we consider the deselection approach in the context of distributional regression models.

For all simulations, the explanatory variables X_1, \dots, X_p were simulated from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with a Toeplitz covariance structure $\Sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq p$, where $\rho \in (0, 1)$ is the correlation between consecutive variables X_j and X_{j+1} . For an alternative block-wise covariance structure, see the corresponding results in Supplemental Material A.1. Overall, we considered two different dimensions of the data problem: (i) a low-dimensional setting ($p < n$) with $n = 500$ observations and $p = 20$ variables and (ii) a high-dimensional setting ($p > n$) with $n = 500$ observations and $p = 1000$ variables. In total, six of the included variables were informative (for the distributional regression, three for each parameter). Furthermore, a low-correlated scenario with $\rho = 0.2$ and a high-correlated scenario with $\rho = 0.8$ was considered for each setting. Additionally, we consider a variation of signal-to-noise ratios (SNRs) and the corresponding effect of different threshold values τ with $\text{SNR} \in \{0.15, 6, 14, 64\}$ and $\tau \in \{0.005, 0.0075, 0.01, 0.025, 0.05, 0.075, 0.1, 0.125\}$.

For evaluation, we generated test data sets with 1000 observations from the same distribution as the training data sets. As in the illustrative examples, the number of boosting iterations was tuned via 10-fold cross-validation. The fixed step size is set to $\nu = 0.1$ and was not varied in the simulation, considering that it does not largely affect the risk reduction as long as the step size is chosen reasonably small.³³ We additionally compared the deselection approach (with $\tau = 0.01$) with the earlier stopping strategies, oSE and RobustC (additional comparison with probing is given in Supplemental Material A.2). The parameter value for RobustC is chosen as $c_{rC} = 1.05$ for a continuous outcome variable and $c_{rC} = 1.1$ for a binary response, following the recommendation of Ellenbach et al.¹³

For each setting, 100 simulation runs were conducted and the data sets were generated from the following models:

Scenario A (linear regression): The true linear model for the continuous outcome variable Y is given by

$$y = -2x_1 - 1.5x_2 - x_3 + x_4 + 1.5x_5 + 2x_6 + \epsilon,$$

with $\epsilon \sim N(0, 1)$. The base-learners correspond to simple linear models and the performance was assessed using the mean squared error of prediction (MSEP).

Scenario B (non-linear regression): The outcome variable Y was generated from the model

$$y = 1.5 \sin(x_1) + x_2 - 0.25x_3^2 - 0.25x_4 - x_5 - 1.5x_6 + \epsilon,$$

with $\epsilon \sim N(0, 1)$. Smooth P-splines were used as base-learners and the MSEP was used for evaluation.

Scenario C (logistic regression): The logistic regression model for covariates with only linear effects on the response was simulated according to

$$\log\left(\frac{\mathbb{P}(y = 1|x)}{\mathbb{P}(y = 0|x)}\right) = -5x_1 - 2.5x_2 - x_3 + x_4 + 2.5x_5 + 5x_6.$$

As evaluation criteria, the Brier score and the area under the curve (AUC) were analyzed on test data.

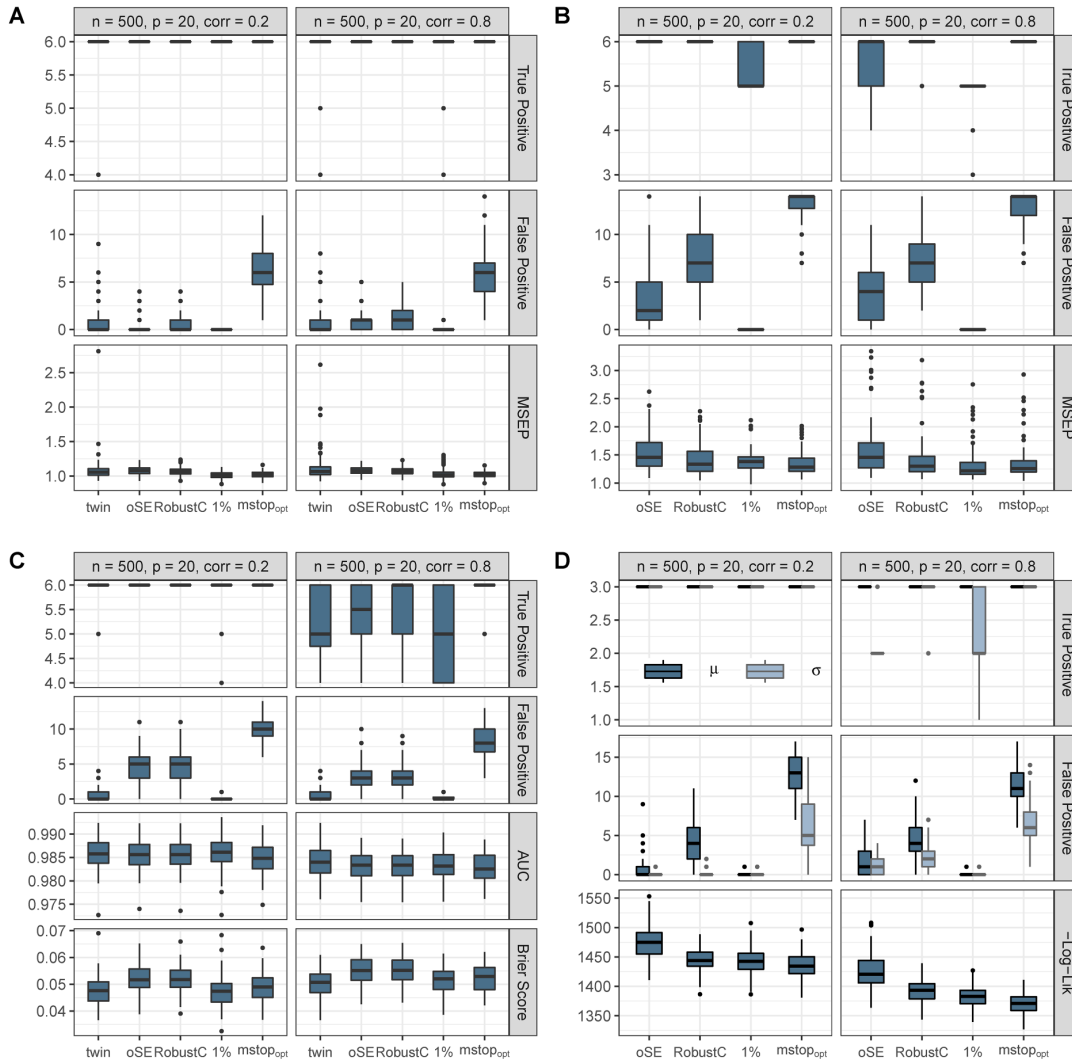


Figure 5. Low-dimensional setting: Comparison of the oSE, RobustC, the deselection approach with $\tau = 0.01$ (1%) and the classical boosted model with $m_{\text{stop_opt}}$ regarding the true positives, false positives and the prediction performance for (A) linear regression, (B) non-linear regression, (C) logistic regression, and (D) distributional regression. Additional comparisons with twin boosting (twin) for the linear and logistic regression scenarios.

Scenario D (distributional regression): For the distributional regression model, we consider a Gaussian regression with expected value μ and scale parameter σ . Both parameters depend on three different covariates

$$\begin{aligned}\mu &= -2x_1 + 1.25x_2 + x_3, \\ \log(\sigma) &= 0.5x_4 - 0.5x_5 + 0.5x_6.\end{aligned}$$

The boosting model was configured with simple linear models as base-learners and the performance was evaluated via the negative log-likelihood.

All simulations were conducted in the statistical computing environment **R**³⁴ using the add-on package **mboost**³⁵ for model-based boosting. The algorithm for fitting GAMLSS models via component-wise gradient boosting is implemented in **gamboostLSS**.³¹ Twin boosting is implemented in the package **bst**.³⁶ The **R** code to reproduce the following simulation results can be found online at GitHub (<https://github.com/AnnikaStr/DeselectBoost>).

Results

Figure 5 shows the results of the low-dimensional simulations regarding the four previously described models for the low- as well as the high-correlated settings, respectively. For each setting, the true positives, false positives, and the predictive performance for the respective model are shown for the earlier stopping strategies, the deselection procedure with $\tau = 0.01$ and the classical boosted model.

In general, the two main strategies (earlier stopping and deselection) resulted in a reduction of false positives. In each of the four models, the fewest false positives were obtained with the proposed deselection procedure; more precisely, almost all false positives were deselected. For some models, one can observe that the selection of informative variables was slightly influenced by the earlier stopping and deselection approach, particularly for the high-correlated settings.

In comparison with classical boosting, the deselection procedure yielded comparable or slightly better predictive performances for simulated data based on Scenario A, Scenario B, and Scenario C. Furthermore, the earlier stopping strategies

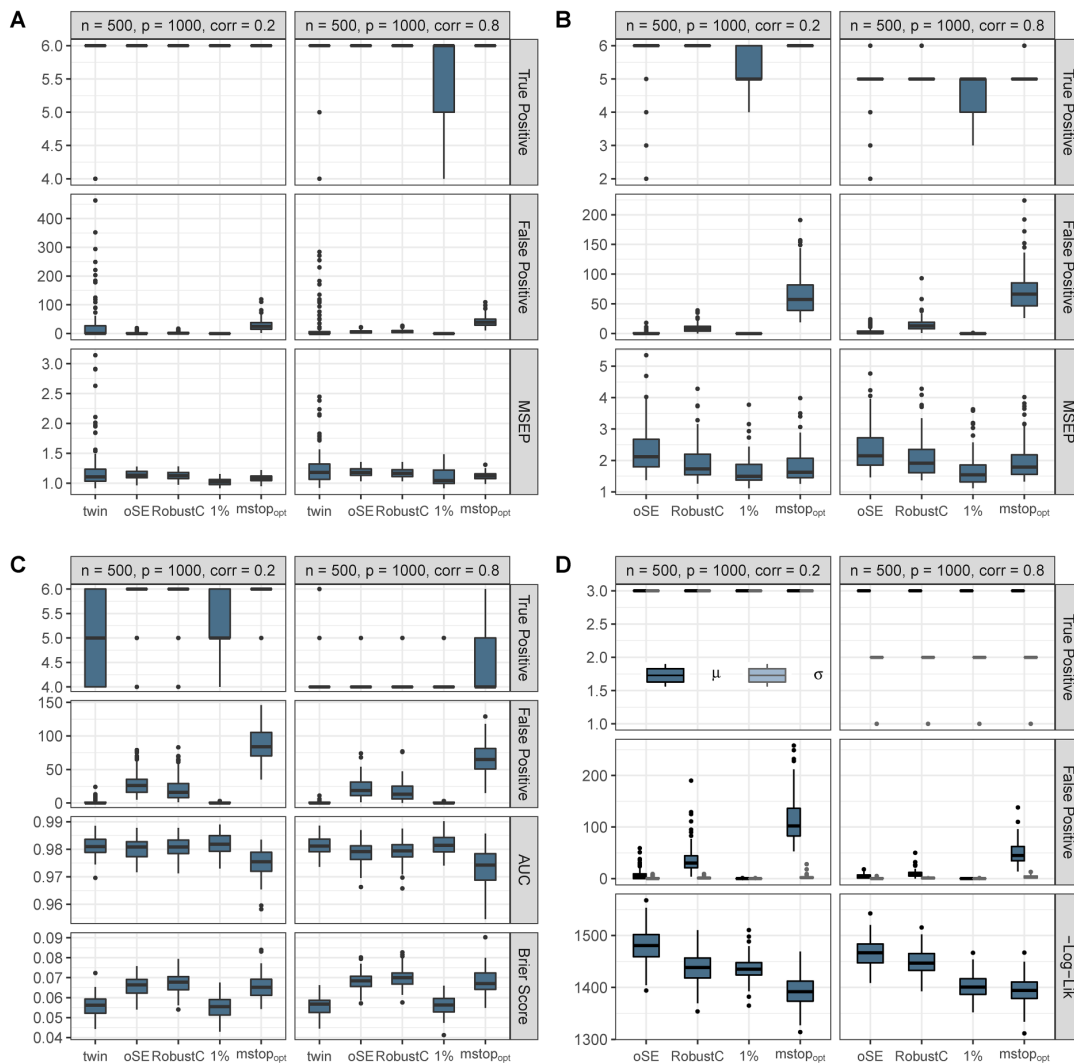


Figure 6. High-dimensional setting: Comparison of the oSE, RobustC, the deselection approach with $\tau = 0.01$ (1%) and the classical boosted model with $m_{\text{stop_opt}}$ regarding the true positives, false positives and the prediction performance for (A) linear regression, (B) non-linear regression, (C) logistic regression, and (D) distributional regression. Additional comparisons with twin boosting (twin) for the linear and logistic regression scenarios.

usually performed not as well as our approach. Only the AUC in Scenario B is very similar. For distributional regression (Scenario D), the classical approach yielded the best results concerning the negative log-likelihood but contained a lot of non-informative variables for both distributional parameters. The deselection approach reduced the false positives almost completely and had only a slightly worse prediction performance.

Figure 6 presents the results of the high-dimensional setting. As in Figure 5, the true positives, false positives, and predictive performances are shown. For the high-correlated cases of Scenario B, Scenario C, and Scenario D, the classical boosting model had already difficulties to select all informative variables. Concerning Scenario C only four of the six true positives were selected on average. In comparison with the classical approaches, the earlier stopping and deselection approaches resulted in an average lower number of true positives. For the false positives, we can observe a noticeable reduction with the earlier stopping strategies, but the number of false positives reduced even more with the deselection procedure and the final models contained almost only informative variables. The greatest reduction can be observed for Scenario D where the classical approach contains 100 false positives on average for parameter μ . After applying the deselection approach, the number of false positives decreased to almost zero with all informative variables still present. Due to the strong reduction of non-informative variables, in most cases, the deselection procedure showed a better predictive performance in comparison to earlier stopping and the classical boosting. Although in most of the simulation runs of Scenario C, not all informative variables were selected by the proposed deselection approach, it yielded a significantly lower Brier score and a better discriminatory power.

Furthermore, we compared the new deselection approach as well as earlier stopping strategies to twin boosting in the context of linear and logistic regression models (see Figures 5 and 6). Considering the results for twin boosting of Scenario A, the number of false positives was reduced (as for oSE, RobustC, and the new deselection procedure), but it shows larger variability, particularly in the high-dimensional setting. In one bootstrap sample, the model contained about 450 false positives (for

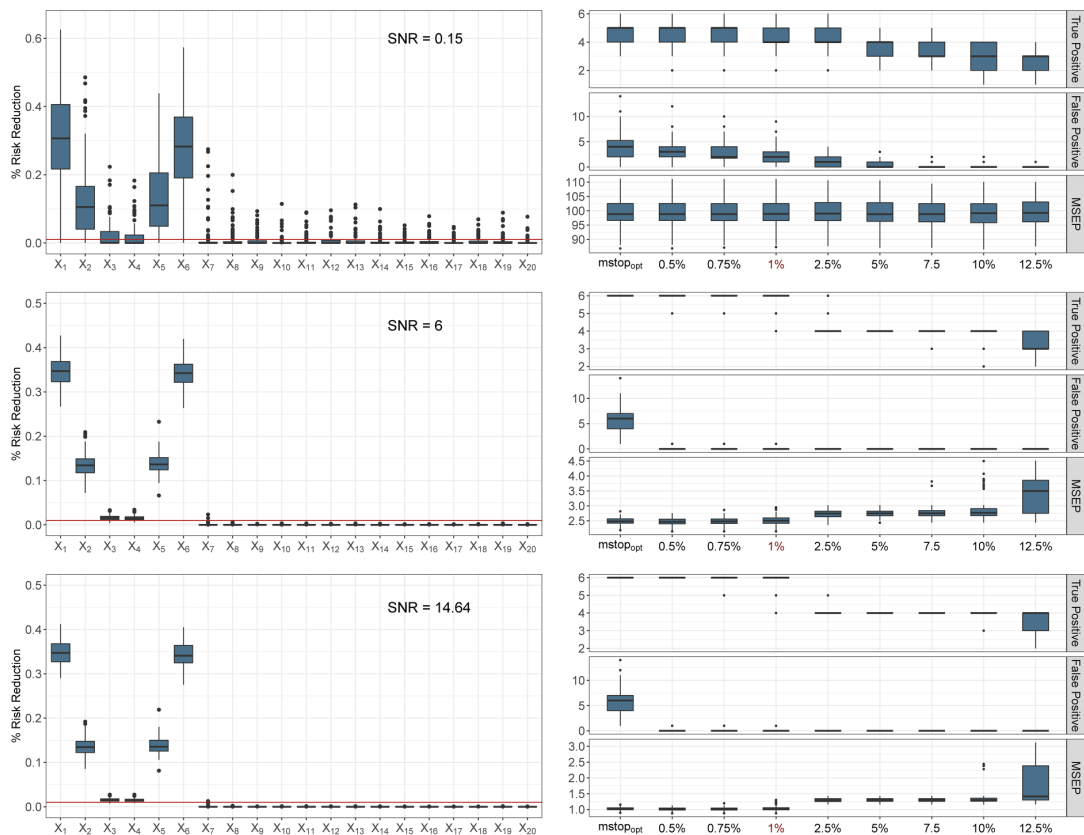


Figure 7. Relative risk reduction (in %) of each base-learner (left) and the variable selection and predictive performance for different τ values (right) for the low-dimensional setting with correlation $\rho = 0.8$ for increasing signal-to-noise ratios (from top with SNR = 0.15 to bottom with SNR = 14.64) for Scenario A.

low correlation). That is much more than we observed with the classical boosting approach, which had the maximum at about 100 selected non-informative variables. However, it should be noted that these different results for twin boosting are related to a different implementation. The highest decrease in false positives was observed for the deselection approach. The prediction accuracy was influenced by the outliers and also showed some higher MSE values for twin boosting. We obtained the best model for the deselection approach regarding the number of false positives as well as predictive performance. The results for logistic regression showed a slight reduction of the selected informative variables for each approach.

On average, twin boosting contained fewer false positives and had a better prediction accuracy than the earlier stopping strategies. Compared with the new deselection procedure, twin boosting tended to include more false positives, but was similar in terms of predictive performance. Here, twin boosting showed favorable properties in comparison to the linear regression model.

Finally, we investigate how the SNR affects the choice of the threshold parameter τ . Figure 7 shows the results for Scenario A concerning different SNRs with $\text{SNR} \in \{0.15, 6, 14.64\}$. In this case, we consider the low-dimensional setting (for illustrative purposes) with a correlation of 0.8, where $\text{SNR} = 14.64$ corresponds to the simulation setting presented before (results for a correlation of 0.2 and the high-dimensional settings are given in Supplemental Material A.3). On the left side, the relative risk reduction (in %) is depicted for each base-learner for the three different SNRs. The red horizontal line represents a threshold τ of 1%. The right side shows the corresponding results of the true positives and false positives as well as the predictive performance for various τ values. Overall, the relative risk reduction for all SNR values was very similar and the highest values always referred to informative variables, of which X_3 and X_4 showed the lowest risk reduction. The risk reduction for a SNR of 0.15 varied more and the non-informative variables showed a higher contribution to the risk reduction.

Considering the variable selection and predictive performance for various τ thresholds, the true positives and false positives for a SNR of 6 and 14.64 were very similar over the different τ values. For a SNR of 0.15, the classical boosting model had larger difficulties to identify all informative variables. Hence, it is also more challenging to deselect the non-informative variables without a further reduction in the true positives. Therefore, smaller τ values are more appropriate, causing less deselection and more noise variables, but the signal variables still remain in the model. For the other SNRs, only variable X_7 contributes to the risk reduction for small threshold values (0.5%, 0.75%, and 1%). Here, the relative risk reduction showed that a small τ value is sufficient to remove almost all false positives (red horizontal line for 1%). Furthermore, a higher τ value can lead not only to a reduction in informative variables included in the model but also to a worse MSE. Due to the noise, the predictive performance for the SNR of 0.15 was very poor and showed no discernible differences between the threshold values. However, τ values above 2.5% lead to a decrease in performance for larger SNRs. Furthermore, the previous simulation results have shown that a low value for τ (in this case 1%) reduces the number of false positives and additionally leads to a comparable predictive performance to classical boosting.

Overall, the number of false positives in the resulting models could be significantly reduced by earlier stopping or deselection as well as twin boosting compared to classical boosting. However, in most cases, the reduction of false positives for oSE and RobustC resulted in worse prediction performance. A comparison with probing for Scenario A, Scenario B, and Scenario C showed similar behavior (given in Supplemental Material A.2). Probing also led to a reduction in the number of false positives, but resulted in worse prediction performance, particularly for Scenario B. Furthermore, the earlier stopping strategies removed a few informative variables from the model in some settings.

The new procedure also deselected some informative variables from time to time, but removed the non-informative variables almost completely and resulted in favorable prediction performance. In some settings, the new approach even resulted in better predictive performance than the classical boosting model. Additional simulation results for the high-dimensional setting with a block structure for the covariance matrix are provided in Supplemental Material A.1 and showed very similar results compared with the Toeplitz covariance structure.

From the consideration of different SNRs, we conclude that the relative risk reduction attributed to a base-learner does not depend much on the overall SNR but on the distribution of the signal among the base-learners. To ensure that not too many informative variables are de-selected and to achieve a favorable predictive performance, our results suggest that the threshold value should be chosen rather small (e.g. 1%). For larger SNR values (6 and 14.64), almost all noise variables were eliminated even for small τ . Higher threshold values resulted in worse performance and a significant reduction of the informative variables. The choice of 1% for the threshold τ resulted in a reasonable tradeoff between sparsity and prediction performance in all considered settings. The *best* threshold, however, will always depend on the actual goal of the analysis and the general data situation.

Additionally, our approach also performed well compared to twin boosting, particularly for the linear regression model. An advantage of our method is the possibility to enhance variable selection for non-linear and distributional regression, which to the best of our knowledge is currently not available for twin boosting.

Table 1. Results for GCKD data in terms of the mean (sd) number of selected variables for the parameters μ and ϕ as well as the negative log-likelihood representing the prediction performance on the 1000 bootstrap replicates.

Model	μ	ϕ	–log-likelihood
classical boosted model	26.43 (7.10)	14.55 (6.00)	–1457.08 (40.24)
deselected ($\tau = 0.01$)	12.58 (1.39)	7.87 (1.61)	–1441.73 (39.41)
oSE	8.06 (2.68)	2.92 (1.63)	–1295.34 (37.88)
RobustC	7.63 (2.31)	2.70 (1.34)	–1290.14 (32.88)

Quality of life of chronic kidney disease patients

The following analysis aims to identify the most important predictors for the QoL of stage III chronic kidney disease patients based on an ongoing German cohort study (German Chronic Kidney Disease Study, GCKD). A similar analysis has already been published (cf. Mayr et al.⁹) and led to the selection of rather large models which partly motivated the current new methodological developments.

The analysis is based on beta regression,³⁷ which is a very flexible approach to model bounded outcome variables like proportions. It is also a well-known tool in the analysis of health-related QoL scores,^{38; 39} which typically range from 0 (lowest possible value) to 100 (highest possible value). The density function of a beta distribution with expected value μ and precision parameter ϕ is given by

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1,$$

where $\Gamma(\cdot)$ denotes the gamma function. In context of distributional beta regression, which refers to a generalized additive model for location, scale and shape (GAMLSS), we model μ and additionally ϕ in terms of several explanatory variables.

The GCKD study⁴⁰ is an ongoing cohort study for patients with stage III chronic kidney disease. We analyzed part of the cross-sectional baseline-data with $n = 3522$ observations and 54 explanatory variables. We aimed to select the most informative variables for the QoL of chronic kidney disease patients⁹ using the **R** add-on package **betaboost**. The effects of the predictors on the quality of life are represented by base-learners. For continuous covariates we incorporated spline effects as base-learners. For factor variables (e.g. education and exercise) we used linear base-learners providing joint updates of the

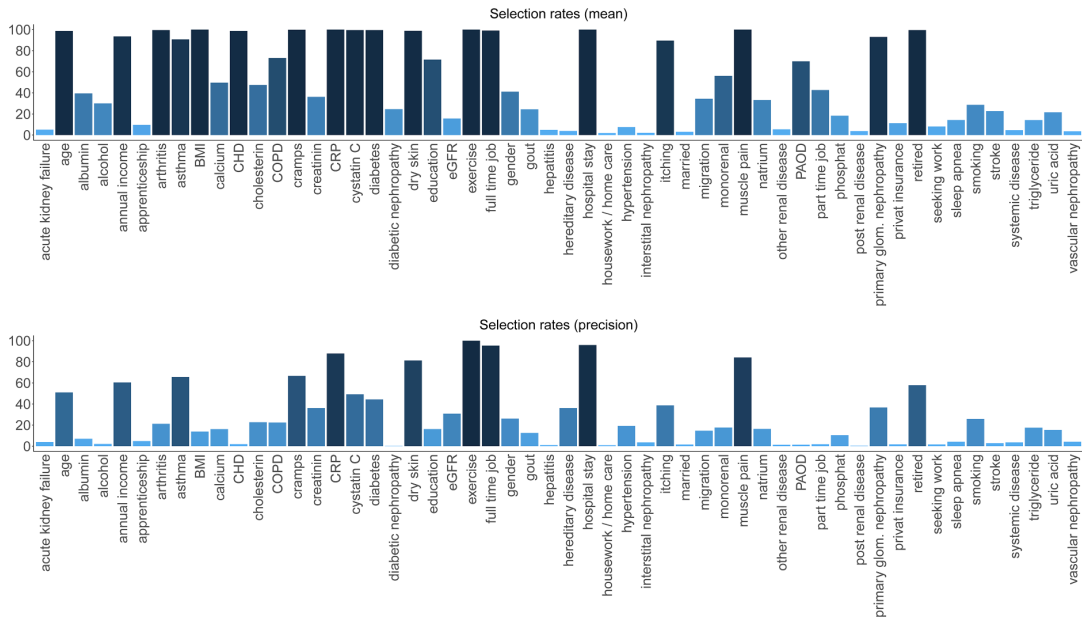


Figure 8. The selection rates of the explanatory variables for μ and ϕ of the classical boosting algorithm in 1000 bootstrap samples.

effects for the different categories in the boosting iterations. Therefore, our approach yields potential deselection (sparsity) on the full factor level and not on the level of different categories of a factor. Alternatively, multi-categorical factors may also be re-coded as several binary dummy variables, so that categories could be selected (and deselected) independently.

We drew 1000 bootstrap replicates and fitted a beta regression model without and with the new deselection procedure using $\tau = 1\%$ for each bootstrap sample (results for different τ values are given in Supplemental Material A.4). To evaluate the predictive performance of the resulting models, the negative log-likelihood was computed on the *out-of-bag* bootstrap samples. The optimal number of boosting iterations were selected via 10-fold cross-validation. For comparison, we additionally considered the oSE and RobustC methods.

Results

Table 1 displays the mean number (with standard deviations) of selected variables for μ and ϕ as well as the average negative log-likelihood for the different models on the 1000 bootstrap replicates. One can observe that more variables are included for the expected value than for the precision parameter. The earlier stopping strategies contain fewer variables than the proposed deselection approach for boosting.

In addition to Table 1, we consider the selection rates for each variable (for μ and ϕ) on the 1000 bootstrap replicates. Figure 8 displays the selection rates of the classical boosting approach. As described in Mayr et al.,⁹ the highest selection rates for parameter μ were obtained for age, body mass index (BMI), exercise, and variables related to pain such as arthritis, cramps, and muscle pain. Furthermore, variables that are indicators of kidney failure and inflammation also had higher rates, for example, cystatin C. For the precision parameter ϕ , 15 variables were included on average, with the highest rates for the variables exercise, employment in a full-time job and hospital stay.

The selection rates after additionally applying the deselection approach in Figure 9 show that the new procedure achieved a significant reduction in the number of included variables; some variables that were rarely selected by classical boosting were never included with the new approach (e.g. alcohol, gender). On the other hand, the variables with the highest selection rates from the classical model were still present at the highest selection rates.

To evaluate the predictive performance of the resulting models, we considered the negative log-likelihood on test data as a scoring rule. The results in Figure 10 suggest that the new deselection procedure outperforms the earlier stopping strategies oSE and RobustC. The smallest negative log-likelihood was obtained with the classical boosting model with an average value of 1457.08 (see Table 1), whereby we achieved a comparable performance for deselection with $\tau = 0.01$.

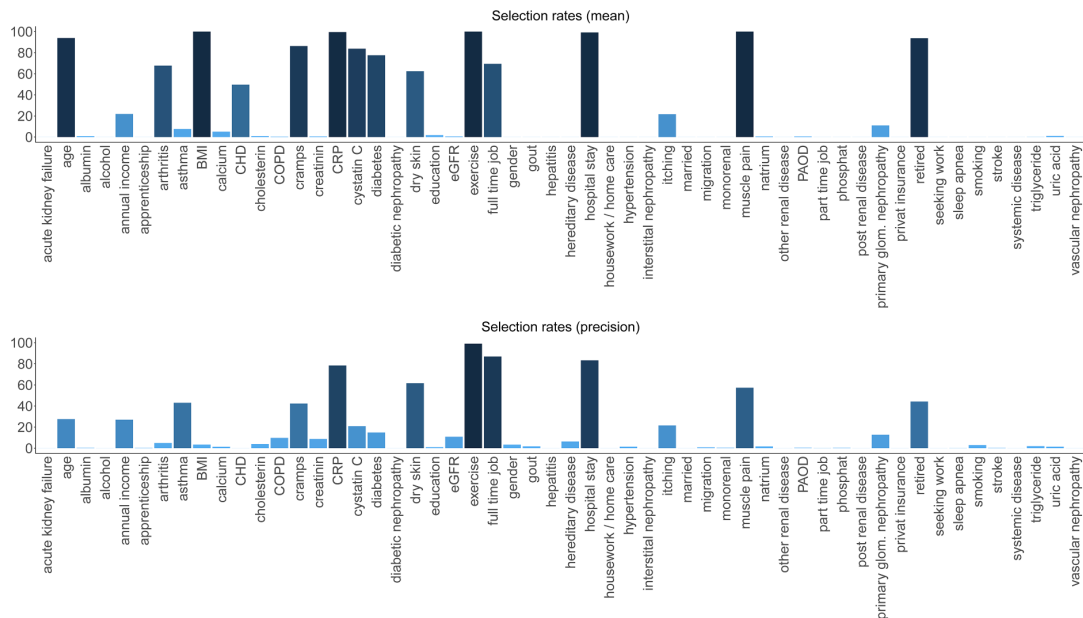


Figure 9. The selection rates of the explanatory variables for μ and ϕ after applying the new deselection approach with $\tau = 0.01$ in 1000 bootstrap samples.

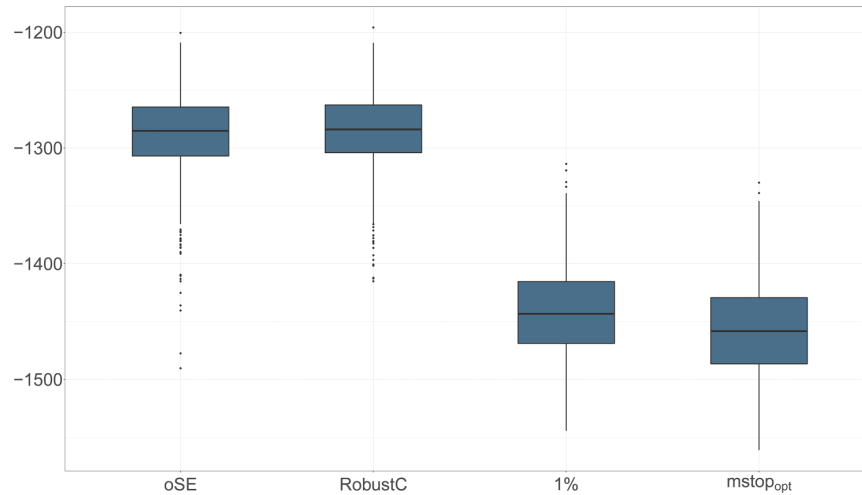


Figure 10. Negative log-likelihood of the oSE, RobustC, the deselection procedure with a threshold value of 1% and the classical boosted model on *out-of-bag* bootstrap samples.

Overall, the deselection approach based on 1% of the total risk reduction was able to enhance the sparsity of boosting models by selecting less predictors for the health-related QoL in chronic kidney disease patients (on average 12.6 for the expected value and 7.9 for the precision parameter from 54 candidate variables) in comparison to the classical boosting approach. However, many predictors contribute to the overall risk reduction to a small extent (see Supplemental Material, Figure A7). That indicates that the "true" underlying model is not as sparse as in the simulations. Earlier stopping strategies can further increase the sparsity, however leading to much poorer predictive performance on *out-of-bag* data (Figure 10). In comparison, the deselection approach leads to a slightly worse predictive performance in comparison to classical boosting, but yields much smaller and more interpretable models.

In practice, we have to deal with the tradeoff between sparsity and predictive performance, which is regulated by the threshold parameter τ . Therefore, higher variable deselection (i.e. larger values of τ) leads to smaller models but, at least in this considered application, also to poorer predictive performance.

Discussion and conclusion

The presented approach to deselect base-learners for enhanced variable selection in statistical boosting is a new technique to obtain sparser models with simpler interpretation via the removal of irrelevant predictors with negligible impact. As the deselection is based on the risk reduction, this approach is suitable for any type of base-learners, for example, linear models, splines, and spatial effects. Furthermore, the deselection approach can also be combined with a wide range of regression settings, including multi-dimensional optimization problems like distributional regression.

Compared to the similar twin-boosting,¹⁰ our approach actively deselects base-learners via a threshold value. This is somehow an analogy to stability selection^{41; 42} but our method focuses on providing a sparse prediction model instead of a set of stable predictors.^{43; 44} Furthermore, it does not include additional resampling steps. Other approaches for enhanced variable selection in the context of boosting focused on strategies for earlier stopping^{27; 13} which typically also increases the amount of shrinkage on effect estimates which might not be necessarily favorable. Our approach hence focuses on a vertical view on the regularization paths (in contrast to the horizontal view with earlier stopping) which has already been discussed in the literature on the lasso.^{15; 16}

The new approach is particularly suitable for high-dimensional data (with more potential predictors than observations) as one can obtain a simplified model with the most relevant variables yielding in many cases almost the same prediction accuracy as the classical boosting approach without deselection. Consequently, the interpretability of resulting prediction models improves, which makes their application in practice more likely.⁴⁵

The results of the simulation study suggest that our procedure can yield much sparser models by deselecting wrongly selected variables; in many cases deselection was associated with an almost complete elimination of false positives. In practice, one could assume that this might often lead also to a decreased prediction accuracy: the standard boosting approach

already selects the optimal prediction model by optimizing the stopping iteration. However, at least in some of the simulation settings, the deselection of false positives led even to a slightly improved prediction performance.

To select the most informative predictors for the health-related QoL of chronic kidney disease patients (GCKD study), the deselection procedure resulted in a drastically reduced set of variables compared to a recent analysis,⁹ which partly motivated the new methodological development. However, we did also observe a slight worsening of the model performance (w.r.t. the likelihood on test data).

The deselection procedure is controlled via a threshold value τ : it represents the minimum amount of total risk reduction which should be attributed to a corresponding base-learner in order to avoid deselection. In the simulation study, a threshold of $\tau = 0.01$ (i.e. 1% of total risk reduction) was considered to be appropriate overall. However, the general tradeoff between a more complex model with the highest possible prediction accuracy and a sparser, more interpretable model (higher *descriptive accuracy*⁴⁶ with potentially reduced prediction accuracy) should be guided by the researcher, depending on the research question and the context of the problem. As an alternative, the threshold parameter could also be chosen via resampling techniques or cross-validation which might further increase the performance but leads also to higher computational burden, particularly for high-dimensional data.

A limitation of our procedure is the assumption of sparsity. In cases where this is not fulfilled, it might deselect too many variables: If, for example, many predictors affect the model with minor impact (e.g. 200 variables with equal importance), our approach with $\tau = 0.01$ may deselect all variables. This is due to the dependency of our approach on the distribution of risk reduction across the base-learners. In theory, it would be beneficial to select τ based on the minimal signal strength, for example, the minimal risk reduction attributed to an informative predictor. As the truly informative variables, however, are unknown—this choice remains challenging in practical applications. An alternative technique, particular for non-sparse settings, could be to consider the *cumulative* risk reduction. Instead of considering only the risk reduction attributed to the corresponding base-learner, the cumulative risk considers the risk reduction of all base-learners that are to be deselected from the model. Thus, this procedure accounts for the complete tail of the base-learners with low importance. This variant would typically yield larger models when used with the same threshold. We investigated also the deselection via the cumulative risk reduction (results for the simulation and the application are given in Supplemental Material B). This alternative version is also implemented and available together with the corresponding code to reproduce the simulations and can be applied by specifying `method = "cumulative"`.

The favorable performance of our new approach motivates research in this direction in the future, in particular for distributional regression,⁴⁷ where sparse and interpretable models are of particular importance. For instance, deselection could be also extended to the level of distribution parameters in order to deselect a complete model-dimension (e.g. decreasing a GAMLSS to a GAM) when the contribution to the overall risk reduction is limited. Another line of potential research could focus on the combination of earlier stopping with deselection to avoid the disadvantage of exaggerated shrinkage.¹⁴

Altogether, we conclude that in our simulation and application the new deselection approach was able to outperform existing methods for earlier stopping, concerning the number of selected variables and the predictive performance. However, it should be noted that these approaches pursue various different goals like variable selection, prediction performance and/or interpretability. Fitting one model, that is able to achieve the best solution for all potential goals, simply might often not be feasible (cf. Hothorn⁴⁸).

Acknowledgements

The authors thank Benjamin Hofner for fruitful discussions on the underlying methodology of the new deselection procedure.


Declaration of conflicting interest


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant nos. 428239776, KL3037/2-1, MA7304/1-1). The GCKD study was funded by grants from the German Ministry of Education and Research (BMBF) (<http://www.gesundheitsforschung-bmbf.de/de/2101.php>; grant no. 01ER0804) and the KfH Foundation for Preventive Medicine (<http://www.kfh-stiftung-praeventivmedizin.de/content/stiftung>).

ORCID iDs

Annika Strömer  <https://orcid.org/0000-0002-1284-3318>

Andreas Mayr  <https://orcid.org/0000-0001-7106-9732>

Supplemental Material

Supplemental material for this article is available online. It contains an alternative deselection approach via the cumulative risk reduction and corresponding results of the simulation study and the GCKD data.

References

1. Fan J and Lv J. A selective overview of variable selection in high dimensional feature space. *Stat Sin* 2010; **20**: 101–148.
2. Chen TH, Chatterjee N, Landi MT, et al. A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information. *J Am Stat Assoc* 2020; **0**: 1–11.
3. Choi SW, Mak TSH and O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* 2020; **15**: 2759–2772.
4. Steyerberg EW and Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014; **35**: 1925–1931.
5. Sauerbrei W, Perperoglou A, Schmid M, et al. State of the art in selection of variables and functional forms in multivariable analysis: outstanding issues. *Diagnostic Prognostic Res* 2020; **4**: 1–18.
6. Bühlmann P and Hothorn T. Boosting algorithms: Regularization, prediction and model fitting. *Stat Sci* 2007; **22**: 477–505.
7. Mayr A, Hofner B, Waldmann E, et al. An update on statistical boosting in biomedicine. *Comput Math Methods Med* 2017; **2017**: 1–12.
8. Bühlmann P, Gertheiss J, Hieke S, et al. Discussion of the evolution of boosting algorithms and extending statistical boosting. *Methods Inf Med* 2014; **53**: 436–445.
9. Mayr A, Weinhold L, Hofner B, et al. The betaboost package – a software tool for modelling bounded outcome variables in potentially high-dimensional epidemiological data. *Int J Epidemiol* 2018; **47**: 1383–1388.
10. Bühlmann P and Hothorn T. Twin boosting: improved feature selection and prediction. *Stat Comput* 2010; **20**: 119–138.
11. Zou H. The adaptive LASSO and its Oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.
12. Breiman L, Friedman J, Stone CJ, et al. *Classification and Regression Trees*. Boca Raton: CRC Press, 1984.
13. Ellenbach N, Boulesteix AL, Bischl B, et al. Improved outcome prediction across data sources through robust parameter tuning. *J Classif*; 2021; **38**: 212–231.
14. Van Calster B, van Smeden M, De Cock B, et al. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Stat Methods Med Res* 2020; **29**: 3166–3178.
15. Zhou S. Thresholding procedures for high dimensional variable selection and statistical estimation. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09)*. Curran Associates Inc., Red Hook, NY, USA, 2009. 2304–2312.
16. Weinstein A, Su WJ, Bogdan M, et al. A Power Analysis for Knockoffs with the Lasso Coefficient-Difference Statistic; 2020. arXiv preprint arXiv:2007.15346.
17. Freund, Y. Boosting a weak learning algorithm by majority. *Information and Computation* 1995; **12**: 256–285.
18. Freund Y and Schapire RE. Experiments with a New Boosting Algorithm. In: *Proceedings of the Thirteenth International Conference on Machine Learning*. Bari, Italie: Morgan Kaufmann Publishers Inc, 1996. pp.148–156.
19. Friedman J, Hastie T and Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann Stat* 2000; **28**: 337–407.
20. Friedman J. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001; **29**: 1189–1232.
21. Mayr A, Binder H, Gefeller O, et al. The evolution of boosting algorithms. *Methods Inf Med* 2014; **53**: 419–427.
22. Mayr A, Binder H, Gefeller O, et al. Extending statistical boosting. *Methods Inf Med* 2014; **53**: 428–435.
23. Hofner B, Mayr A, Robinzonov N, et al. Model-based boosting in R: A hands-on tutorial using the R package mboost. *Comput Stat* 2014; **29**: 3–35.
24. Mayr A, Hofner B and Schmid M. The importance of knowing when to stop: A sequential stopping rule for component-wise gradient boosting. *Methods Inf Med* 2012; **51**: 178–186.
25. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33**: 1–22.
26. Hastie T, Tibshirani R and Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.
27. Thomas J, Hepp T, Mayr A and Bischl B. Probing for Sparse and Fast Variable Selection with model-based boosting. *Computational and Mathematical Methods in Medicine* 2017; **2017**: 1421409.
28. Hofner B, Boccutto L and Göker M. Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics* 2015; **16**: 144.
29. Wood SN. *Generalized Additive Models: An Introduction with R*. 2nd ed. Boca Raton: CRC Press, 2017.
30. Rigby RA and Stasinopoulos DM. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2005; **54**: 507–554.
31. Mayr A, Fenske N, Hofner B, et al. Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2012; **61**: 403–427.

32. Thomas J, Mayr A, Bischl B, et al. Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Stat Comput* 2017; **28**: 1–15.
33. Schmid M and Hothorn T. Boosting additive models using component-wise P-splines. *Comput Stat Data Anal* 2008; **53**: 298–311.
34. R Core Team.: R: A Language and Environment for Statistical Computing. Vienna, Austria; 2019.
35. Hothorn T, Bühlmann P, Kneib T, et al. mboost: Model-Based Boosting; 2020. R package version 2.9-2.
36. Wang Z. bst: Gradient Boosting; 2020. R package version 0.3-23.
37. Ferrari S and Cribari-Neto F. Beta regression or modeling rates and proportions. *J Appl Stat* 2004; **31**: 799–815.
38. Hunger M, Baumert J and Holle R. Analysis of SF-6D index data: is beta regression appropriate? *Value Health* 2011; **14**: 759–767.
39. Hunger M, Döring A and Holle R. Longitudinal beta regression models for analyzing health-related quality of life scores over time. *BMC Med Res Methodol* 2012; **12**: 144.
40. Eckardt KU, Bärthlein B, Baid-Agrawal S, et al. The German chronic kidney disease (GCKD) study: design and methods. *Nephrol Dial Transplant* 2012; **27**: 1454–1460.
41. Meinshausen N and Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010; **72**: 417–473.
42. Shah RD and Samworth RJ. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2013; **75**: 55–80.
43. Hofner B, Boccuto L and Göker M. Controlling false discoveries in high-dimensional situations: boosting with stability selection. *BMC Bioinformatics* 2015; **16**: 144.
44. Mayr A, Hofner B and Schmid M. Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection. *BMC Bioinformatics* 2016; **17**: 288.
45. Wyatt JC and Altman DG. Prognostic models: clinically useful or quickly forgotten? *Br Med J* 1995; **311**: 1539–1541.
46. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, and Yu B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 2019; **116**: 22071–22080.
47. Stasinopoulos MD, Rigby RA, Heller GZ, et al. *Definitions, methods, and applications in interpretable machine learning*. CRC Press, 2017.
48. Hothorn T. Invited discussion on “Meinshausen and Bühlmann: Stability selection”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010; **72**: 463–464.

3.5 Publication E: Enhanced variable selection for boosting sparser and less complex models in distributional copula regression

Strömer A, Klein N, Staerk C, Faschingbauer F, Klinkhammer H, Mayr A. Enhanced variable selection for boosting sparser and less complex models in distributional copula regression. *Statistics in Biosciences* 2025.

<https://doi.org/10.1007/s12561-025-09491-8>

Supplementary information can be found at:

<https://doi.org/10.1007/s12561-025-09491-8>

Implementations are available on GitHub:

<https://github.com/AnnikaStr/ComplRedBoostCop>



Enhanced Variable Selection for Boosting Sparser and Less Complex Models in Distributional Copula Regression

Annika Strömer^{1,2} · Nadja Klein³ · Christian Staerk^{4,5} ·
Florian Faschingbauer⁶ · Hannah Klinkhammer^{2,7} · Andreas Mayr¹

Received: 17 April 2024 / Revised: 17 March 2025 / Accepted: 18 May 2025
© The Author(s) 2025

Abstract

Structured additive distributional copula regression allows to model the joint distribution of multivariate outcomes by relating all distribution parameters to covariates. Estimation via statistical boosting enables accounting for high-dimensional data and incorporating data-driven variable selection, both of which are useful given the complexity of the model class. However, as known from univariate (distributional) regression, the standard boosting algorithm tends to select too many variables with minor importance, particularly in settings with large sample sizes, leading to complex models with difficult interpretation. To counteract this behavior and to avoid selecting base-learners with only a negligible impact, we combine the ideas of probing, stability selection, and a new deselection approach with statistical boosting for distributional copula regression. In simulations and an application to the joint modeling of weight and length of newborns, we find that all proposed methods enhance variable selection by reducing the number of false positives. However, only stability selection and the deselection approach yield similar predictive performance to classical boosting. Finally, the deselection approach is better scalable to larger datasets and leads to competitive predictive performance, which we further illustrate in a genomic cohort study from the UK Biobank by modeling the joint genetic predisposition for two phenotypes.

Keywords Distributional regression · Multiple outcomes · Probing · Stability selection · UK Biobank · Variable selection

1 Introduction

Statistical boosting, an iterative, sequential fitting algorithm for statistical models originating from machine learning [6], has gained increasing interest as an alternative to classical (penalized) maximum likelihood estimation (PMLE) or

Extended author information available on the last page of the article

Bayesian inference. Since boosting is well suited for high-dimensional and complex data problems, it is also a useful tool for distributional regression, where the number of candidate models is typically large. Distributional regression generally has the aim of estimating complete conditional distributions of a quantity of interest as a function of covariates (see e.g., [14], for a recent review). A convenient framework for univariate distributional regression is the class of generalized additive models for location, scale, and shape (GAMLSS [27]), which allow relating each distribution parameter of a parametric response distribution to covariates. However, in the classical PMLE-based implementation, the response is restricted to be univariate and the complexity of the predictors is limited due to numerical instabilities when it comes to selecting smoothing parameters for regularization. While Bayesian estimation of GAMLSS based on Markov chain Monte Carlo simulations [15, 39] allows to overcome the latter issues, it is notoriously slow when the number of observations n and/or the number of covariates p is large. In such scenarios, statistical boosting is particularly beneficial as also demonstrated in various applications and extensions of the original boosting algorithm (see e.g., [16, 31]).

In this paper, we are not only concerned with situations, where either n or p is large or where even $p \gg n$, but particularly when the outcome Y is multivariate. By modeling dependent outcomes together, we can gain a better understanding of the relationship and identify relevant factors affecting their association. An example is the consideration of multiple phenotypes from deeply phenotyped cohort studies in genetic epidemiology. For distributional regression modeling of multiple outcomes, one approach is to consider the joint parametric distribution. A popular alternative in this context are copulas that offer increased flexibility by allowing the use of different marginals and different dependence structures through the copula function. This approach has been the subject of ongoing research and there is rich literature on copula modeling with regression data (see e.g., [24, 40, 43], for recent examples).

Statistical boosting was extended to multivariate distributional regression towards parametric distributions [34] and also using copula [9]. However, while this is done conceptually in these works, some practical aspects are still challenging. One challenge is that while boosting has an implicit variable selection mechanism, it often leads to relatively *large* models, that is, models with many included covariates despite having small to negligible effects. This happens because the algorithm typically optimizes prediction accuracy without explicitly considering sparsity. This was also observed for boosting copula regression, where especially the sub-models for the location parameter did result in rather large numbers of selected covariates [9]. This behavior is particularly undesirable in situations with a large number of candidate variables p , where sparse and interpretable models are practically relevant. Therefore, variable selection is of great importance, not only in the context of boosting (see e.g., [13, 38]). In distributional copula regression, a further interesting question that arises is how to decide in a data-driven manner if the overall complexity of the model could be reduced. In this context, the aim of statistical modeling is to select a model that is as complex as needed but also as simple as possible to facilitate efficient estimation and interpretability. For example, if certain distribution parameters like the association parameters do

not depend on covariates, then simpler univariate models might also be suitable. Statistical boosting can help to answer this question.

To tackle these yet unaddressed practical challenges in boosting distributional copula regression, we incorporate three existing approaches for refining variable selection within this framework. All three approaches have been already proposed or extended to boosting, but have never been integrated into boosting multivariate distributional regression via copulas. This new combination aims to reduce the complexity of the model, particularly when dealing with high-dimensional data. The three considered existing approaches for enhanced variable selection are the following: (i) Stability selection [22], which has been extended to boosting univariate distributional regression [37]; (ii) probing [36], which was proposed for boosting simple mean regression models, shifts the focus of early stopping from prediction accuracy directly to variable selection; and (iii) deselection [33], the newest approach, which pragmatically deselects base-learners that do not contribute enough to the overall model performance.

We initially investigate the performance of these three methods on simulated data, where the true data-generating process is known. Afterward, we consider two real data applications: first, the joint modeling of the weight and length of newborns (as in [9]), and second, the modeling of the joint genetic disposition towards continuous phenotypes based on a large cohort study (UK Biobank).

2 Methods

2.1 Boosting Distributional Copula Regression

In this section, we briefly review distributional copula regression models focusing on the bivariate case of two continuous outcomes and how statistical boosting algorithms can be applied for model estimation.

2.1.1 Distributional Copula Regression Models

A flexible modeling approach for the joint analysis of two continuous response variables $\mathbf{Y} = (Y_1, Y_2)^T$ in terms of covariates are bivariate copula regression models, which describe the dependence structure through a copula [23]. According to Sklar's theorem, the joint conditional cumulative distribution function (CDF) of two responses given covariate information \mathbf{x} can be written as

$$F(y_1, y_2 | \boldsymbol{\theta}) = C[F_1(y_1 | \boldsymbol{\theta}^{(1)}), F_2(y_2 | \boldsymbol{\theta}^{(2)}) | \boldsymbol{\theta}^{(c)}],$$

where $F_1(\cdot | \boldsymbol{\theta}^{(1)})$ and $F_2(\cdot | \boldsymbol{\theta}^{(2)})$ are the marginal conditional CDFs of the two responses $Y_1 = y_1$ and $Y_2 = y_2$ which are uniformly distributed on $[0, 1]$. The copula function $C(\cdot, \cdot | \boldsymbol{\theta}^{(c)})$ contains the information about the dependence structure between the two outcomes and is unique when the responses are continuous. The vector $\boldsymbol{\theta} = \{(\boldsymbol{\theta}^{(1)})^T, (\boldsymbol{\theta}^{(2)})^T, (\boldsymbol{\theta}^{(c)})^T\}^T$ contains the model parameters $k = 1, \dots, K$ of the marginal distributions and the copula, whereby all components of $\boldsymbol{\theta} \equiv \boldsymbol{\theta}(\mathbf{x})$ can be linked to a covariate vector via additive predictors and appropriate link functions.

The representation of the joint conditional CDF via a copula allows the separation of the marginal distributions and the dependence structure; different copula functions allow different structures to be modeled. The Clayton copula, for example, can capture asymmetric dependence (so-called lower tail dependence), where the two responses show a stronger positive association for smaller values than for larger values. In our work we will focus on Gaussian, Clayton, and Gumbel copulas (cf. [23]) to represent no, lower, and upper tail dependencies.

The joint density $f(y_1, y_2 | \theta)$ of a distributional copula regression model can be expressed via

$$\begin{aligned} f(y_1, y_2 | \theta) &= \frac{\partial^2}{\partial F_1 \partial F_2} F(y_1, y_2 | \theta) \\ &= c[F_1(y_1 | \theta^{(1)}), F_2(y_2 | \theta^{(2)}) | \theta^{(c)}] f_1(y_1 | \theta^{(1)}) f_2(y_2 | \theta^{(2)}), \end{aligned}$$

where $f_1(\cdot | \theta^{(1)})$ and $f_2(\cdot | \theta^{(2)})$ are the marginal probability density functions and $c(\cdot, \cdot | \theta^{(c)})$ is the copula density of C . Based on our applications, the most relevant marginal distributions in this work are the log-logistic and the log-normal distributions.

Finally, for a dataset of n independent pairs $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^n$ of bivariate responses $\mathbf{y}_i = (y_{i1}, y_{i2})^\top$ with covariate information \mathbf{x}_i , the joint log-likelihood function is given by

$$l(\theta) = \sum_{i=1}^n \log \{c[F_1(y_{i1} | \theta^{(1)}), F_2(y_{i2} | \theta^{(2)}) | \theta^{(c)}]\} + \sum_{i=1}^n \sum_{d \in \{1,2\}} \log \{f_d(y_{id} | \theta^{(d)})\}.$$

2.1.2 Structured Additive Predictors

In distributional copula regression, each distribution parameter θ_k , $k = 1, \dots, K$ is modeled via a structured additive predictor η_k [3, 32] with parameter-specific monotonic link functions g_k , such that $g_k(\theta_k) = \eta_k$ and $g_k^{-1}(\eta_k) = \theta_k$, where g_k^{-1} is the inverse of g_k . The additive predictors η_k depend on (possibly different) subsets of \mathbf{x} ,

$$g_k(\theta_k) = \eta_k = \beta_{0k} + \sum_{j=1}^{p_k} f_{jk}(\mathbf{x}_{jk}), \text{ for } k = 1, \dots, K,$$

where β_{0k} are the intercepts and each f_{jk} , $j = 1, \dots, p_k$, represents functional effects of covariates \mathbf{x}_{jk} , whereby \mathbf{x}_{jk} is a covariate subset of \mathbf{x} . The effects can be chosen in a flexible manner [4], for instance we incorporate linear and non-linear effects in Sections 3 and 4. Linear effects can be represented by $f_{jk}(\mathbf{x}) = \mathbf{x}_{jk}^T \boldsymbol{\beta}_{jk}$ where $\boldsymbol{\beta}_{jk}$ are the regression coefficients. Non-linear effects can be included using appropriate basis functions, such as B-splines.

2.1.3 Estimation Via Model-Based Boosting

Component-wise gradient boosting with regression-type base-learners, also referred to as *statistical boosting* [18], originates from the gradient boosting approach of [6], who translated the original concept from the machine learning literature to statistical modeling. Its basic idea is to iteratively minimize a pre-specified loss function \mathcal{L} by fitting the so-called base-learners separately to the negative gradient \mathcal{L} and by then adding only a small amount of the “best-fitting” base-learner—that is, the base-learner that yields the steepest descent in the direction of the current gradient—to the overall regression predictor in each step of the boosting algorithm. In our case, a base-learner represents one effect in the additive regression predictor (see [12] for a detailed overview). In this way, the overall predictor is built sequentially, where more and more variables are selected the longer the algorithm runs, such that *early stopping* yields implicit variable selection. In likelihood-based statistical boosting, the loss \mathcal{L} is the negative log-likelihood $l \equiv l(\theta)$, but more general functionals such as proper scoring rules are possible.

The boosting algorithm is a flexible alternative to classical estimation approaches. It has several practical advantages, such as dealing with high-dimensional data in which classical inferential methods are no longer applicable. As mentioned above, the algorithm performs data-driven variable selection, which is controlled by the number of boosting iterations m_{stop} [19]: Variables whose corresponding base-learner has never been selected until m_{stop} is reached are excluded from the final model. Therefore, the number of boosting iterations is the main tuning parameter and is typically optimized by cross-validation or resampling techniques. Another parameter of the algorithm is the fixed step length ν , with which the best-fitting base-learner is multiplied before being included into the predictor. This parameter is set to a small fixed value within the range of $0 < \nu < 1$ [28]. For boosting copula regression, [9] suggests a value of $\nu = 0.01$.

In the boosting approach for distributional copula regression [9], all distribution parameters are modeled simultaneously by combining the properties of GAMLSS and the main features of statistical boosting. In every iteration, the partial derivatives $u_k = \partial l / \partial \theta_k$ of the negative log-likelihood l with respect to the different distribution parameters θ_k are calculated and each base-learner $h_{jk} \equiv f_{jk}(\mathbf{x}_{jk})$ is separately fitted to the gradient. Then, the best-fitting base-learner (and the corresponding update) for each distribution parameter is determined and compared across the different dimensions. Only the overall best-performing update is finally performed using a non-cyclic version of the algorithm [37]. For more details on fitting distributional copula regression via boosting, we refer to [9].

2.2 Complexity Reduction and Enhanced Variable selection

In the following, we present different techniques for enhanced variable selection that we will integrate in our boosting distributional copula framework. Probing (Sect. 2.2.1) had been introduced to statistical boosting by [36] and since then been applied or used as a

benchmark approach for mean regression models with only one dimension [2, 31] or joint models of time-to-event and longitudinal data [8]. Stability selection (Sect. 2.2.2) is a more general approach [22] and has been introduced to boosting mean regression models by [11], before the approach was extended to the context of univariate distributional regression [37]. Deselection (Section 2.2.3) is the most recent enhancement and was directly introduced for boosting mean regression as well as distributional regression [33]. None of the three approaches have ever been extended towards boosting multivariate distributional regression or even to copula regression. In the process of integrating these enhanced variable selection approaches in our framework, we also allow for constant distribution parameters that do not depend on covariates. This is particularly attractive in copula regression, where for example copula parameters not depending on covariates reflect situations in which the dependence structure between the outcomes does not vary across observations with distinct feature values. This can lead to a substantial reduction in the complexity of the final model.

2.2.1 Probing

Probing is based on the inclusion of random noise variables, the so-called probes, to determine the stopping iteration by stopping when the algorithm starts selecting those (Algorithm 1): First, randomly generated shuffled versions (probes) of the covariates are added to the original dataset. Second, a boosting model is fitted on the expanded dataset and the algorithm stops when the first probe is selected. The idea is that, in each iteration, the base-learner with the highest loss reduction is updated and the selection of a probe means that the best possible improvement is based on information known to be unrelated to the outcome. Because each parameter may depend on a potentially different set of variables, the randomly shuffled probes are simply added for each of the distribution parameters. In our model class, the distribution parameters may represent parameters of the marginal distributions or the copula. In each boosting iteration, a single base-learner is updated, i.e., the algorithm stops when the first probe is selected for any of the distribution parameters. While probing does not require optimizing the stopping criterion via computationally expensive cross-validation or resampling, it optimizes towards sparse models and does not maximize prediction performance. As a consequence, probing typically yields sparse models with strongly regularized predictor effects [36].

Algorithm 1 Probing for boosting distributional copula regression.

-
- 1: Shuffle probes \tilde{x}_{jk} for each of the covariates x_{jk} with $j = 1, \dots, p_k$ and $k = 1, \dots, K$.
 - 2: Perform boosting on the expanded set of variables $x_1, \dots, x_{p_k}, \tilde{x}_1, \dots, \tilde{x}_{p_k}$ for each distribution parameter $\theta_k, k = 1, \dots, K$.
 - 3: Stop when the first probe \tilde{x}_{jk} of any distribution parameter is selected.
 - 4: Use final model from the previous iteration (containing only original variables).
-

2.2.2 Stability selection

A popular enhanced variable selection technique is *stability selection*, which yields a stable set of covariates by repeated model fitting using subsamples of the original dataset [22, 29]. In the context of boosting, Thomas et al. [37] introduced stability selection for boosted GAMLSS. As outlined in Algorithm 2, the general idea is to draw B random subsets of the data with size $\lfloor n/2 \rfloor$ of the original dataset and to fit separate boosting models for each subset. The boosting algorithm runs on each subset until a pre-specified number of covariates q have been selected. Every variable has a selection frequency defined by the fraction of subsets in which the variable j was selected. If the selection frequency exceeds the threshold π_{thr} , the variable is considered stable and is included in the final model fit [11]. Stability selection provides a sparse solution, controlling the number of false discoveries by defining an upper bound for the per-family error rate (PFER), i.e., the expected number $\mathbb{E}(V)$ of noninformative variables included in the final model. The upper bound is given by $\mathbb{E}(V) \leq q^2 / ((2\pi_{\text{thr}} - 1)p)$, where $p = \sum_{k=1}^K p_k$ is the total number of predictor variables and q the number of selected variables.

For practical use, the most important aspect is the choice of the parameters q , π_{thr} and PFER, whereby the PFER can be derived from the upper bound and visa versa. It is recommended to specify PFER and either q or π_{thr} [12]. Meinshausen and Bühlmann [22] state that the number of selected base-learners q should be chosen sufficiently large concerning the informative variables, or at least as high as the number of informative variables, which, however, are usually unknown. The threshold π_{thr} should be in the range of $\pi_{\text{thr}} \in (0.6, 0.9)$, meaning a variable should be selected in more than half of the fitted models in order to be considered stable. The choice of B is of minor importance as long as it is sufficiently large to ensure accurate estimation of $\hat{\pi}_j$ across various scenarios [22].

Algorithm 2 Stability selection for boosting distributional copula regression.

-
- 1: **for** $b = 1, \dots, B$ **do**
 - 2: Select a random subset from the data of size $\lfloor n/2 \rfloor$.
 - 3: Fit a boosting model until q base-learner are selected.
 - 4: **end for**
 - 5: Compute the relative selection frequencies per base-learner $\hat{\pi}_j = \frac{1}{B} \sum I_{j \in \hat{S}_b}$, where \hat{S}_b denotes the set of selected base-learner.
 - 6: Select the stable set of base-learner $\hat{S}_{\text{stable}} := \{j : \hat{\pi}_j \geq \pi_{\text{thr}}\}$.
 - 7: Fit a boosting model with the stable set of base-learners.
-

2.2.3 Deselection of Base-Learners

Another approach to encourage variable selection and sparsity is to deselect and remove base-learners with a negligible impact on the model's predictive

performance. The general idea is to start with a classical boosted model tuned by cross-validation or resampling techniques. Then, the base-learners that were selected but only have a minor impact on the model are identified and deselected. Afterward, the model is boosted again with the remaining variables. This idea was introduced by [33] for univariate GAMLSS and is now extended to distributional copula regression. The importance of a base-learner is based here on the risk reduction and can be defined for base-learner j after m_{stop} boosting iterations with

$$R_j = \sum_{m=1}^{m_{\text{stop}}} I(j = j^{*[m]})(r^{[m-1]} - r^{[m]}), \quad j = 1, \dots, p,$$

where I denotes the indicator function and $j^{*[m]}$ is the selected base-learner in iteration m . Furthermore, $r^{[m-1]} - r^{[m]}$ represents the risk reduction in iteration m , for risks $r^{[m]}$ and $r^{[m-1]}$ at iterations m and $m-1$, respectively. Note that in the case of distributional copula regression, all distribution parameters are considered together and each parameter $\theta_k, k = 1, \dots, K$ can depend on a different number of variables p_k . Here, we do not distinguish between the different parameters, such that $p = \sum p_k$.

For a given threshold $\tau \in (0, 1)$, we deselect base-learner j if

$$R_j < \tau \cdot (r^{[0]} - r^{[m_{\text{stop}}]}),$$

where $r^{[0]} - r^{[m_{\text{stop}}]}$ represents the total risk reduction and R_j denotes risk reduction attributable to base-learner j . In other words, only base-learners whose contribution R_j to the total risk reduction is larger than the relative τ threshold (e.g., 1% [33]) will remain in the model after the deselection step.

Algorithm 3 Deselection for boosting distributional copula regression.

-
- 1: Initial boosting:
Tune m_{stop} based on cross-validation or resampling techniques (early stopping).
 - 2: Deselection:
Identify the base-learners with minor impact on the risk reduction according to $R_j < \tau \cdot (r^{[0]} - r^{[m_{\text{stop}}]})$ and remove them from the model.
 - 3: Final boosting:
Boost again with the remaining variables and the m_{stop} of Step 1.
-

2.2.4 Illustration of the Different Approaches

Figure 1 displays the coefficient paths resulting from the classical boosted copula regression and the final models after applying the different approaches for reducing the model complexity on simulated data (more details on this example can be found in the supplement, Section A). Overall the coefficient paths of the different approaches yield similar final models. Applying probing leads to earlier

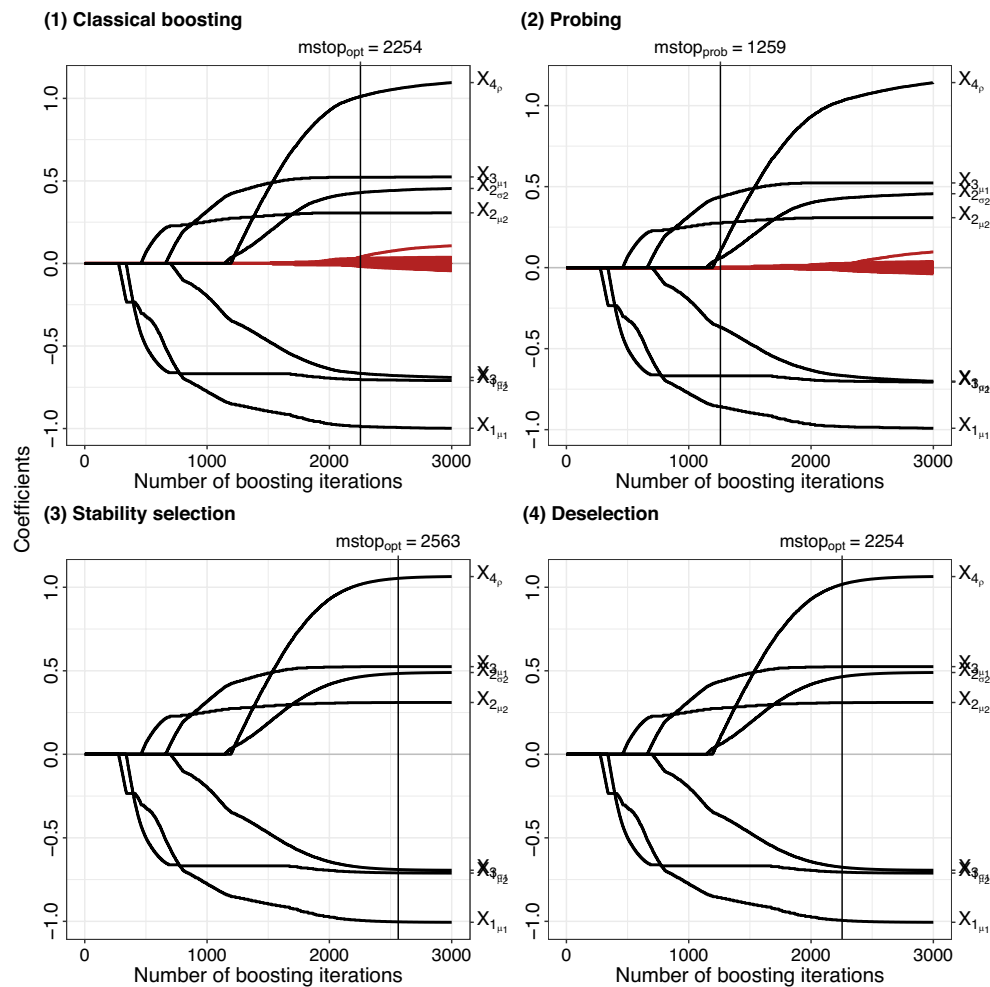


Fig. 1 The resulting coefficient paths along the number of boosting iterations for a simulated example (for more details see the supplement, Section A for the classical boosting, probing, stability selection, and the deselection approach (1%). The coefficient paths of the informative variables are colored in black, the noninformative in red. The intercept was removed for clarity. For stability selection and deselection, only the final model is plotted

stopping than the classical model with a stopping iteration at 1259 iterations. Therefore, the effect estimates are shrunk and fewer variables are included in the model (all informative but also one noninformative variable). As described in Sect. 2.2.1, the shrinkage of the effect estimates might not be optimal for predictive performance. The resulting model for stability selection is shown in the third plot, with selection frequencies across the B subsets for the different base-learners provided in the supplement (Figure A1). The performance of stability selection depends strongly on the choice of the parameters, here we choose $q = 20$ and $PFER = 5$, but for example smaller q and $PFER$ would lead to worse results as most informative variables would not be included, leading to poorer predictive performance. The choice of q is informed by our comprehensive simulations

detailed in Sect. 3. For a representation of how different q values impact the results, we refer to Appendix B (Table A3), for details.

The deselection approach with a threshold value of 1% is similar to stability selection. Stable covariates are the ones with the highest risk reduction in the deselection approach. The final deselection model contains also here only the informative variables with the same number of boosting iterations as the classical model and similar coefficient estimates. The corresponding risk reduction for the different variables can be found in Figure A2 in the supplement with different threshold values (0.1 and 1%). Higher threshold values would lead to the elimination of informative variables, whereby for smaller values such as 0.1% (dotted line in the risk reduction plot Figure A2 in the supplement) noninformative variables would remain in the model.

2.3 Computational Details and Implementation

Boosting for distributional copula regression is implemented via the R package **gamboostLSS**. For tuning of m_{stop} , cross-validation, resampling techniques, or evaluation on a single test dataset can be used. The process is facilitated using a provided function that directly works on the model object.

From a computational and implementation perspective, probing can be very easily utilized, because no computationally intensive techniques for optimizing the stopping iteration and no additional tuning parameters are required. Stability selection for copula regression can be realized using the fitted boosting model and the `stabs()` function in the package **gamboostLSS**. One needs to specify two of the parameters beforehand, the per-family error rate and either the number of base-learners q or the threshold π_{thr} . The stopping iteration of the boosting model has to be chosen sufficiently large so that the q base-learners can be selected. The function returns the stable set of base-learner for each distribution parameter. To obtain the final model, one can again run a boosting model with only these stable base-learners. As in any classical statistical boosting model, the stopping iteration needs to be optimized by cross-validation, resampling techniques, or on an additional validation data set (if available). Moreover, the function encompasses various options for assumptions. It is important to note that the described approach in Sect. 2.2.2 does not involve any additional assumptions (`assumption = "none"`). The number of subsamples B should be sufficiently large to ensure reliable results. Typically, it is set to $B = 100$ [22]. By default, the implementation uses complementary pairs for subsampling with $B = 50$, which means $2 \cdot B$ subsamples in total.

The implementation of the deselection approaches is available at GitHub <https://github.com/AnnikaStr/ComplRedBoostCop> and is accessed with the `DeselectBoost()` function, which requires a boosting model with early stopping and the specification of an appropriate threshold value (e.g., 1%). The refitting of the model with the remaining base-learners to obtain the final model is already included in the function.

3 Simulations

To evaluate the performance of the different approaches for reducing the model complexity of boosted bivariate distributional copula regression models, we conducted a simulation study. We compared probing, stability selection, and the deselection of base-learners with a focus on their variable selection properties, the prediction performance, and runtime. Our specific objectives were to determine the following: (i) Can the variable selection approaches identify the truly informative variables while decreasing the number of false positives? (ii) How do the approaches perform in comparison to each other? (ii) Can the complexity of the model be reduced by simplifying complete additive predictors for distribution parameters to an intercept?

A detailed description of the simulation design of the following scenarios can be found in the supplement, Section B. Furthermore, here we provide a descriptive summary of the simulation results, while detailed numerical results can be also found in the supplement, Section B. All codes to reproduce the results can be found on GitHub <https://github.com/AnnikaStr/ComplRedBoostCop>. The simulations were conducted in R using the add-on package **gamboostLSS** for estimating the copula regression models. The **copula** and **gamlss** packages were used for data generation.

3.1 Simulation Design

To investigate these questions, we considered four different bivariate scenarios for continuous outcomes, each with five distribution parameters (marginal means μ_1 and μ_2 , marginal variances σ_1^2 and σ_2^2 , and association parameter ρ):

Scenario A Same simulation setup as in [9] with four informative variables x_1, \dots, x_4 . Cubic P-splines with 20 equidistant knots were included as base-learners. The log-normal and log-logistic distributions were used as marginal distributions.

Scenario B Modification of Scenario A:

1. σ_1 does not depend on explanatory variables.
2. ρ does not depend on explanatory variables.

Scenario C More informative variables: Ten informative variables for each distribution parameter $p_k = 10, k = 1, \dots, 5$ with x_1, \dots, x_{50} with Gaussian marginal distributions for both outcomes. The base-learners correspond to simple linear models.

Detailed insights for each scenario are provided in Supplement B. A total of 100 simulation runs were performed for each simulation setting. For each scenario, $n = 1000$ observations were considered, where the covariates x_1, \dots, x_p were independently drawn from a uniform distribution on $(-1, 1)$. The simulations cover a

low-dimensional case ($p < n$) with $p = 20$ variables for **Scenario A**, B.1, and B.2 and $p = 200$ for **Scenario C**. Furthermore, a high-dimensional case ($p > n$) with $p = 1000$ variables was investigated for each scenario. The Gaussian, the Clayton, and the Gumbel copula were considered. For fitting the models, all covariates were considered for each distribution parameter simultaneously. The stopping iteration m_{stop} was optimized by minimizing the empirical risk on an additional validation dataset with 1500 observations. For all simulations, the step length of the boosting algorithm was set to a fixed value of $\nu = 0.01$ as suggested in [9] for boosting copula regression. For the deselection approach, we specified the threshold parameter τ with 0.1 and 1%. For stability selection, the number of variables to be included in the model was set as $q = 20$ and the per-family error rate was chosen to be $\text{PFER} = 5$. We employed $B = 50$ complementary pairs for subsampling. The m_{stop} for the boosting model for stability selection was set to five times the number of observations ($5 \cdot n$) ensuring that q base-learners can be selected. For the final model with only the stable covariates, the optimal stopping iteration was determined using an additional dataset, as in the classical boosted model. Note that due to the high computational cost, stability selection could not be applied for the high-dimensional settings.

To evaluate the prediction performance we used multivariate proper scoring rules, namely, the negative log-likelihood and the energy score. The energy score generalizes the continuous ranked probability score to multivariate quantities [7] and is defined as follows. Let $\mathbf{y} = (y_1, y_2)^T \in \mathbb{R}^2$ represent the vector of observations and let \hat{F} denote a forecast distribution on \mathbb{R}^2 . Assume $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are n independent realizations from \hat{F} , where each realization is given by $\mathbf{Y}_i = (Y_{i1}, Y_{i2}) \in \mathbb{R}^2$ for $i = 1, \dots, n$. The energy score is defined as

$$\text{ES}(F, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n \|\mathbf{Y}_i - \mathbf{y}\| - \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{Y}_i - \mathbf{Y}_j\|,$$

where $\|\cdot\|$ denotes the Euclidean norm on \mathbb{R}^2 .

3.2 Summary of Simulation Results

In Scenario A, B.1, and B.2, classical boosting is able to correctly select the informative variables for each distribution parameter, while noninformative variables were included mainly for the mean parameters (supplement, Section B.1, and B.2). However, in Scenario C, not all informative variables were selected for the dependence parameter and many noninformative variables were included for the mean and scale parameters (see the supplement, Section B.3).

In comparison, probing, stability selection, and deselection led to much sparser models in both low- and high-dimensional cases. Specifically, in Scenario A and Scenario B.1, the final models generally contained all informative variables except when using probing, which occasionally missed the informative variable for the dependence parameter ρ . Stability selection also occasionally missed the informative variable with a Clayton copula. With the deselection approach, all informative variables remained in the model. The fewest false

positives were obtained when τ was set to 1%, almost completely eliminating them. With a threshold value of 0.1% also many noninformative variables were excluded but to a lower extent than with 1% (see the supplement, Section B.1 and B.2). Similarly, in Scenario B.2, all approaches perform well in terms of true positive selection, although again probing failed to include all informative variables in some cases for σ_1 . As before, none of the approaches can eliminate all false positives and excluding false positives for the association parameter is particularly challenging. However, this depends on the strength of the association between the outcomes. With stronger association, it is more difficult to eliminate the noninformative variables. With weaker association, the classical boosting tends to include few false positives for ρ , particularly for high-dimensional scenarios (see the additional setting for Scenario B.2 in supplement Section B.2 for more details).

In Scenario C, only for the Gaussian copula all informative variables were selected by classical boosting; for the other copulas, it was already difficult to select all true positives for the dependence parameter. Most false positives were included for the mean and scale parameters. The resulting models for probing and stability selection for Scenario C had difficulties in selecting the informative variables. The average number of true positives is relatively low for both approaches. The deselection approach with a threshold value of 0.1% only slightly influenced the average number of true positives. For all other parameters, the informative variables remained in the model in every simulation run. The number of noninformative variables were considerably reduced but there are still false positives left in the model. A higher threshold value would lead to a higher decrease in false positives but also to a reduction of correctly identified informative variables (see the supplement, Section B.3).

For the predictive performance on test data, evaluated with the negative log-likelihood and the energy score (smaller values are better), the deselection approach as well as stability selection had a comparable predictive performance and led to an improvement in the negative log-likelihood compared to the classical boosting for Scenario A, B.1 and B.2. For the energy score, the approaches resulted in similar values. Only probing showed a worse performance compared to the classical boosted model for the negative log-likelihood and the energy score (see the supplement, Section B.1 and B.2). For Scenario C, probing and stability selection led to a worse predictive performance due to the exclusion of informative variables. The deselection approach yielded an improvement in the negative log-likelihood for a threshold value of 0.1% and provided comparable performance to the classical approach regarding the energy score (see the supplement, Section B.3).

Overall, all approaches can drastically reduce the number of false positives in the final boosting model, whereby probing yielded the smallest runtime, as there is no need for an additional optimization of the stopping iteration. Due to its second boosting step, the deselection approach took slightly longer than the classic approach (≈ 1 – 2 min). Stability selection had the longest runtime because B boosting models have to be fitted on the subsamples.

3.3 Characteristics of Enhanced Variable Selection Approaches

Based on our simulation study, we have summarized the key aspects of the different variable selection approaches to guide researchers in choosing a suitable method for specific data challenges. Table 1 provides an overview of probing, stability selection and deselection approaches, evaluating each method across several important characteristics, including the number of parameters to be specified, computational cost, and ease of use. Note that not every category can be considered by itself.

Probing emerges as the simplest method in terms of parameter specification and computational efficiency. It requires no specification of additional parameters and offers a low computational burden. However, this simplicity leads to limitations in variable selection, coefficient estimation and prediction performance. As discussed in Sect. 2.2.1, probing is intended to yield sparse models rather than optimized predictive performance. In addition, while probing is generally computationally efficient regarding runtime, it can cause memory problems when applied to large or high-dimensional data. This is because the method generates the additional probe variables, effectively doubling the dimensionality of the data set. As a result, even larger matrices must be stored and processed. This makes probing less practical for extremely high-dimensional scenarios, despite its otherwise efficient nature. Stability selection, on the other hand, provides effective variable selection and, after refitting with the stable covariates, provides reasonable coefficient estimation and comparable predictive performance to the classical boosting model. It offers error control, which is an advantage in many applications. However, these advantages come at the cost of increased complexity. Stability selection requires the specification of multiple parameters and is computationally intensive due to the repeated sub-sampling, making it more suitable for low-dimensional data sets. Deselection provides effective variable selection and coefficient estimation and achieves predictive performance comparable to the classical boosting model. It is relatively easy to use, similar to probing, and shows particular strengths when handling many potential variables, i.e., it is better scalable for large or high-dimensional data than probing

Table 1 Comparison of enhanced variable selection approaches: ✓ indicates an advantage, ○ represents a moderate or neutral characteristic and ✕ signifies a limitation or disadvantage

Characteristic	Probing	Stability selection	Deselection
Number of parameters	✓	✕	○
Computational cost	✓	✕	○
True positive selection	○	✓	✓
False positive reduction	✓	✓	✓
Coefficient estimates	○	✓	✓
Predictive performance	✕	✓	✓
Per-family error rate control	✕	✓	✕
High number of potential variables	○	✕	✓
True model not sparse	✕	✕	○
Simple to use	✓	○	✓

and stability selection. While deselection requires the specification of one parameter, this parameter is generally more intuitive to select. In general, $\tau = 0.01$ serves as a reasonable default for many scenarios. This combination of characteristics makes deselection a practical option for a variety of data problems.

In general, the selection of an appropriate variable selection method should be guided by a careful consideration of the specific requirements, the characteristics of the dataset and the available computational resources. As expected, all three methods perform better when the true underlying model is sparse.

4 Real Data Illustrations

4.1 Analysis of Fetal Ultrasound Data

Motivated by the analysis of fetal ultrasound data using boosted copula regression of [9], which resulted in rather large sub-models for the different distribution parameters, we examined and compared the variable selection and the predictive performance of this analysis with the models resulting from the enhanced variable selection techniques introduced in Sect. 2.2. The considered dataset was collected from 2006 to 2016 at the Department of Obstetrics and Gynecology of the Erlangen University Hospital and contains 6103 observations and 36 variables, including sonographic variables, e.g., abdominal anteroposterior diameter, abdominal transverse diameter, the interaction between these sonographic variables, and clinical variables, e.g., weight, height and body-mass index (BMI) of the mother. For more details on the data, we refer to [5].

The response variables of interest are the birth length and weight, which were modeled via copula regression with log-logistic marginal distributions and the Gaussian copula. We split the dataset into a training dataset with $n = 4,103$ observations and a test dataset for evaluation with 2000 observations. The step length was set to $\nu = 0.01$ and the stopping iteration was optimized by 10-fold cross-validation. All variables were considered for each distribution parameter. For continuous variables, cubic P-splines with 20 equidistant knots, a second-order difference penalty and 4 degrees of freedom were used as base-learners. Sex of the fetus and gestational diabetes were included via linear base-learners. Furthermore, we applied a gradient stabilization to ensure comparable gradients for the distribution parameters [12]. For the deselection approach, threshold values of 0.1 and 1% were considered. The parameters for stability selection were specified as $q = 20$ for the number of variables to be included in the model and $\text{PFER} = 5$ for the per-family error rate.

Table 2 shows the numbers of selected variables for each distribution parameter, the predictive performance in terms of the negative log-likelihood as well as the resulting optimal m_{stop} . An overview of the included variables for the different approaches can be found in the supplement, Section D. The classical boosted copula model selected almost all considered variables for the mean parameters μ_1 and μ_2 . Fewer variables were selected for the shape parameters and the dependence parameter. The approaches for enhanced variable selection reduced the model complexity substantially and led to fewer included variables in the final models. The deselection

Table 2 Numbers of selected variables for distribution parameters μ_1 , σ_1 , μ_2 , σ_2 and ρ , negative log-likelihood values and stopping iteration m_{stop} for classical boosting, deselection with threshold values of 0.1 and 1%, probing and stability selection

Method	μ_1	σ_1	μ_2	σ_2	ρ	−Log-Lik	m_{stop}
Classic	33	15	30	13	9	4156.58	5536
Deselection 0.1%	6	9	8	7	4	4184.07	5536
Deselection 1%	–	5	2	4	1	4843.41	5536
Probing	9	9	8	9	5	4654.58	808
Stability selection	5	3	3	3	–	4273.18	4194

approach with a threshold value of 1% deselected all variables for the mean parameter μ_1 . Still, it contained variables for the other distribution parameters, more precisely interactions of the sonographic variables and the gestational age for the scale parameters. As expected, it resulted in a slightly worse negative log-likelihood compared to the classical approach.

With a smaller threshold value (0.1%), the final model contained variables for each distribution parameter and led to a comparable predictive performance than the classical boosted model. Here the model included mostly interactions of the sonographic variables as well but also a few other variables, e.g., sex for the location and gestational age for each distribution parameter except the dependence parameter. Probing resulted in a similar model as the deselection approach with 0.1%, but led to worse predictive performance. The most likely reason is the stronger shrinkage of the effect estimates due to the much smaller number of iterations. Via stability selection no covariates were selected as stable for the dependence parameter implying conditional independence of the two responses. This resulted in a poorer predictive performance compared to the classical model.

4.2 Joint Modeling of Cholesterol Phenotypes

We analyzed data from the UK Biobank (application number 81202), which is a large biomedical cohort study containing genetic and health information from over half a million British participants [35]. Using the boosting algorithm for distribution copula regression, we aim to model the polygenic contribution to the individual distributions of different phenotypes, but also to estimate the dependence between these phenotypes as a function of genetic variants. We want to identify the most relevant variants and therefore apply the methods presented in Sect. 2.2 to obtain sparse solutions.

The focus in the following is on three bivariate combinations of phenotypes, namely LDL (*Low-Density Lipoprotein*) and ApoB (*Apolipoprotein B*), LDL and cholesterol, and HDL (*High-Density Lipoprotein*) and ApoA (*Apolipoprotein A*). We considered these combinations because they have high empirical association based on an analysis of genetic blood and urine biomarkers in the UK Biobank [30], suggesting potential benefits in modeling these phenotypes jointly. All of these phenotypes are components of cholesterol metabolism. Cholesterol can be split mainly into two groups: i) LDL cholesterol, which is responsible for the transportation of cholesterol from the liver to various tissues and can be attached to specific receptors

on the cell surface with the help of ApoB, ii) HDL cholesterol, the counterpart of LDL which is accountable for the removal of excess LDL cholesterol from the body, with ApoA supporting this process [41].

The considered dataset for each combination of phenotypes consists of $n = 20,000$ randomly sampled observations with white British ancestry. Additionally, 15,000 observations were used for validation and 20,000 observations were used to evaluate the prediction performance via the negative log-likelihood. For each phenotype, 1000 variants were selected in a pre-screening step based on the largest marginal associations between the variants and the phenotype, which were computed with the PLINK2 function `-variant-score` [1, 26]. Variants with minor allele frequency not less than 1% were randomly sampled with the `-thin-count` function. Missing genotypes were imputed by the reference allele using the R package **bigsnpr** [25]. After the pre-screening, the dataset contains 1156 variants for LDL and ApoB (844 variants were selected for both phenotypes), 1179 variants for LDL and cholesterol (821 common variants), and 1249 variants for HDL and ApoA (751 common variants).

For each combination of two phenotypes, the marginal distributions and copulas were chosen which minimize the predictive risk (see Table 3). All variants were considered for each distribution parameter and incorporated with linear base-learners and step length $\nu = 0.01$. Stability selection unfortunately could not be applied to these data because of the high computational cost.

Table 3 shows the results for the joint analysis of the different combinations of phenotypes. Furthermore, Fig. 2 displays Manhattan-type plots for the phenotype combination LDL and cholesterol for every distribution parameter of the copula model. For each combination of phenotypes, the classical boosting approach selected several variants for each distribution parameter. Most genetic variants were selected for the location parameters. Each model included variants for the dependence

Table 3 Number of selected variants for distribution parameters μ_1 , σ_1 , μ_2 , σ_2 and ρ , negative log-likelihood values and stopping iteration m_{stop} for the classical boosted model, the deselection approach with threshold values of 0.1 and 1%, and probing for the different combinations of phenotypes

Phenotype	Marginals	Copula	Method	μ_1	σ_1	μ_2	σ_2	ρ	−Log-Lik	m_{stop}
LDL	Log-logistic	Gaussian	Classic	441	26	386	67	47	10535.16	4965
ApoB	Gamma		Deselection 0.1%	121	2	71	9	15	10749.28	4965
			Deselection 1%	8	–	3	–	–	11647.92	4965
			Probing	–	–	–	–	–	14792.60	863
LDL	Log-logistic	Gumbel	Classic	286	89	266	100	44	31105.03	13,975
Cholesterol	Log-logistic		Deselection 0.1%	81	5	54	7	9	31090.53	13,975
			Deselection 1%	12	–	6	–	–	31817.39	13,975
			Probing	45	–	15	3	–	32834.92	5314
HDL	Log-normal	Gaussian	Classic	171	40	197	69	28	79820.43	1954
ApoA	Log-normal		Deselection 0.1%	81	15	83	24	9	79868.42	1954
			Deselection 1%	8	–	9	–	–	80337.71	1954
			Probing	113	20	185	36	6	80074.35	905

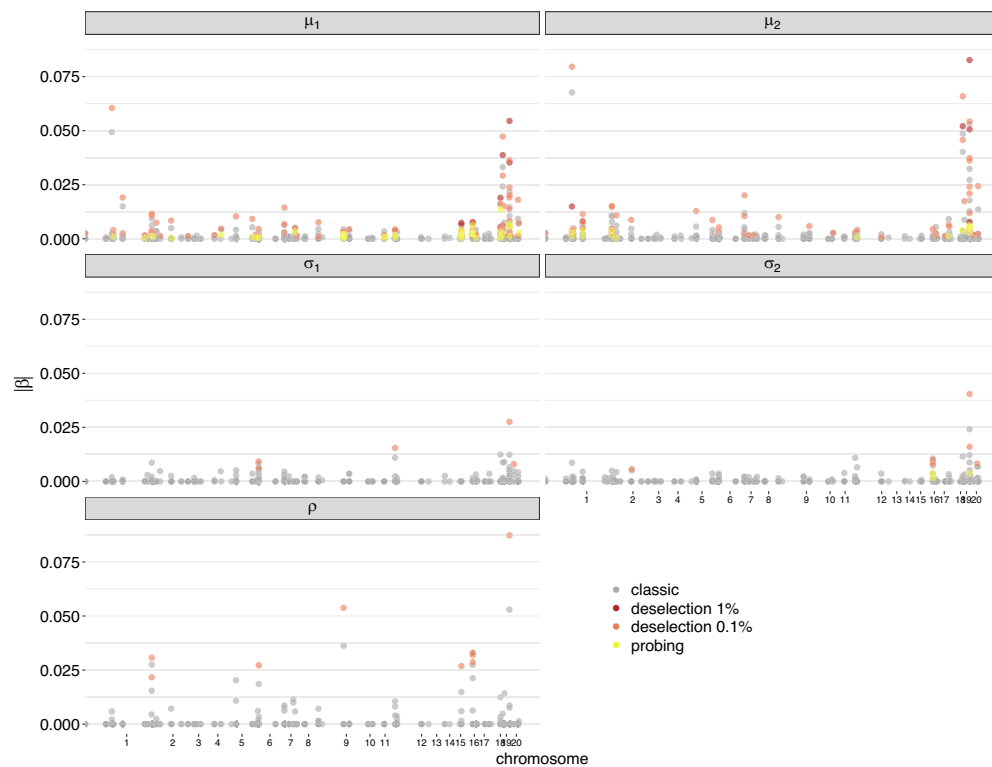


Fig. 2 Manhattan-type plots (chromosomes on x-axis) for the absolute coefficients of boosted copula regression for the joint analysis of LDL and cholesterol

parameter, indicating that different variants affect the associations between phenotypes and the potential benefit of modeling these phenotypes together. Considering the total number of selected variants, a relatively high number of the pre-filtered variants were included in the classical boosting model. In particular, for LDL and ApoB, almost half of all variants were included for the mean parameters. Despite the intrinsic variable selection of the boosting algorithm, we still obtain large models with a potentially difficult interpretation. Therefore, we aim to reduce the model complexity by enhancing variable selection.

With the deselection approach, the model complexity could be drastically reduced. When considering a threshold value of 1%, for all phenotype combinations only variants for the location parameters remained after deselection, which resulted in two univariate models. One can argue that this threshold value may be too strong for the data situation as there are several variants with only a small-to-medium effect (see for example Fig. 2) and therefore a minor impact on risk reduction. Also, owing to the pre-filtering, all variants in our analysis have some association with one of the outcomes, making it harder for single variants to pass the relative threshold. Using a smaller threshold value (0.1%) also led to sparser models, but for each distribution parameter several variants remained in the final model. The negative log-likelihood indicated a comparable predictive performance, whereby even a slight improvement in the performance for the phenotype combination LDL and cholesterol could be observed.

Probing also resulted in sparser models for each combination. In fact, for the phenotypes LDL and ApoB, no variants were included in the resulting model: in the first iterations, only the intercept was updated and stopping after 863 boosting iterations (when the first probe was selected) resulted in an intercept model, which led to a considerably worse predictive performance. For the other phenotypes, several variants were included after stopping when the first probe was selected. However, due to the smaller number of boosting iterations, the effect estimates were more shrunk (see Fig. 2 for LDL and cholesterol) and therefore the predictive performance deteriorated in comparison to the classical boosted model but also to the deselection approach, particularly for a threshold value of 0.1%.

5 Discussion and Conclusion

To reduce model complexity and to enhance variable selection for boosting multivariate distributional copula regression, we have integrated probing [36], stability selection [22], and also the recent deselection approach [33] in the boosting framework for this model class. This combination of classical boosting with all three approaches leads to considerably sparser models, thereby improving the interpretability of the obtained prediction models, which is desirable in practice [17, 42].

Regarding the specific approaches, the results of stability selection show similarities to the ones from deselection, even though the initial goals of the two methods differ. All three approaches perform better when the true model is sparse, whereby deselection can still lead to reasonable results when many variables are informative. The probing approach is the most favorable regarding computational runtime, but typically stops the algorithm also very early, leading often to underfitting and reduced predictive performance. As also observed in our first application on the weight and length of newborns, stability selection and deselection are more often able to maintain the predictive performance with smaller models. However, only deselection is also scalable to large, high-dimensional data as in our genetic application.

Our results additionally suggest that deselection not only yields much sparser models but can even lead to simpler univariate regression models in comparison to the classical boosted copula model in situations where the association parameter is close to zero. The proposed methods for enhanced variable selection could hence also represent tools for data-driven model choice [21]. The prediction performance typically does not improve after deselection but can lead to comparable accuracy as the classical boosting model with fewer predictors. Further improvements could be achieved in the future by optimizing the stopping iteration of the final boosting model, potentially leading to reduced shrinkage and slightly higher predictive performance. Stability selection works similarly, providing stable covariates that are then re-fitted in a final model with an optimized tuning parameter. The same principle could be applied to the deselection approach, which would, however, increase again the computational burden. In addition, also probing could be considered only as an extended method of variable selection followed by refitting the model only using the selected base-learners and tuning the stopping iteration. These

methodological extensions are beyond the scope of our current comparison, but offer promising directions for future investigation.

The deselection procedure is controlled via a threshold value τ , which represents the minimum amount of total risk reduction that should be attributed to a corresponding base-learner to avoid deselection. This can be interpreted as a threshold value for the importance of the particular predictor variable. Depending on the data situation, different thresholds may be appropriate; however, tuning is not straightforward because the true number of informative variables is not known in practice and the best model regarding predictive risk is naturally the one without any deselection. Further research is warranted on how to specify the threshold τ in this context.

Besides the practical advantages of the proposed tools, the natural limitation of all boosting algorithms applies: Due to early stopping and therefore shrinkage of the effect estimates, providing standard errors of the resulting coefficients is not an easy task as there are no closed formulas. To overcome this, one could apply permutation tests to carry out significance testing and provide p -values [20], but this would drastically increase the computational cost.

In conclusion, while statistical models should be as complex as needed to be able to capture the underlying nature of the data-generating process, they should also remain as simple as possible to facilitate interpretation [10]. To navigate this conceptual trade-off, we have proposed three competing approaches to simplify distributional copula regression models by reducing the model complexity and to enhance the variable selection properties of statistical boosting without considerably reducing the prediction accuracy of the resulting models.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12561-025-09491-8>.

Funding Open Access funding enabled and organized by Projekt DEAL. The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776, KL3037/2-1, MA7304/1-1).

Data Availability The code used for the simulations and the biomedical applications is available at GitHub <https://github.com/AnnikaStr/ComplRedBoostCop>. The fetal ultrasound data are not publicly available. However, to facilitate reproducibility, an artificial dataset that mimics the characteristics of the original data is also available in the GitHub repository. The genomic cohort data are available upon request from the UK Biobank at <https://www.ukbiobank.ac.uk/>.

Declarations

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM et al (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4(1):7. <https://doi.org/10.1186/s13742-015-0047-8>
2. Dikheel TR, Alwa SH (2022) Using cross-validation, probing, and lasso in gradient boosting variable selection. *Nat Volatiles Essent Oils* 9:13620–13630. <https://doi.org/10.53555/nveo.v9i1.5583>
3. Fahrmeir L, Kneib T, Lang S (2004) Penalized structured additive regression for space-time data: a Bayesian perspective. *Stat Sin* 14(3):731–761
4. Fahrmeir L, Kneib T, Lang S, Marx B (2013) *Regression: models, methods and applications*, 1st edn. Springer, Berlin
5. Faschingbauer F, Dammer U, Raabe E, Kehl S, Schmid M et al (2016) A new sonographic weight estimation formula for small-for-gestational-age fetuses. *J Med Ultrasound* 35(8):1713–1724. <https://doi.org/10.7863/ultra.15.09084>
6. Freund Y (1995) Boosting a weak learning algorithm by majority. *Inf Comput* 12(2):256–285. <https://doi.org/10.1006/inco.1995.1136>
7. Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA (2008) Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST* 17:211–235. <https://doi.org/10.1007/s11749-008-0114-x>
8. Griesbach C, Mayr A, Bergherr E (2023) Variable selection and allocation in joint models via gradient boosting techniques. *Mathematics* 11(2):411. <https://doi.org/10.3390/math11020411>
9. Hans N, Klein N, Faschingbauer F, Schneider M, Mayr A (2022) Boosting distributional copula regression. *Biometrics* 79(3):2298–2310. <https://doi.org/10.1111/biom.13765>
10. Heller GZ (2024) Simple or complex statistical models: non-traditional regression models with intuitive interpretations. *Stat Model* 24(6):503–519. <https://doi.org/10.1177/1471082X241274405>
11. Hofner B, Boccuto L, Göker M (2015) Controlling false discoveries in high-dimensional situations: boosting with stability selection. *BMC Bioinformatics* 16(1):144. <https://doi.org/10.1186/s12859-015-0575-3>
12. Hofner B, Mayr A, Robinzonov N, Schmid M (2014) 02. Model-based boosting in R: A hands-on tutorial using the R package mboost. *Comput Stat* 29:3–35. <https://doi.org/10.1007/s00180-012-0382-5>
13. Keil AP, O'Brien KM (2023) Considerations and targeted approaches to identifying bad actors in exposure mixtures. *Stat Biosci* 16(2):459–481. <https://doi.org/10.1007/s12561-023-09409-2>
14. Klein N (2023) Distributional regression for data analysis. *Ann Rev Stat Appl*. <https://doi.org/10.1146/annurev-statistics-040722-053607>
15. Klein N, Kneib T, Klasen S, Lang S (2015) Bayesian structured additive distributional regression for multivariate responses. *J R Stat Soc C Appl* 64(4):569–591. <https://doi.org/10.1111/rssc.12090>
16. Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A (2023) A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Front Genet* 13:1076440. <https://doi.org/10.3389/fgene.2022.1076440>
17. Markowitz F (2024) All models are wrong and yours are useless: making clinical prediction models impactful for patients. *NPJ Precision Oncol* 8:54. <https://doi.org/10.1038/s41698-024-00553-6>
18. Mayr A, Binder H, Gefeller O, Schmid M (2014) The evolution of boosting algorithms - from machine learning to statistical modelling. *Methods Inf Med* 53(06):419–427. <https://doi.org/10.3414/ME13-01-0122>
19. Mayr A, Hofner B, Schmid M (2012) The importance of knowing when to stop. A sequential stopping rule for component-wise gradient boosting. *Methods Inf Med* 51(2):178–186
20. Mayr A, Schmid M, Pfahlberg A, Uter W, Gefeller O (2017) A permutation test to analyse systematic bias and random measurement errors of medical devices via boosting location and scale models. *Stat Methods Med Res* 26(3):1443–1460. <https://doi.org/10.1177/0962280215581855>
21. Mayr A, Wistuba T, Speller J, Gude F, Hofner B (2023) Linear or smooth? enhanced model choice in boosting via deselection of base-learners. *Stat Model* 23(5–6):441–455. <https://doi.org/10.1177/1471082X231170045>
22. Meinshausen N, Bühlmann P (2010) Stability selection. *J R Stat Soc B Stat Methodol* 72(4):417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
23. Nelsen RB (2006) *An introduction to copulas*. Springer, New York

24. Nguyen PH, Herring AH, Engel SM (2023) Power analysis of exposure mixture studies via Monte Carlo simulations. *Stat Biosci* 16(2):321–346. <https://doi.org/10.1007/s12561-023-09385-7>
25. Privé F, Aschard H, Ziyatdinov A, Blum MGB (2018) Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* 34(16):2781–2787. <https://doi.org/10.1093/bioinformatics/bty185>
26. Purcell S, Chang C (2015) Plink 2.0. www.cog-genomics.org/plink/2.0/
27. Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. *J R Stat Soc C Appl* 54(3):507–554. <https://doi.org/10.1111/j.1467-9876.2005.00510.x>
28. Schmid M, Hothorn T (2008) Boosting additive models using component-wise P-splines. *Comput Stat Data Anal* 53(2):298–311. <https://doi.org/10.1016/j.csda.2008.09.009>
29. Shah RD, Samworth RJ (2013) Variable selection with error control: another look at stability selection. *J R Stat Soc Ser B Methodol* 75(1):55–80. <https://doi.org/10.1111/j.1467-9868.2011.01034.x>
30. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C et al (2021) Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* 53(2):185–194. <https://doi.org/10.1038/s41588-020-00757-z>
31. Staerk C, Mayr A (2021) Randomized boosting with multivariable base-learners for high-dimensional variable selection and prediction. *BMC Bioinformatics* 22:1–28. <https://doi.org/10.1186/s12859-021-04340-z>
32. Stasinopoulos DM, Rigby RA, Heller GZ, De Bastiani F (2023) P-splines and GAMLSS: a powerful combination, with an application to zero-adjusted distributions. *Stat Model* 23(5–6):510–524. <https://doi.org/10.1177/1471082X231176635>
33. Strömer A, Staerk C, Klein N, Weinhold L, Titze S, Mayr A (2022) Deselection of base-learners for statistical boosting - with an application to distributional regression. *Stat Methods Med Res* 31(2):207–224. <https://doi.org/10.1177/09622802211051088>
34. Strömer A, Klein N, Staerk C, Klinkhammer H, Mayr A (2023) Boosting multivariate structured additive distributional regression models. *Stat Med* 42(11):1779–1801. <https://doi.org/10.1002/sim.9699>
35. Sudlow C, Gallacher J, Allen N, Beral V, Burton P et al (2015) UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 12(3):e1001779. <https://doi.org/10.1371/journal.pmed.1001779>
36. Thomas J, Hepp T, Mayr A, Bischl B (2017) Probing for sparse and fast variable selection with model-based boosting. *Comput Math Methods Med* 2017:1421409. <https://doi.org/10.1155/2017/1421409>
37. Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B (2018) 05. Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Stat Comput* 28:673–687. <https://doi.org/10.1007/s11222-017-9754-6>
38. Tian T, Sun J (2024) Variable selection for nonlinear covariate effects with interval-censored failure time data. *Stat Biosci* 16(1):185–202. <https://doi.org/10.1007/s12561-023-09391-9>
39. Umlauf N, Klein N, Zeileis A (2018) Bamlss: Bayesian additive models for location, scale, and shape (and beyond). *J Comput Graph Stat* 27(3):612–627. <https://doi.org/10.1080/10618600.2017.1407325>
40. Verhasselt A, Flórez AJ, Molenberghs G, Van Keilegom I (2024) Copula-based pairwise estimator for quantile regression with hierarchical missing data. *Stat Model* 25(2):129–149. <https://doi.org/10.1177/1471082x231225806>
41. Walldius G, Jungner I (2004) Apolipoprotein B and apolipoprotein A-I: Risk indicators of coronary heart disease and targets for lipid-modifying therapy. *J Intern Med* 255(2):188–205. <https://doi.org/10.1046/j.1365-2796.2003.01276.x>
42. Wyatt JC, Altman DG (1995) Prognostic models: clinically useful or quickly forgotten? *Br Med J* 311(7019):1539–1541. <https://doi.org/10.1136/bmj.311.7019.1539>
43. Yang L, Czado C (2022) Two-part d-vine copula models for longitudinal insurance claim data. *Scand J Stat* 49(4):1534–1561. <https://doi.org/10.1111/sjos.12566>

Authors and Affiliations

**Annika Strömer^{1,2}  · Nadja Klein³ · Christian Staerk^{4,5} ·
Florian Faschingbauer⁶ · Hannah Klinkhammer^{2,7} · Andreas Mayr¹ **

✉ Annika Strömer
annika.stroemer@uni-marburg.de

Nadja Klein
nadja.klein@kit.edu

Christian Staerk
staerk@statistik.tu-dortmund.de

Florian Faschingbauer
Florian.Faschingbauer@uk-erlangen.de

Hannah Klinkhammer
klinkhammer@imbie.uni-bonn.de

Andreas Mayr
andreas.mayr@uni-marburg.de

¹ Institute for Medical Biometry and Statistics, University of Marburg, Marburg, Germany

² Department of Medical Biometrics, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany

³ Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany

⁴ IUF - Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany

⁵ Department of Statistics, TU Dortmund University, Dortmund, Germany

⁶ Department of Obstetrics and Gynecology, University Hospital of Erlangen, Erlangen, Germany

⁷ Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn, Germany

4 Discussion with references

This cumulative dissertation introduces novel modeling approaches designed to address the multiple challenges presented by modern biomedical data. Furthermore, a central focus is on how to balance model complexity with interpretability, ensuring that sophisticated statistical methods remain accessible and meaningful for practical applications.

Advanced statistical models are crucial for identifying subtle relationships between multiple outcomes, as demonstrated by *Publication A* (Hilbert et al., 2025), which revealed distinct patterns of change over time in physical and mental health, demonstrating that various outcomes do not constantly evolve in parallel and that sustainability can differ among patient subgroups. The results indicate that considering multiple health indicators provided better outcome predictions than weight alone. Therefore, this emphasizes the importance of modeling approaches that capture the association among multiple outcomes and the identification of risk factors that influence the strength of these dependencies in order to understand disease mechanisms and inform targeted interventions.

While classical maximum likelihood or Bayesian approaches (Klein et al., 2015) for multivariate distributional regression can not handle high-dimensional data problems, *Publication B* (Strömer et al., 2023) addresses this by presenting a boosting algorithm for multivariate responses (i.e., for binary, count and continuous outcomes) by combining the properties of GAMLSS and the main features of statistical boosting. As a special merit, the boosting framework can be used directly for high-dimensional data, allowing for data-driven variable selection mechanisms that enable sparse models for all parameters of a multivariate distribution. *Publication B* describes the algorithm in detail and illustrates its potential to consider multiple outcomes simultaneously and hence allows accessing, for example, the genetic predisposition for the association between several phenotypes, such as heart disease and high cholesterol. To the best of our knowledge, this is the first time multivariate distributional regression has been adapted to model the joint genetic liability for multiple phenotypes. Building on this, recent work has continued to address the underexplored area of multivariate distributional regression models (Kneib et al., 2023). For example, Gioia et al. (2025) extended these methods for

Gaussian distributions with more than two response dimensions. In contrast, Kock and Klein (2025) proposes a more flexible multivariate distributional model within the GAMLSS framework, enabling the analyses of more than two outcomes of varying types. However, the latter approach does not yet support high-dimensional data. Gioia et al. (2025) addresses high-dimensional parameter and covariate spaces by employing a semi-automatic model selection strategy motivated by *Publication D* (Strömer et al., 2022).

Another key focus in biomedical research is the analysis of time-to-event data, where a major challenge arises from dependent censoring, for instance, when patients with poorer health are more likely to drop out early. While existing copula-based and joint frailty-copula models have improved the handling of this issue, they typically rely on strong assumptions and therefore lack the flexibility to model the dependence structure or accommodate high-dimensional predictors (Emura and Chen, 2018; Emura et al., 2019). *Publication C* (Strömer et al., 2025b) addresses these limitations by introducing a boosting distributional copula regression approach for dependent censoring that explicitly models the dependence between event and censoring times as a function of covariates. Therefore, this approach provides additional insights into the dependence between survival and censoring times not captured by previous models (Midtjord et al., 2022; Czado and Van Keilegom, 2023; Deresa et al., 2022). While the method is currently based on parametric margins and copulas, future extensions to semi-parametric margins, as in Deresa and Van Keilegom (2024) or transformation models (Deresa and Van Keilegom, 2025), would increase flexibility. Additionally, integrating other censoring mechanisms into the current framework, such as left censored or truncated data, would further enhance its applicability.

Flexible modeling, combined with advanced statistical techniques, has expanded our ability to analyze complex biomedical data and uncover underlying mechanisms. However, as these models become increasingly complex and incorporate a large number of potential covariates, interpretability and variable selection become increasingly challenging. Therefore, one of the main issues is the effective selection of variables, particularly in high-dimensional settings. While earlier stopping strategies, such as probing (Thomas et al., 2017) or the one standard

error rule (Hastie et al., 2009; Friedman et al., 2010), are easy to implement and encourage sparsity, they lead to suboptimal variable selection and poorer predictive accuracy. Stability selection (Hofner et al., 2015; Meinshausen and Bühlmann, 2010) offers more reliable covariate selection in low-dimensional settings, but its computational demands limit its practicality for large-scale biomedical data.

In this context, *Publication D* (Strömer et al., 2022) introduces a deselection approach, which systematically refines variable selection in boosting algorithms by leveraging the attributable risk reduction of individual base-learners. By removing variables that contribute minimally to overall model performance, deselection produces much sparser models without sacrificing predictive power. Importantly, because this method is based on risk reduction, it is broadly applicable across a range of base-learners, such as linear models, splines, and spatial effects, and can be adapted to various regression settings. This approach was first applied to boosting GAMs and GAMLSS with different outcome and effect types, where it proved to be both reliable and effective. Building on this foundation, *Publication E* (Strömer et al., 2025a) extends the deselection framework to multivariate distributional copula regression. Here, deselection not only reduces dimensionality but also supports model simplification by identifying which components truly benefit from covariate effects. This targeted complexity control is particularly relevant when modeling multiple correlated outcomes, where overparameterization can easily obscure scientific insights. A notable strength of the proposed deselection framework is its versatility, which has already been demonstrated in various contexts, for example, in applications to vine copula models (Jobst et al., 2024). Beyond this, recent modifications have further enhanced its applicability. For instance, the deselection approach has been used directly as a pre-filtering strategy to efficiently reduce the number of covariates before model fitting, allowing the model fitting process to remain computationally feasible in high-dimensional settings where standard approaches would fail (Gioia et al., 2025). Further extension of the framework enables data-driven model choice between linear and smooth spline effects for continuous covariates (Mayr et al., 2023).

Despite the advantages of newly developed statistical methods and their ability to handle high-

dimensional data, certain limitations remain. The approaches introduced in this work maintain interpretability while offering flexibility in modeling complex data structures. However, challenges arise when moving toward ultra-high-dimensional settings – such as genomic datasets with hundreds of thousands of predictors – where model estimation becomes computationally infeasible without substantial resources or specialized batch-wise techniques (Klinkhammer et al., 2023; Wetscher et al., 2024). Finally, while the proposed deselection framework effectively reduces model complexity and enhances sparsity, pushing these methods toward even more complex modeling frameworks may come at the expense of interpretability. While the results of the models presented here are still interpretable, the increasing complexity of the models makes it more difficult to provide a meaningful explanation of individual model components. As Heller (2024) emphasizes, statistical models should be as complex as necessary to capture the underlying data-generating process yet remain as simple as possible to ensure meaningful interpretation.

4.1 Conclusion

In summary, this cumulative dissertation develops flexible, interpretable and scalable statistical frameworks to address the complexities of modern biomedical data. By advancing methods for multivariate distributional regression, survival analysis under dependent censoring and high-dimensional variable selection, this work provides robust tools for uncovering complex relationships and dependencies in health research. Furthermore, the deselection framework offers a versatile and practical approach to model simplification and variable selection in various challenging settings.

These contributions provide a solid foundation for future research. Moving forward, it will be essential to extend these frameworks to accommodate even more complex data structures and further enhance dependency modeling (e.g., asymmetric dependence or modeling associations among three or more outcomes) while carefully balancing model complexity, interpretability and computational efficiency.

4.2 References

- Czado C, Van Keilegom I. Dependent censoring based on parametric copulas. *Biometrika*, 2023; 110 (3): 721–738
- Deresa N, Van Keilegom I, Antonio K. Copula-based inference for bivariate survival data with left truncation and dependent censoring. *Insurance: Mathematics and Economics*, 2022; 107: 1–21
- Deresa NW, Van Keilegom I. Copula Based Cox Proportional Hazards Models for Dependent Censoring. *Journal of the American Statistical Association*, 2024; 119 (546): 1044–1054
- Deresa NW, Van Keilegom I. Semiparametric transformation models for survival data with dependent censoring. *Annals of the Institute of Statistical Mathematics*, 2025; 77 (3): 425–457
- Emura T, Chen YH. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer, 2018
- Emura T, Matsui S, Rondeau V. *Survival analysis with correlated endpoints: Joint Frailty-Copula models*. Springer, 2019
- Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 2010; 33 (1): 1–22
- Gioia V, Fasiolo M, Browell J, Bellio R. Additive Covariance Matrix Models: Modeling Regional Electricity Net-Demand in Great Britain. *Journal of the American Statistical Association*, 2025; 120 (549): 107–119
- Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009
- Heller GZ. Simple or complex statistical models: Non-traditional regression models with intuitive interpretations. *Statistical Modelling*, 2024; 24 (6): 503–519
- Hilbert A, Strömer A, Staerk C, Schreglmann B, Mansfeld T, Sander J, Seyfried F, Kaiser S, Stroh C, Dietrich A, Schmidt R, Mayr A. Multivariate Trajectories of Weight and Mental Health and Their Prognostic Significance 6-Years After Obesity Surgery. *International Journal of Eating Disorders*, 2025; 58 (11): 2214–2226

- Hofner B, Boccuto L, Göker M. Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, 2015; 16 (1): 144
- Jobst D, Möller A, Groß J. Gradient-Boosted Generalized Linear Models for Conditional Vine Copulas. *Environmetrics*, 2024; 35: e2887
- Klein N, Kneib T, Klasen S, Lang S. Bayesian Structured Additive Distributional Regression for Multivariate Responses. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 2015; 64 (4): 569–591
- Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, 2023; 13
- Kneib T, Silbersdorff A, Säfken B. Rage Against the Mean – A Review of Distributional Regression Approaches. *Econometrics and Statistics*, 2023; 26: 99–123
- Kock L, Klein N. Truly Multivariate Structured Additive Distributional Regression. *Journal of Computational and Graphical Statistics*, 2025; 0 (0): 1–13
- Mayr A, Wistuba T, Speller J, Gude F, Hofner B. Linear or smooth? Enhanced model choice in boosting via deselection of base-learners. *Statistical Modelling*, 2023; 23 (5-6): 441–455
- Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2010; 72 (4): 417–473
- Midtfjord AD, De Bin R, Huseby AB. A copula-based boosting model for time-to-event prediction with dependent censoring. 2022; arXiv preprint arXiv: 2210.04869 (stat.ME).
- Strömer A, Klein N, Staerk C, Faschingbauer F, Klinkhammer H, Mayr A. Enhanced variable selection for boosting sparser and less complex models in distributional copula regression. *Statistics in Biosciences*, 2025a
- Strömer A, Klein N, Staerk C, Klinkhammer H, Mayr A. Boosting multivariate structured additive distributional regression models. *Statistics in Medicine*, 2023; 42 (11): 1779–1801
- Strömer A, Klein N, Van Keilegom I, Mayr A. Modelling dependent censoring in time-to-event data by boosting copula regression. Under Review by *Lifetime Data Analysis*, 2025b

- Strömer A, Staerk C, Klein N, Weinhold L, Titze S, Mayr A. Deselection of base-learners for statistical boosting – with an application to distributional regression. *Statistical Methods in Medical Research*, 2022; 31 (2): 207–224
- Thomas J, Hepp T, Mayr A, Bischl B. Probing for Sparse and Fast Variable Selection with Model-Based Boosting. *Computational and Mathematical Methods in Medicine*, 2017
- Wetscher M, Seiler J, Stauffer R, Umlauf N. Stagewise Boosting Distributional Regression. 2024; arXiv preprint arXiv: 2405.18288 (stat.ME).

Publication list during the PhD

- Enengl S, Rath W, Kehl S, Oppelt P, Mayr A, Strömer A, Eichinger T, Lasting J, Stelzl P. Differences between Current Clinical Practice and Evidence-Based Guideline Recommendations Regarding Tocolysis– an Austria-wide Survey. *Geburtshilfe und Frauenheilkunde*, 2025; 85 (02): 180–189
- Grabert J, Mohsen G, Diepenseifen C, Heister U, Breil M, Rohner M, Graeff I, Kappler J, Gutbrod K, Schewe J, Kunsorg A, Mayr A, Strömer A, Stoppe C, Zimmer S, Hoffmann U, Duerr G, Wittmann M, Velten M. The impact of preexisting omega-3 fatty acid serum levels on outcomes following out-of-hospital cardiac arrest: A comprehensive investigation. *Life Sciences*, 2025: 123770
- Hilbert A, Staerk C, Strömer A, Mansfeld T, Sander J, Seyfried F, Kaiser S, Dietrich A, Mayr A. Nonnormative Eating Behaviors and Eating Disorders and Their Associations With Weight Loss and Quality of Life During 6 Years Following Obesity Surgery. *JAMA Network Open*, 2022; 5 (8): e2226244
- Hilbert A, Strömer A, Staerk C, Schreglmann B, Mansfeld T, Sander J, Seyfried F, Kaiser S, Stroh C, Dietrich A, Schmidt R, Mayr A. Multivariate Trajectories of Weight and Mental Health and Their Prognostic Significance 6-Years After Obesity Surgery. *International Journal of Eating Disorders*, 2025; 58 (11): 2214–2226
- Martynov I, Tobi L, Strömer A, Mayr A, Heinz A, Ebinger M, Sparber-Sauer M, Vahdad R, Seitz G. Prevalence of Locoregional and Distant Lymph Node Metastases in Children and Adolescent/Young Adults with Soft Tissue Sarcomas: a Bayesian Meta-analysis of Proportions. *eClinicalMedicine*, 2025
- Mohsen G, Strömer A, Mayr A, Kunsorg A, Stoppe C, Wittmann M, Velten M. Effects of Omega-3 Fatty Acids on Postoperative Inflammatory Response: A Systematic Review and Meta-Analysis. *Nutrients*, 2023; 15 (15): 3414
- Pondorfer SG, Heinemann M, Wintergerst MWM, Pfau M, Strömer A, Holz FG, Finger RP. Detecting vision loss in intermediate age-related macular degeneration: A comparison of visual function tests. *PLOS ONE*, 2020; 15 (4): 1–12

- Simonini C, Strizek B, Strömer A, Gembruch U, Geipel A. Prenatal diagnosis and outcome of fetal urinomas in relation to the underlying etiology. *Prenatal Diagnosis*, 2024; 44 (2): 138–147
- Strömer A, Klein N, Staerk C, Faschingbauer F, Klinkhammer H, Mayr A. Enhanced variable selection for boosting sparser and less complex models in distributional copula regression. *Statistics in Biosciences*, 2025a
- Strömer A, Klein N, Staerk C, Klinkhammer H, Mayr A. Boosting multivariate structured additive distributional regression models. *Statistics in Medicine*, 2023; 42 (11): 1779–1801
- Strömer A, Klein N, Van Keilegom I, Mayr A. Modelling dependent censoring in time-to-event data by boosting copula regression. Under Review by Lifetime Data Analysis, 2025b
- Strömer A, Staerk C, Klein N, Weinhold L, Titze S, Mayr A. Deselection of base-learners for statistical boosting – with an application to distributional regression. *Statistical Methods in Medical Research*, 2022; 31 (2): 207–224
- Thudium M, Braun L, Strömer A, Mayr A, Menzenbach J, Saller T, Soehle M, Kornilov E, Hilbert T. Cerebral Overperfusion Despite Reduced Cortical Metabolism Is Associated with Postoperative Delirium in Cardiac Surgery Patients: A Prospective Observational Study. *Journal of Clinical Medicine*, 2024; 13 (21)

5 Acknowledgements

First of all, I would like to thank my supervisor, Prof. Dr. Andreas Mayr, for his invaluable feedback and for creating such a welcoming and supportive research environment. I am also grateful to Prof. Dr. Nadja Klein for her constructive critiques and valuable suggestions, which improved my work. I am thankful to the members of my dissertation committee, Prof. Dr. Matthias Schmid and Prof. Dr. Robert Finger, for their time and input.

Many thanks to my former colleagues at IMBIE for the supportive and friendly atmosphere that made my time there especially enjoyable and to my colleagues at IMBS for contributing to a positive research environment. Special thanks go to Moritz and Jan for their careful proofreading and to Nikolai for assistance with formalities. I am particularly grateful to Jan, who was a wonderful office mate and always willing to help with any questions. I would also like to thank office 607 – Nikolai, David and Jenny – for their support in various ways.

Finally, and most importantly, I would like to thank my family and friends for their unwavering encouragement, patience and support throughout this journey. This dissertation would not have been possible without them.

6 Statement

The work was carried out at the Institute of Medical Biometry, Informatics and Epidemiology under the supervision of Prof. Dr. Andreas Mayr.

Publication A: "Multivariate trajectories of weight and mental health and their prognostic significance six years after obesity surgery".

I conceptualized together with CS and AM the statistical methods for the analysis in consultation with AH. This study is part of the prospective Psychosocial Registry for Obesity Surgery (PRAC) study, which is being implemented at six surgical treatment centers in Germany and is registered in the German Clinical Trials Register (DRKS00006749). AH, BS, TM, JS, FS, SK, CS, AD, and RS carried out the data acquisition. I preprocessed the data for the following analysis, including the imputation of missing values and conducted statistical analysis, model development and refinement in close collaboration with CS and AM. AH, AM and CS provided feedback within regular discussions about the results of the analyses. I drafted the Data Analytic Plan and participated in editing and refining the manuscript in collaboration with AH, AM, CS and RS, with critical revisions and final approval by all authors.

Publication B: "Boosting multivariate structured additive distributional regression model".

I conceptualized and implemented the methodological approach in R in collaboration with AM, CS, and NK. HK provided the data from the UK Biobank (Application No. 81202). The Healthcare data from Australia is openly available. The Nigeria dataset is available upon request from the Demographic and Health Survey (DHS). I have pre-processed the data for the following analysis. I implemented and conducted all statistical analyses and generated all visualizations (e.g., figures, tables). AM, CS and NK provided feedback within regular discussions about the results of all analyses. I created the first draft of the manuscript and answered most of the reviewers' questions (including changes in the main manuscript) during the revision process. AM, CS and NK provided valuable feedback and suggested changes for the final version of the manuscript.

Publication C: "Modelling dependent censoring in time-to-event data by boosting copula regression".

I conceptualized the methodological approach and implemented it in R, with all co-authors contributing to the conceptualization. I performed pre-processing of the application data. I implemented and conducted all simulation studies. The dataset is publicly available in the R package `dirttee`. I evaluated the simulation study and the applications and generated all visualizations (e.g., figures, tables). The co-authors provided feedback within regular discussions about the results of all analyses. I created the first draft of the manuscript. All co-authors provided valuable feedback and suggested changes for the final version of the manuscript.

Publication D: "Deselection of base-learners for statistical boosting — with an application to distributional regression".

I conceptualized the methodological approach and implemented it in R in collaboration with AM and CS. ST provided the GCKD data and LW performed data preprocessing. I implemented and conducted all statistical analyses and generated all visualizations (e.g., figures, tables). AM and CS provided feedback within regular discussions about the results of all analyses. ST contributed feedback specifically regarding the application of the GCKD data. I created the first draft of the manuscript and answered most of the reviewers' questions (including changes in the main manuscript) during the revision process. All co-authors provided valuable feedback and suggested changes for the final version of the manuscript.

Publication E: "Enhanced variable selection for boosting sparser and less complex models in distributional copula regression".

I conceptualized and implemented the methodological approach in R in collaboration with AM and CS. I designed the data-generating process for the simulations and pre-processed the data for all applications. HK provided the data from the UK Biobank (Application No. 81202). FF provided the fetal ultrasound data. I implemented and conducted all statistical analyses and generated all visualizations (e.g., figures, tables). AM and CS provided feedback during regular discussions about the results of all analyses. FF provided input specifically on

the analysis of the fetal ultrasound data. I created the first draft of the manuscript and answered most of the reviewers' questions. (including changes in the main manuscript) of the revision process. All co-authors provided valuable feedback and suggested changes for the final version of the manuscript.

I confirm that I have written this thesis independently and have not used any sources or aids other than those specified by me.