Smooth risk-based predictive approaches to treatment effect heterogeneity: A simulation study

Alexandros Rekkas[1], Peter R. Rijnbeek[1], . . . , Ewout W. Steyerberg[2], David van Klaveren[3]

[1] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands

[2] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

[3] Department of Public Health, Erasmus University Medical Center, Rotterdam, Netherlands

Abstract

Objective: Simulation study to compare different risk-based approaches to estimating individualized treatment effects within the RCT setting. **Study Design and Setting:** Starting from a base case scenario that assumes a true constant treatment effect, we considered a total of 66 scenarios for introducing non-constant effects and evaluating methods under different sample sizes and baseline risk prediction performance. We compared 7 methods for predicting absolute benefit: A constant treatment effect model, a risk stratified approach, a model including a linear interaction of the baseline risk linear predictor with treatment, 3 restricted cubic spline smoothing models of increasing flexibility (3, 4 and 5 knots) and an adaptive model selection method based on Akaike's Information Criterion. We evaluated performance using root mean squared error, discrimination for benefit and calibration for benefit (i.e., observed vs. predicted risk difference in treated vs. untreated). **Results:** The model including a linear interaction of the risk linear predictor with treatment had adequate performance that was robust under the majority of the simulation scenarios. Methods using restricted cubic spline smoothing required larger sample sizes and higher prediction AUC to achieve adequate performance. The adaptive approach's performance was comparable to the performance of the best model in each scenario. **Conclusion:** In most cases using a model just including a linear interaction of the risk linear predictor with treatment adequately predicts absolute benefit.

1 Introduction

Within the setting of patient-centered outcomes research, predictive approaches for assessing heterogeneity of treatment effects (HTE) aim at the development of models predicting either individualized effects or which of two (or more) treatments is better for an individual [1]. In prior work, we divided such methods in three broader categories based on the reference class used for defining patient similarity when making individualized predictions or recommendations [2]. Risk-modeling approaches use prediction of baseline risk as the reference; treatment effect modeling approaches also model treatment-covariate interactions, in addition to risk factors; optimal treatment regime approaches focus on developing treatment assignment rules and therefore rely heavily on modeling treatment effect modifiers.

Risk-modeling approaches to predictive HTE analyses provide a viable option in the absence of well-established treatment effect modifiers [3,4]. In simulations, modeling of effect modifiers in the form of treatment-covariate interactions often led to miscalibrated predictions of benefit, while risk-based methods proved quite robust [5]. Most often, risk-modeling approaches are carried out in two steps: first a risk prediction model is developed externally or internally on the entire RCT population, "blinded" to treatment; then the RCT population is stratified using this prediction model to evaluate risk-based treatment effect variation [6]. However, even though estimates at the risk subgroup level are accurate, this does not apply on the individual level, especially for patients with predicted risk at the boundaries of the risk intervals. Therefore, the risk-stratified approach should be used for exploring and presenting an overview of HTE, while inferences on the individual level should be made with caution.

We aimed to provide an overview of methods that can be used to move from a risk-stratified approach to a continuous one using common smoothing techniques. These methods extend the risk-based framework of predictive HTE analyses to allow predictions on the individual level, within the RCT setting. We carried out a simulation study to compare the performance of these methods under different settings of increasing non-linearity of treatment effects. Finally, we carried out an application on real data as a demonstration of the considered techniques.

2 Methods

2.1 Simulation scenarios

In the simulated datasets based on the base-case scenario treatment was allocated at random using a 50/50 split. For each patient we simulated 8 baseline covariates, where $x_1, \ldots, x_4 \sim N(0,1)$ and $x_5, \ldots, x_8 \sim B(1, 0.2)$.

Outcomes for patients in the control arm were generated from a logistic regression model including all baseline covariates. Coefficient values were such, so that the prediction model had an AUC of 0.75 and an event rate of 20 in the control arm was achieved. Outcomes in the treatment arm were created using the same logistic regression model, including a constant treatment effect odds ratio (OR) of 0.8. The generated samples of the base-case scenario were of size $n = 4,250$ (80 power for the detection of an unadjusted OR of 0.8).

We evaluated the effect of sample size considering additional scenarios with sample sizes of $1,064$ and $17,000$. We also evaluated the effect of prediction performance, adjusting the baseline covariate coefficients, so that AUC values of 0.65 and 0.80 were achieved when validating in a simulated dataset of $500,000$ patients.

A true logistic regression model with a constant treatment effect (constant OR) implies that outcome risk in the treatment arm is a straight line parallel to the first diagonal on the *log-odds* scale, with distance equal to $\log(\text{OR})$. (See Figure XX of the supplementary material). We assessed the effect of stronger and absent relative treatment effects (OR $= 0.5$ or OR $= 1$). We also relaxed the assumption that the line should be parallel to the diagonal, considering moderate and stronger linear deviations. Finally, we dropped the assumption of linearity allowing for quadratic deviations.

We also considered scenarios with treatment-covariate interactions. These scenarios include 4 weak interactions ($\text{OR}_{t_x=1}/\text{OR}_{t_x=0} = 0.82$), 4 strong interactions ($\text{OR}_{t_x=1}/\text{OR}_{t_x=0} = 0.61$), and 2 weak and 2 strong interactions. Combining all these different settings resulted in a simulation study of 66 scenarios. The exact settings for each scenario are available in the supplementary material.

2.2 Individualized risk-based benefit predictions

All methods assume that a risk prediction model is available and can be used to assign individualized predictions. For the simulations we developed the prediction models internally and blinded to treatment using logistic regression including main effects for all baseline covariates and treatment. Predictions on individuals were made setting treatment to 0.

The **stratified HTE method** was suggested as an alternative to traditional subgroup analyses. Patients are stratified into equally-sized risk strata—in this case based on risk quartiles. Absolute effects are estimated using the differences in event rates between treatments within risk quarters. We considered this approach as a reference, expecting it to perform worse than the other candidates, as its objective is not individualized benefit prediction.

We also considered a set of **linear methods**. We fit separate models within treatment arms using only the treatment indicator and the linear predictor of the internal risk prediction model. In the simpler case, we assume

a constant relative treatment effect (OR). Absolute benefit is then estimated from expit$(lp + \log(\text{OR}))$, where expit$(x) = \frac{e^x}{1+e^x}$ and *lp* is the linear predictor of the prediction model. A different approach fits a logistic regression using treatment, risk linear predictor and their interaction. In this case, absolute benefit is estimated from expit$(\beta_0 + \beta_{lp} lp) -$ expit$(\beta_0 + \beta_{t_x} + (\beta_{lp} + \beta_*) lp)$. We will refer to this method as the linear interaction approach.

Finally, we used restricted cubic splines (RCS) to relax the linearity assumption on the effect of the linear predictor [7]. We compared the results for 3, 4 and 5 knots when fitting the splines to introduce increasing flexibility to the methods considered.

2.3 Evaluation metrics

Assuming that $\tau(\boldsymbol{x}) = E\{y \mid \boldsymbol{x}, t_x = 0\} - E\{y \mid \boldsymbol{x}, t_x = 1\}$ is the true benefit for each patient and $\hat{\tau}(x)$ is the estimated benefit from a method under study, the ideal loss function to use for the considered methods would be the unobservable root mean squared error $E\big\{(\hat{\tau} - \tau)^2 \mid \boldsymbol{x}\big\}$. However, in simulations the true patient-level benefit is available. Therefore, we estimated RMSE from

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^{n} \big(\tau(\boldsymbol{x}_i) - \hat{\tau}(\boldsymbol{x}_i)\big)^2$$

We also compared the predictive performance of the methods under study. We assessed discrimination using the c-for-benefit statistic described in [8]. Patients in each treatment arm are ranked based on their predicted benefit and then are matched 1:1, dropping patients in the larger treatment arm without a pair. We define *observed* pair-specific treatment benefit as the difference of observed outcomes between the untreated and the treated patient of each pair. Pair-specific *predicted* benefit is defined as the average of predicted benefits within each pair. Then, c-for-benefit is defined as the probability that from two randomly chosen matched patient pairs with unequal observed benefit, the pair with greater observed benefit also has a higher predicted benefit.

We evaluated calibration in a similar manner, using the integrated calibration index (ICI) for benefit [9]. After creating pairs based on predicted benefit, observed benefits are regressed on the predicted benefits using a locally weighted scatterplot smoother (loess). The ICI is the area between the loess fit and the diagonal, with values closer to 0 indicating better calibration.

3 Results

3.1 Simulations

The model including a constant relative treatment effect had the lowest median RMSE in scenarios where the base case of true constant relative treatment effect (OR = 0.8, N = 4,250 and AUC = 0.75) or moderate relative deviations were considered (Figure 1; Panel A). However, when we considered strong linear and quadratic deviations from the base case the linear interaction model performed best (Figure 1; Panels B and C). Only in the case of strong quadratic deviations models including RCS smoothing (3 knots) performed equally well to the linear interaction method. Increasing the number of knots in RCS smoothing resulted in higher error rates across all scenarios. The performance of the adaptive approach was quite comparable with the best performing model in each scenario.

When we increased the sample size (N = 17,000), in all scenarios the model more in agreement with the underlying settings had the lowest error rates. Under the base case of constant relative treatment effects (OR = 0.8), the model including a constant relative treatment effect had the lowest RMSE (Figure 2; Panel A). When introducing moderate and strong linear deviations the linear interaction model performed best (Figure 2; Panels B and C). In the case of true strong quadratic deviations from the base case of constant relative effects, the more flexible RCS smoothing models (3 knots) had the lowest RMSE. Again, the increased flexibility of RCS smoothing with higher number of knots resulted in overfitting and worse performance (Figure 2; Panel D).

When we increased the true prediction AUC to 0.85, models including RCS smoothing had the lowest RMSE in the presence of strong quadratic deviations from the base case of true constant relative treatment effects (Figure 3; Panel D). However, with milder deviations, the linear interaction model had the lowest RMSE with the RCS smoothing methods (3 knots) being a close second (Figure 3; Panels B and C). Increasing the number of knots of RCS smoothing resulted in increased RMSE, which was less pronounced in the case of strong quadratic deviations.

All candidate methods demonstrated comparable discrimination for benefit in all scenarios where linear and quadratic deviations from the base case of constant treatment effect were considered (N=4,250; AUC = 0.75). However, models including a linear interaction with the risk linear predictor tended to present much lower variability compared to all other model-based and smoothing approaches (Figure 4). We also observed an increasing trend of discrimination for benefit variability with increasing number of restricted cubic spline knots in all scenarios.

When focusing on calibration for benefit, the linear interaction model had the lowest median ICI for benefit in the majority of the scenarios except for the scenarios where moderate linear deviations from the base case were considered. In that case constant treatment effect models demonstrated the best performance, very comparable

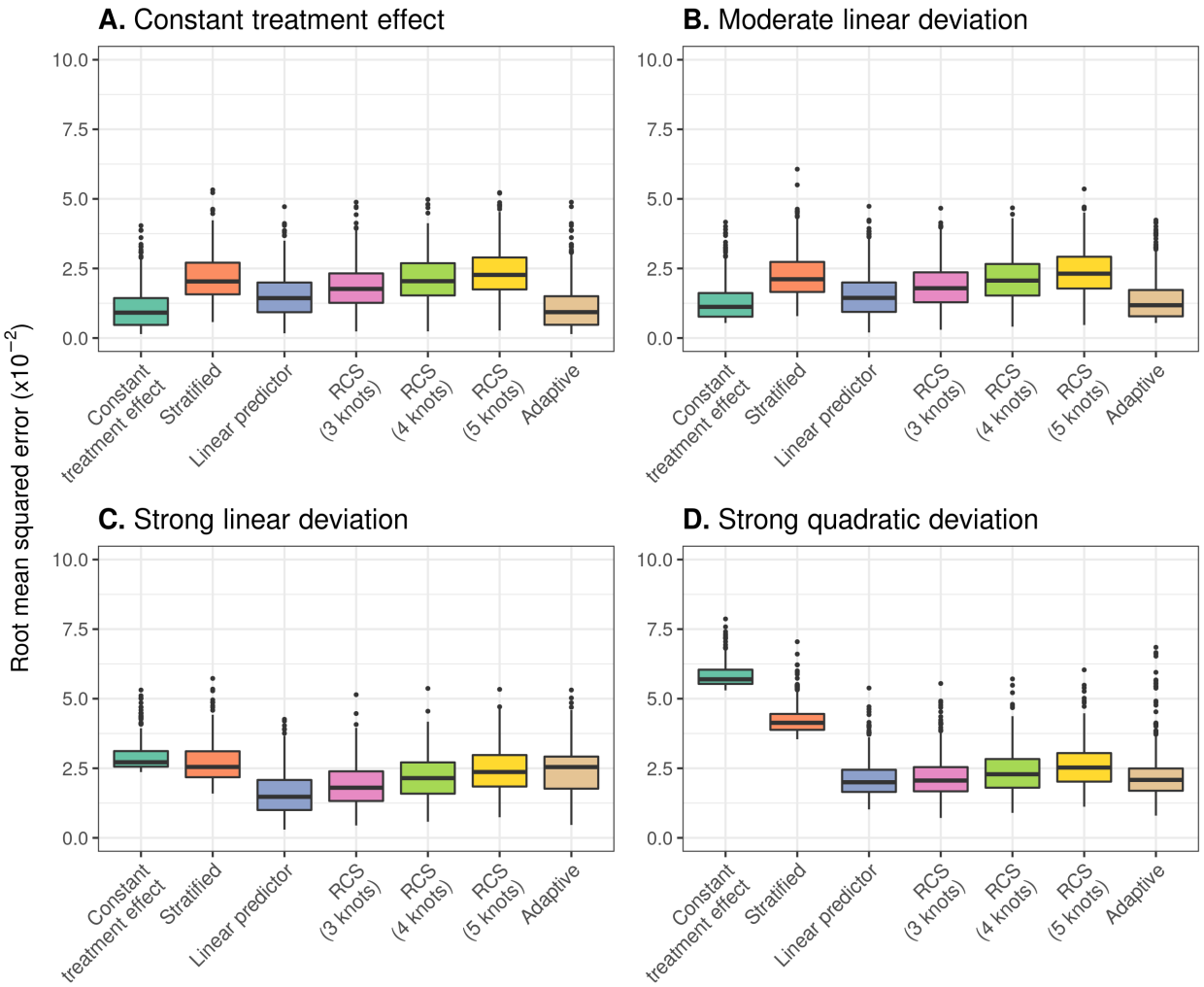

Figure 1: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000. *Panel A* presents the results under the base case scenario of true constant relative treatment effect, with a true prediction AUC of 0.75 and sample size of 4250; *Panel B* presents the results under moderate linear deviations from the base case of constant treatment effects, while holding true prediction AUC and sample size constant; *Panel C* presents the results for strong linear deviations from the base case scenario of constant relative treatment effects; *Panel D* presents the results for strong quadratic deviations from the base case scenario of constant relative treatment effects.

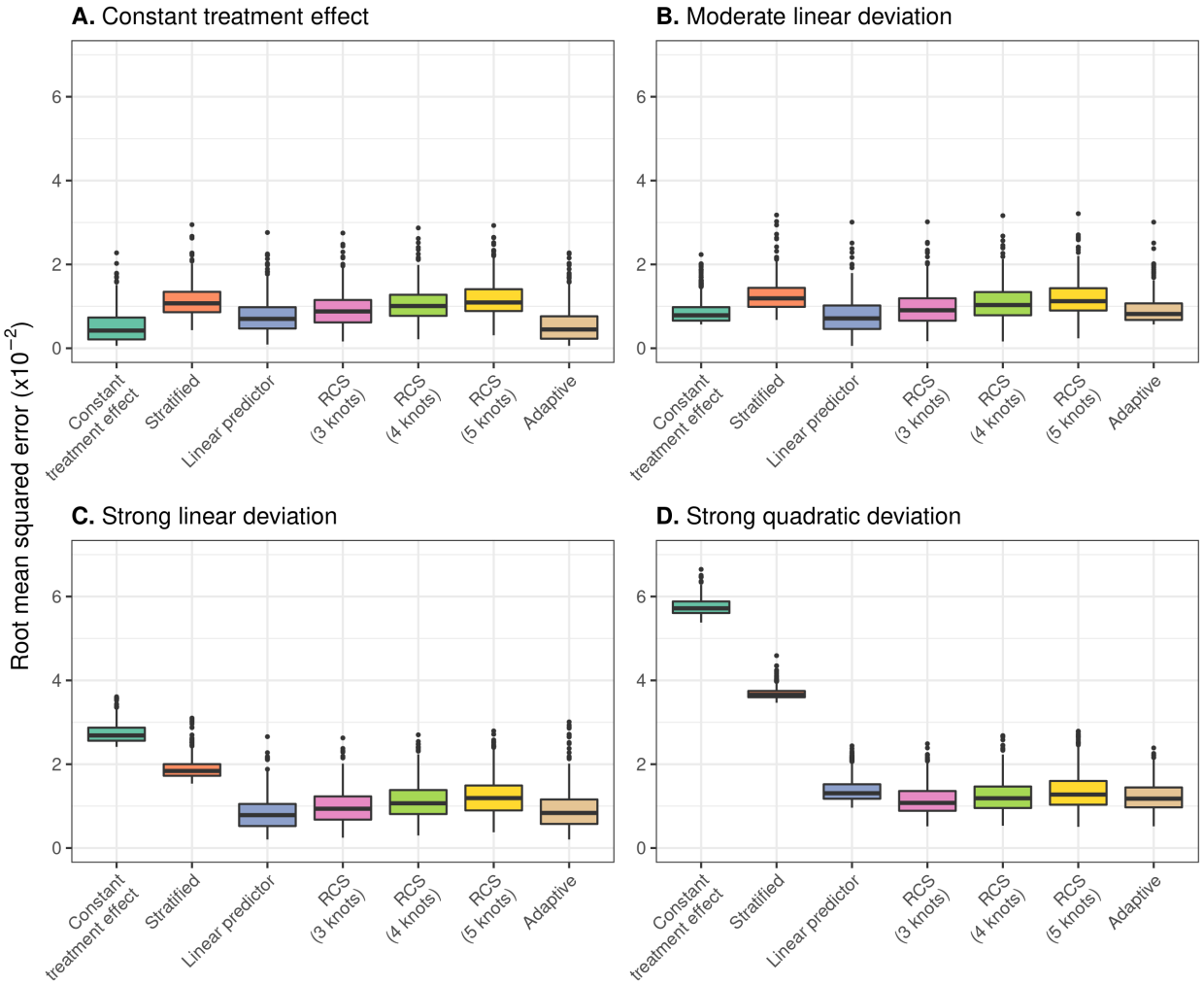

Figure 2: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000. *Panel A* presents the results under the base case scenario of true constant relative treatment effect, with a true prediction AUC of 0.75 and sample size of 17,000; *Panel B* presents the results under moderate linear deviations from the base case of constant treatment effects, while holding true prediction AUC and sample size constant; *Panel C* presents the results for strong linear deviations from the base case scenario of constant relative treatment effects; *Panel D* presents the results for strong quadratic deviations from the base case scenario of constant relative treatment effects.

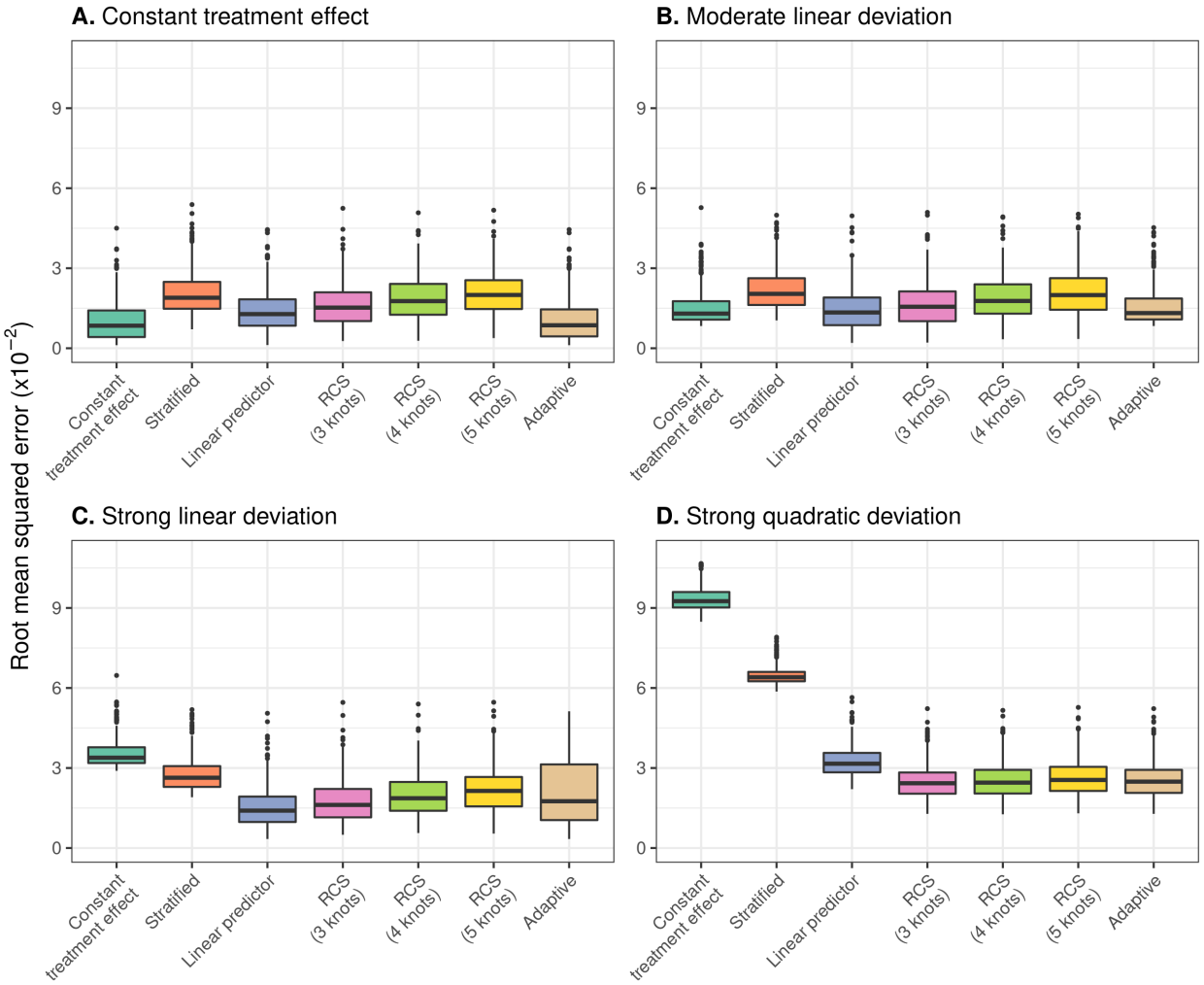

Figure 3: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000. *Panel A* presents the results under the base case scenario of true constant relative treatment effect, with a true prediction AUC of 0.85 and sample size of 4,250; *Panel B* presents the results under moderate linear deviations from the base case of constant treatment effects, while holding true prediction AUC and sample size constant; *Panel C* presents the results for strong linear deviations from the base case scenario of constant relative treatment effects; *Panel D* presents the results for strong quadratic deviations from the base case scenario of constant relative treatment effects.

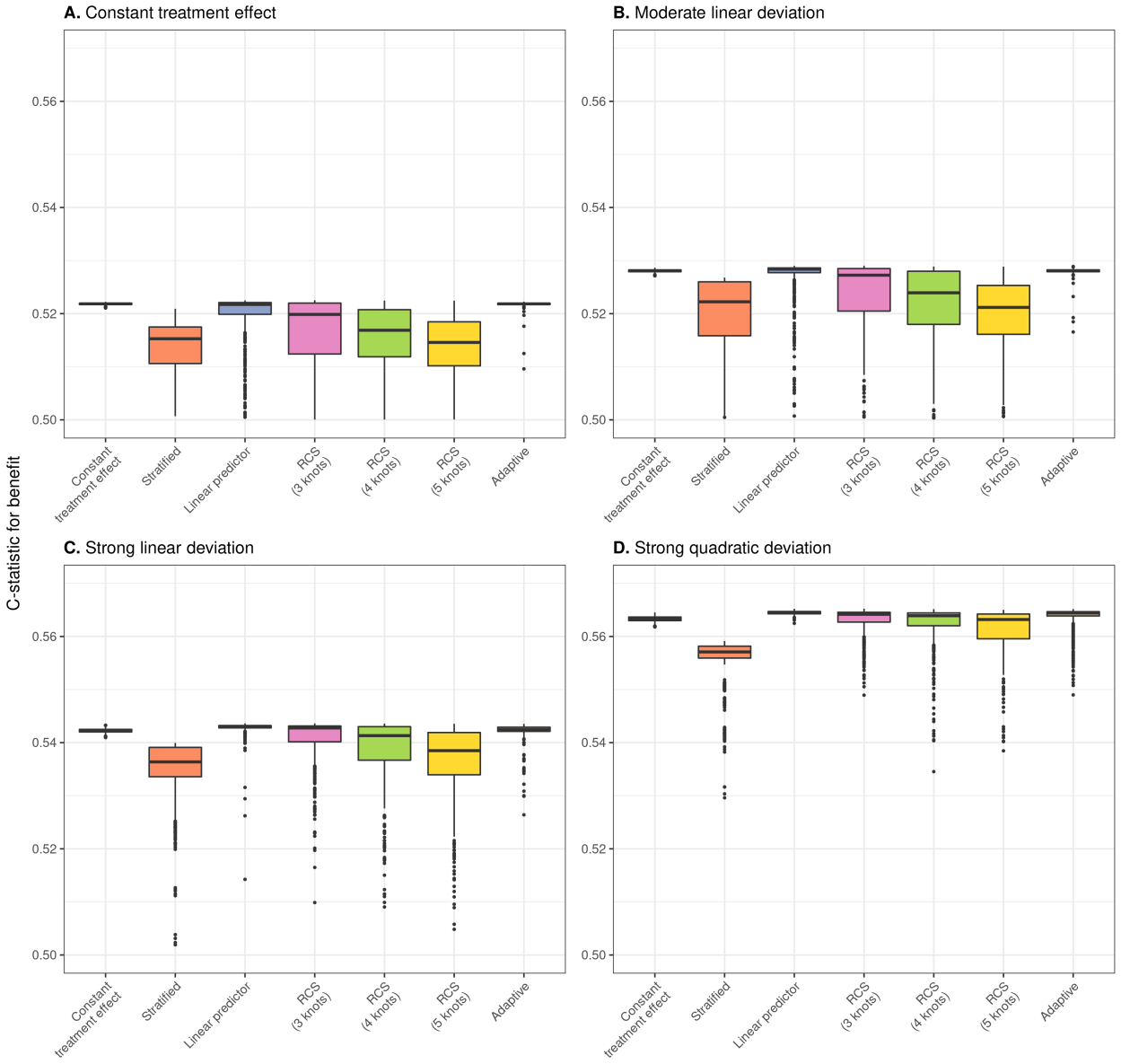

Figure 4: Discrimination for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. *Panel A* presents the results under the base case scenario of true constant relative treatment effect, with a true prediction AUC of 0.75 and sample size of 4250; *Panel B* presents the results under moderate linear deviations from the base case of constant treatment effects, while holding true prediction AUC and sample size constant; *Panel C* presents the results for strong linear deviations from the base case scenario of constant relative treatment effects; *Panel D* presents the results for strong quadratic deviations from the base case scenario of constant relative treatment effects.

to the linear interaction model's performance, nonetheless (Figure 5; Panels A, B, and C).

3.2 Real data

We demonstrate the different methods for individualizing treatment benefits using data from 30,510 patients with an acute myocardial infarction (MI) included in the GUSTO-I trial. 10,348 patients were randomized to tissue plasminogen activator (tPA) treatment and 20,162 were randomized to streptokinase. The outcome of interest was 30-day mortality, recorded for all patients.

In line with previous analyses [10,11], we fitted a logistic regression model with 6 baseline covariates, i.e. age, Killip class, systolic blood pressure, heart rate, an indicator of previous MI, and the location of MI, to predict 30-day mortality risk. A constant effect of treatment was included in the model. When deriving risk predictions for individuals we set the treatment indicator to 0. More information on model development can be found in the supplement.

We used the risk linear predictor to fit the the proposed methods under study for individualizing absolute benefit predictions. All methods had quite comparable results, in the sense that we predicted increasing benefits for patients with higher baseline risk predictions. In terms of c-for benefit (validated internally) all models had quite comparable performance ranging from 0.519 (RCS smoothing with 3 knots) to 0.542 (RCS smoothing with 4 knots). Similar conclusions could be drawn in terms of ICI-for-benefit which ranged from 0.0039 (linear interaction approach) to 0.0053 (RCS smoothing with 4 knots).

The adaptive approach picked the model with RCS smoothing with 4 knots, which had quite comparable performance to the smooth fit with 5 knots. However, for lower baseline risk these 2 models predicted implausible benefits and maybe should be avoided when applying such models in practice. The linear interaction model and the model with RCS smoothing (3 knots) made very similar predictions, while also followed quite closely the evolution of the stratified estimates. In this case, we observed that even using a simple model with a constant relative treatment effect can result in very similar benefit predictions as other more complex approaches.

4 Discussion

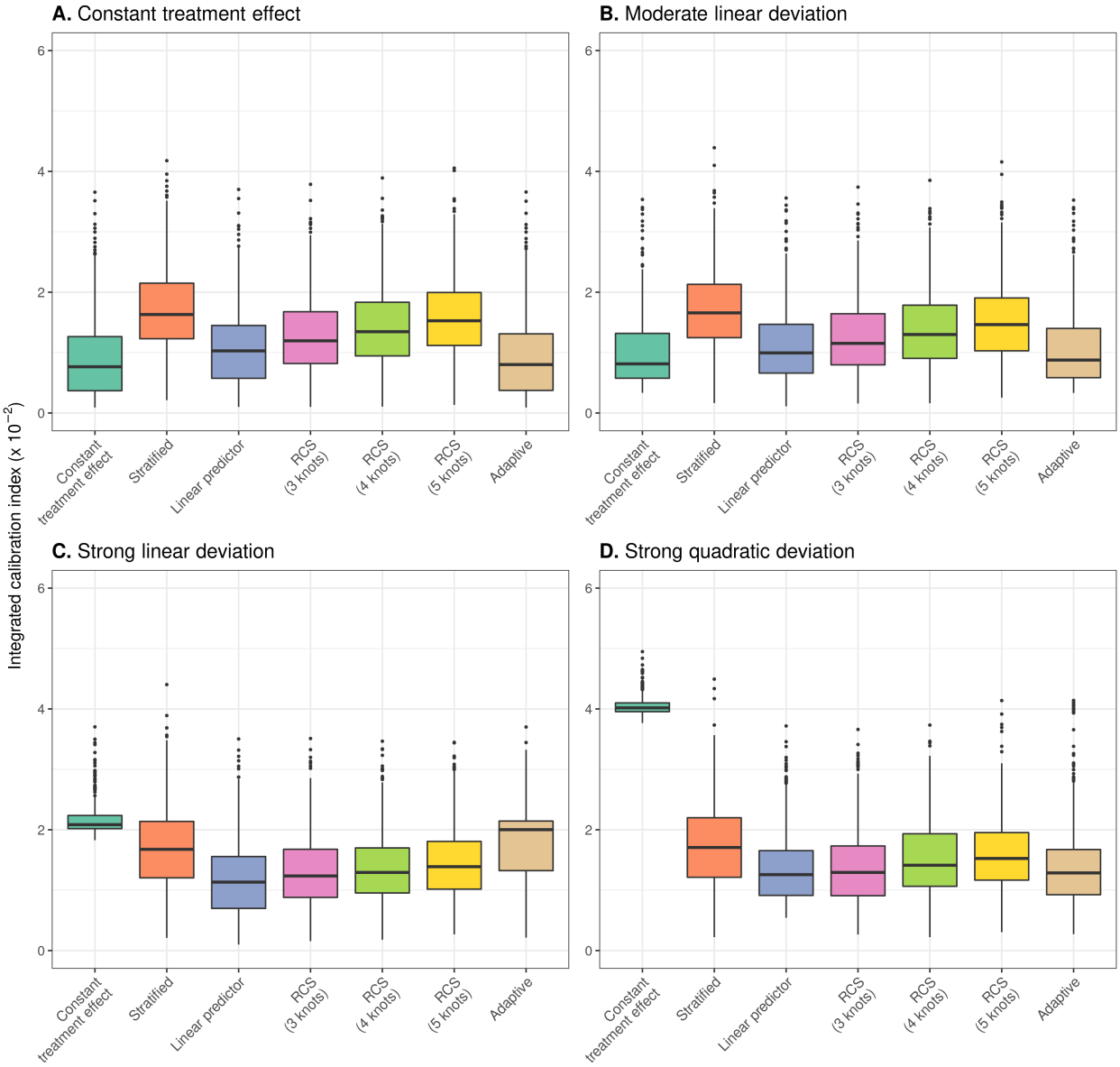

Figure 5: Calibration for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. *Panel A* presents the results under the base case scenario of true constant relative treatment effect, with a true prediction AUC of 0.75 and sample size of 4250; *Panel B* presents the results under moderate linear deviations from the base case of constant treatment effects, while holding true prediction AUC and sample size constant; *Panel C* presents the results for strong linear deviations from the base case scenario of constant relative treatment effects; *Panel D* presents the results for strong quadratic deviations from the base case scenario of constant relative treatment effects.

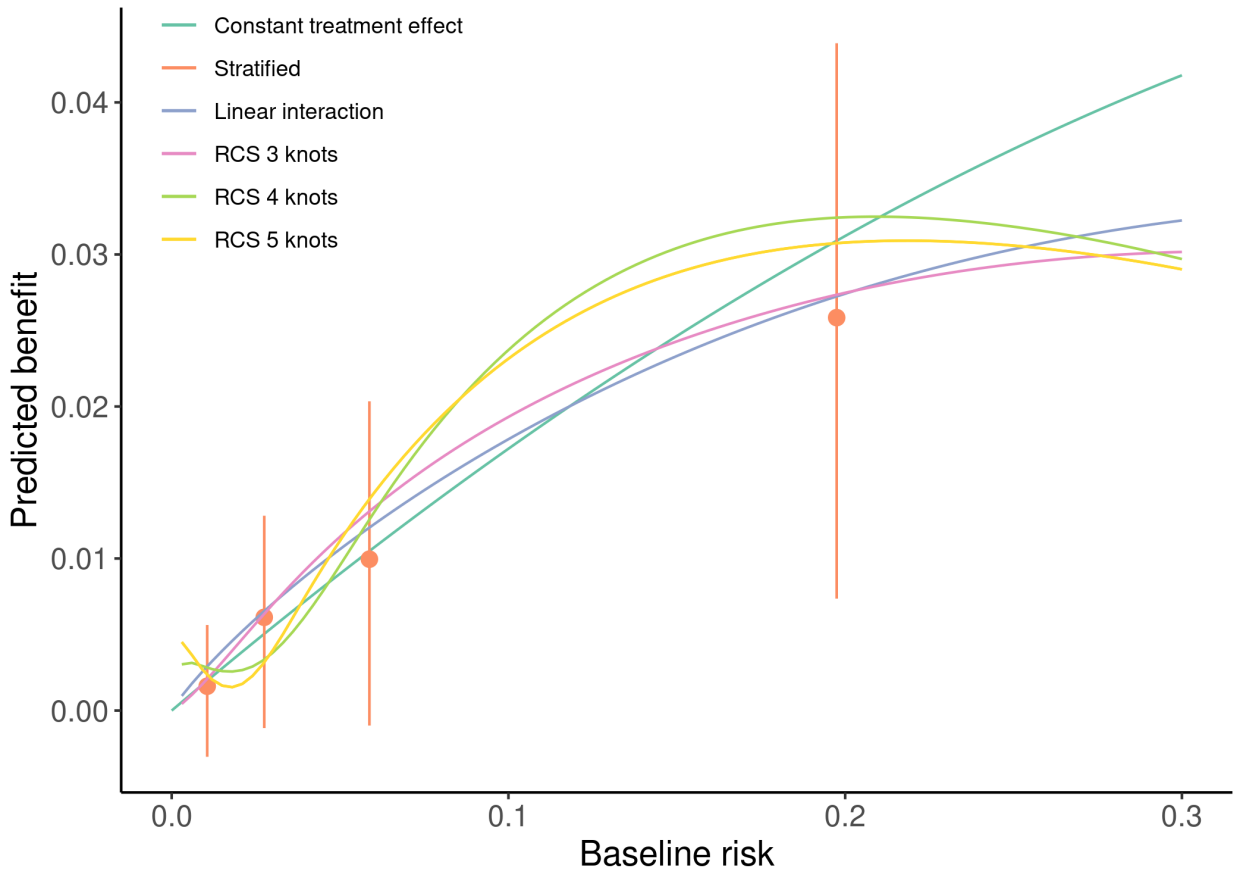

Figure 6: Individualized absolute benefit predictions based on baseline risk when using a constantn treatment effect approach, a linear interaction approach and RCS smoothing using 3,4 and 5 knots. Risk stratified estimates of absolute benefit are presented within quartiles of baseline risk as reference.

5 References

[1] Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. Journal of Clinical Epidemiology 2013;66:818–25. https://doi.org/10.1016/j.jclinepi.2013.02.009.

[2] Rekkas A, Paulus JK, Raman G, Wong JB, Steyerberg EW, Rijnbeek PR, et al. Predictive approaches to heterogeneous treatment effects: A scoping review. BMC Medical Research Methodology 2020;20. https://doi.org/10.1186/s12874-020-01145-1.

[3] Kent DM, Paulus JK, Klaveren D van, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Annals of Internal Medicine 2019;172:35. https://doi.org/10.7326/m18-3667.

[4] Kent DM, Klaveren D van, Paulus JK, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement: Explanation and elaboration. Annals of Internal Medicine 2019;172:W1. https://doi.org/10.7326/m18-3668.

[5] Klaveren D van, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. Journal of Clinical Epidemiology 2019;114:72–83. https://doi.org/10.1016/j.jclinepi.2019.05.029.

[6] Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. Trials 2010;11. https://doi.org/10.1186/1745-6215-11-85.

[7] Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: Determining relationships between predictors and response. JNCI Journal of the National Cancer Institute 1988;80:1198–202. https://doi.org/10.1093/jnci/80.15.1198.

[8] Klaveren D van, Steyerberg EW, Serruys PW, Kent DM. The proposed "concordance-statistic for benefit" provided a useful metric when modeling heterogeneous treatment effects. Journal of Clinical Epidemiology 2018;94:59–68. https://doi.org/10.1016/j.jclinepi.2017.10.021.

[9] Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine 2019;38:4051–65. https://doi.org/10.1002/sim.8281.

[10] Califf RM, Woodlief LH, Harrell FE, Lee KL, White HD, Guerci A, et al. Selection of thrombolytic therapy for individual patients: Development of a clinical model. American Heart Journal 1997;133:630–9. https:

//doi.org/10.1016/s0002-8703(97)70164-9.

[11] Steyerberg EW, Bossuyt PMM, Lee KL. Clinical trials in acute myocardial infarction: Should we adjust for baseline characteristics? American Heart Journal 2000;139:745–51. https://doi.org/10.1016/s0002-8703(00)90001-2.