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

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 $\log(\text{OR})$. 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 individual 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 within each treatment arm. 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

3 Results

3.1 Simulations

Under the base case of constant relative treatment effect, the model assuming a constant treatment effect had the lowest median RMSE, regardless of true prediction AUC and sample size (Figure 1). Linear interaction models demonstrated comparable performance. Among the RCS smoothing methods, the one fitted with 3 knots always performed best, while the increased flexibility achieved when increasing the knots resulted in overfitting and worse performance. The adaptive approach under all scenarios performed similar to the model with smaller RMSE in all scenarios.

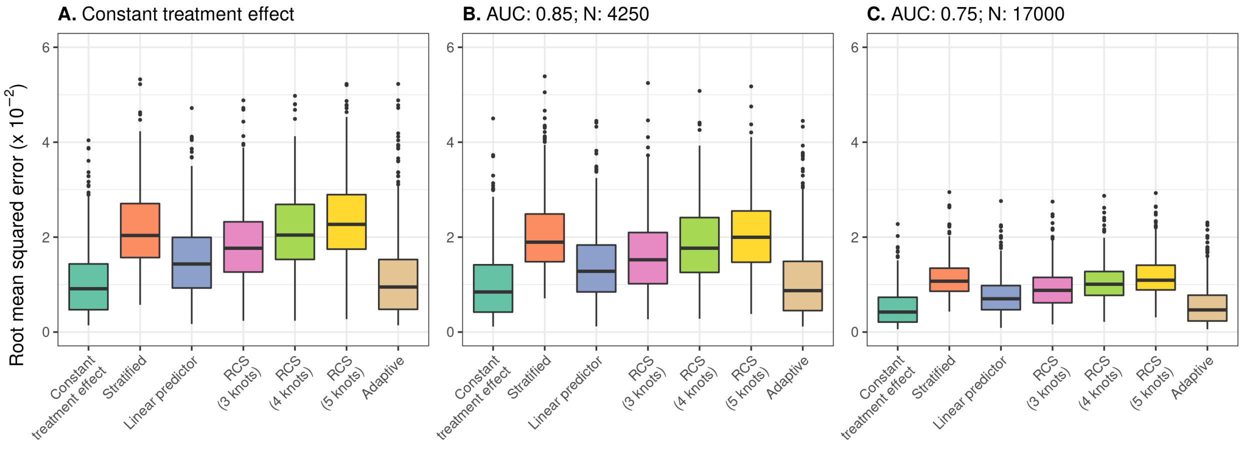

Figure 1: Caption

When we introduced deviations from the base case of constant relative treatment effects, while keeping fixed both the sample size (N = 4250) and the true prediction AUC (0.75) the linear interaction model had the lowest RMSE (Figure 2; Panels A, D, and G). When these deviations were moderate (Figure 2; Panel A) the constant treatment effect model had comparable performance to the linear interaction model. This can be attributed to the fact that such deviations are quite mild and absolute benefits maintain similar patterns across baseline risk. On the contrary, when strong quadratic deviations were considered the constant effect model's RMSE sharply increased, while the more flexible method of RCS smoothing (3 knots) preformed very well (Figure 2; Panel G). Again, increasing the number of knots increased RMSE, indicating overfitting.

When we increased the true prediction AUC to 0.85, models including RCS smoothing had the lowest RMSE when strong quadratic deviations from the base case of constant relative treatment effects were assumed (Figure 2; Panel H). However, with milder deviations, the linear interaction model had the lowest RMSE with the RCS smoothing methods (3 knots) being a close second (Figure 2; Panels B and E). Increasing the number of knots of RCS smoothing resulted in increased RMSE, which was less pronounced in the case of strong quadratic deviations. We observed similar results when we increased the sample size to 17000, while keeping the true prediction AUC constant at 0.75 (Figure 2; Panels C, F, and I).

When focusing on the different scenarios where true treatment-covariate interactions were considered all methods had similar RMSE performance, (Figure 3; Panels B, C and D). In case of strong treatment-covariate interactions the constant effect model had slightly increased RMSE (0.096; [0.092, 0.103]) compared to the other methods. The linear interaction model with the risk linear predictor had the lowest RMSE (0.088; [0.08, 0.095]).

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 (Figure 4). 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. We also observed an increasing trend of discrimination for benefit variability with increasing number of restricted cubic spline knots in all scenarios. This is evidence that the increased flexibility of these methods often led to overfitting.

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

3.2 Real data

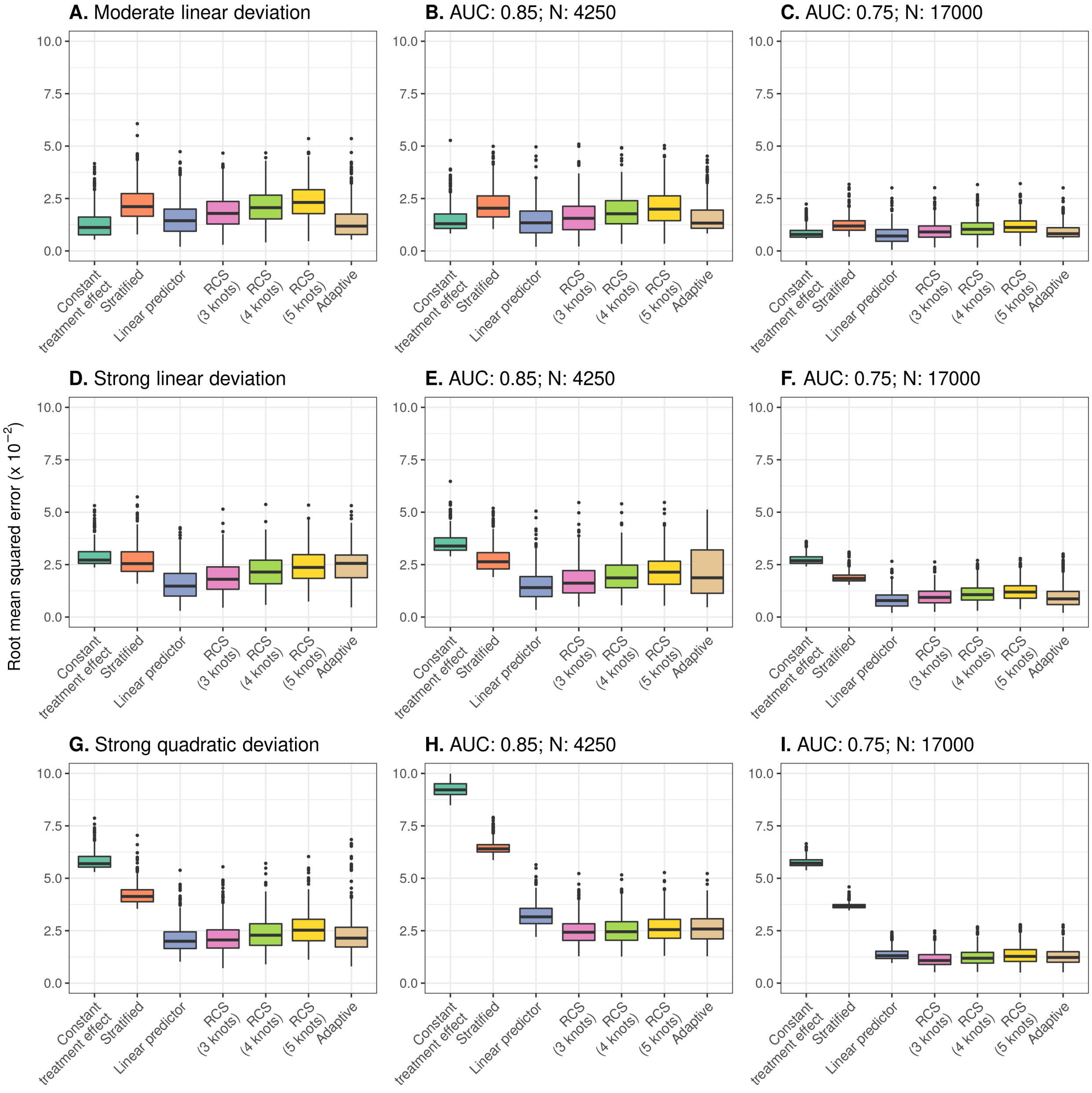

Figure 2: Caption

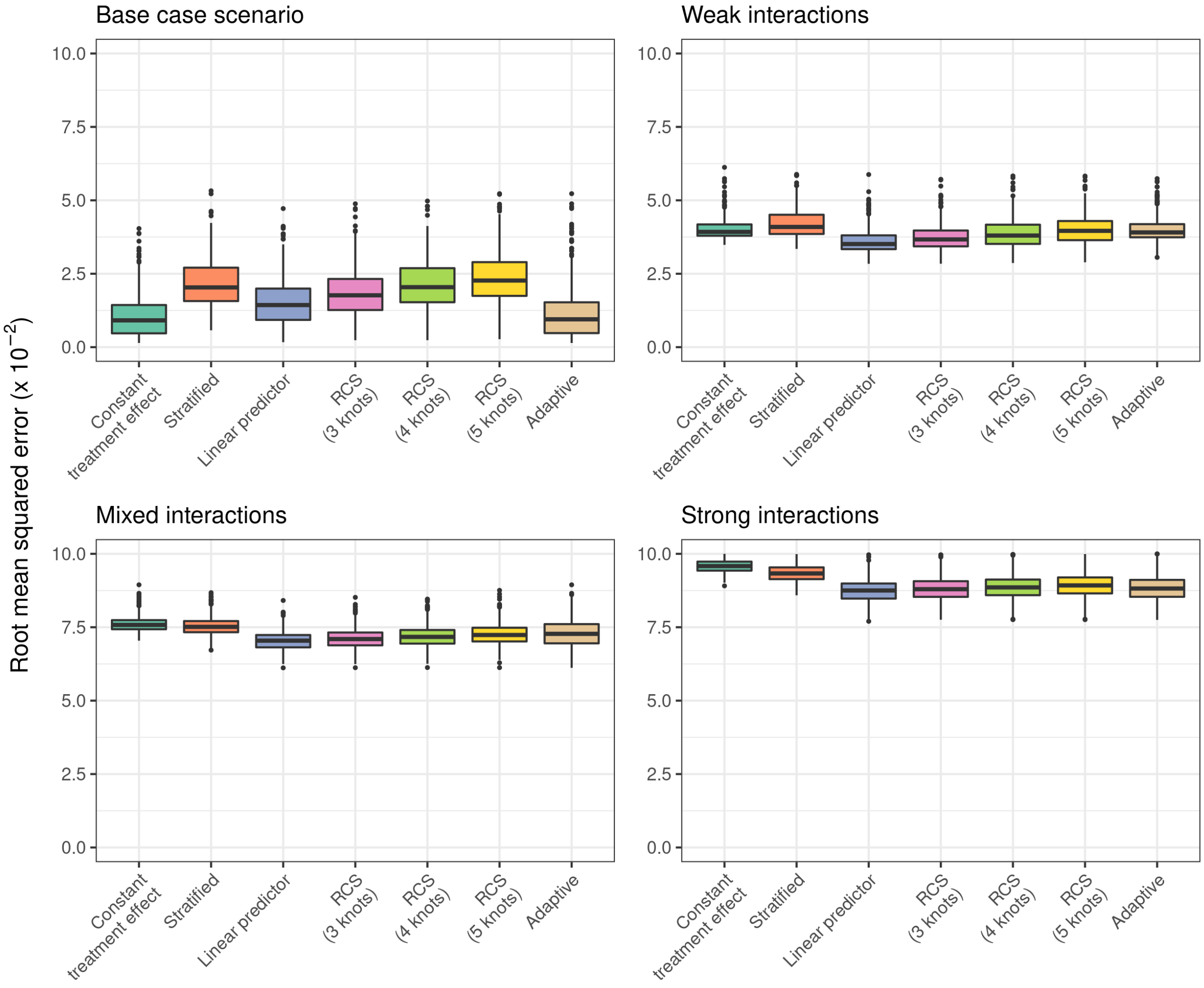

Figure 3: Root mean squared error of the different methods when focusing on the effect treatment-covariate interactions. In all scenarios sample size is fixed to 4250 and the true prediction AUC is 0.75. *Panel A* presents the results for the base case scenario, where no interactions are present. *Panel B* presents the results for the scenario with 4 weak interactions. *Panel C* presents the results for the scenario with mixed interactions (2 weak and 2 strong). *Panel D* presents the results for the scenario with 4 strong interactions.

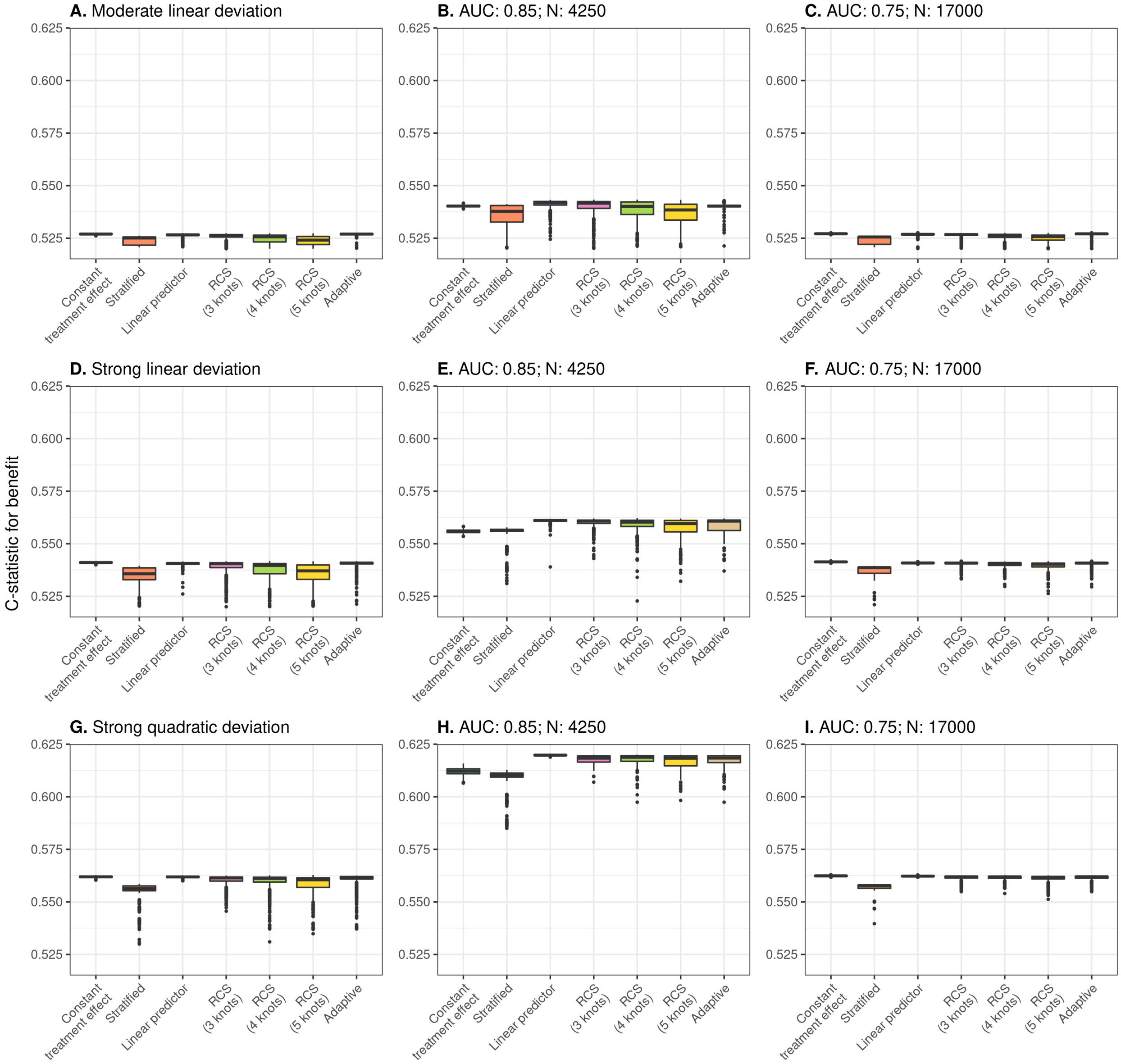

Figure 4: Discrimination for benefit of the different methods. *Panel A* presents the results when moderate linear deviations from the base case scenario of constant treatment effects are considered, sample size is 4250 and true prediction AUC is 0.75. *Panel B* presents the results when moderate linear deviations are considered and true prediction AUC is 0.85. Sample size is 4250. *Panel C* presents the results when moderate linear deviations are considered and the sample size is 0.75. *Panel D* presents the results when strong linear deviations from the base case are considered, sample size is 4250 and true prediction AUC is 0.85. *Panel E* presents results for true prediction AUC of 0.85 and *Panel F* for sample size of 17000 (assuming moderate linear deviations from base case). *Panel G* presents the results when strong quadratic deviations from the base case are considered, sample size is 4250 and true prediction AUC is 0.85. *Panel H* presents the results for true prediction AUC of 0.85 and *Panel I* for sample size of 17000 (assuming strong quadratic deviations from base case)

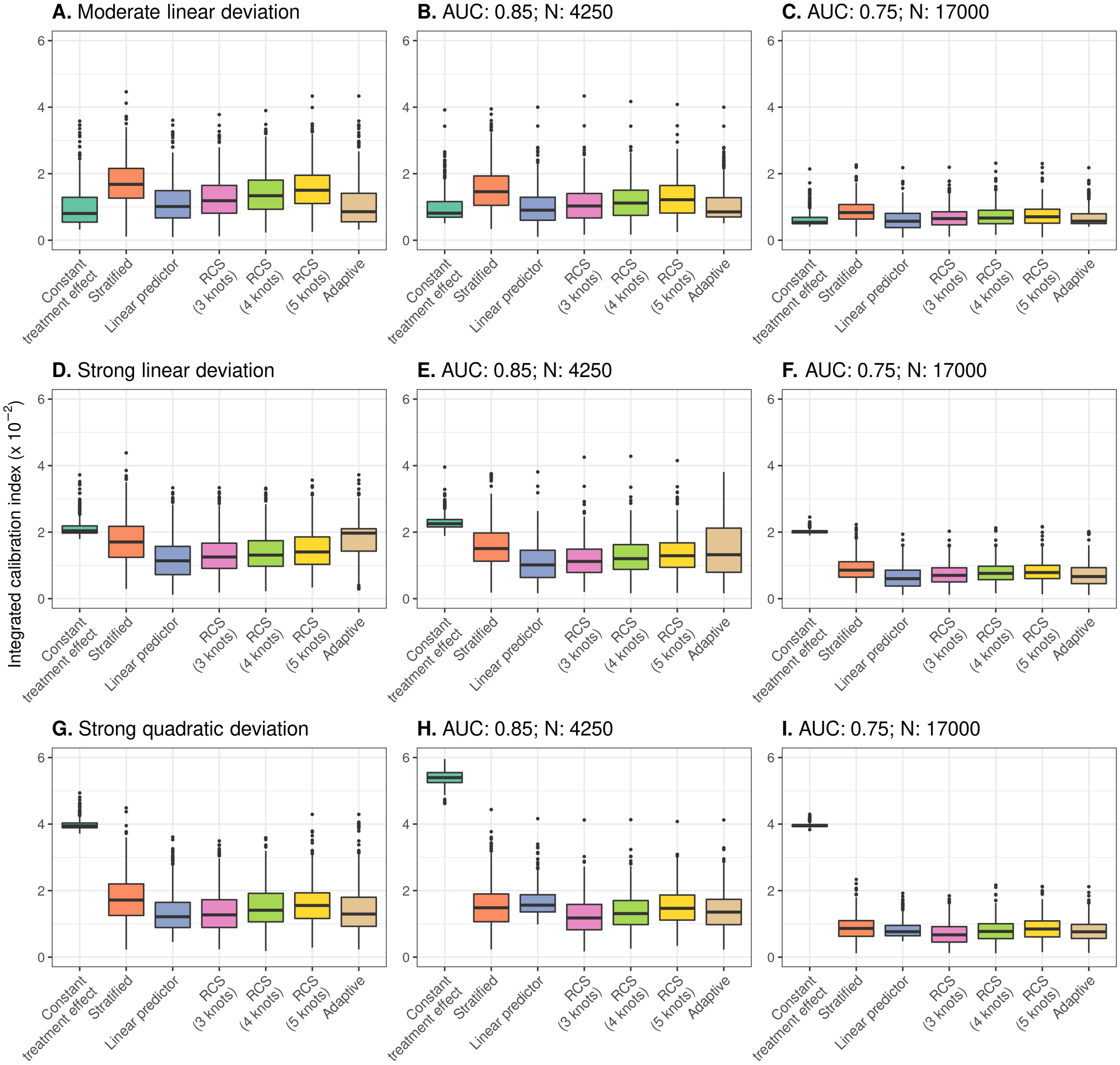

Figure 5: Calibration for benefit of the different methods. *Panel A* presents the results when moderate linear deviations from the base case scenario of constant treatment effects are considered, sample size is 4250 and true prediction AUC is 0.75. *Panel B* presents the results when moderate linear deviations are considered and true prediction AUC is 0.85. Sample size is 4250. *Panel C* presents the results when moderate linear deviations are considered and the sample size is 0.75. *Panel D* presents the results when strong linear deviations from the base case are considered, sample size is 4250 and true prediction AUC is 0.85. *Panel E* presents results for true prediction AUC of 0.85 and *Panel F* for sample size of 17000 (assuming moderate linear deviations from base case). *Panel G* presents the results when strong quadratic deviations from the base case are considered, sample size is 4250 and true prediction AUC is 0.85. *Panel H* presents the results for true prediction AUC of 0.85 and *Panel I* for sample size of 17000 (assuming strong quadratic deviations from base case)

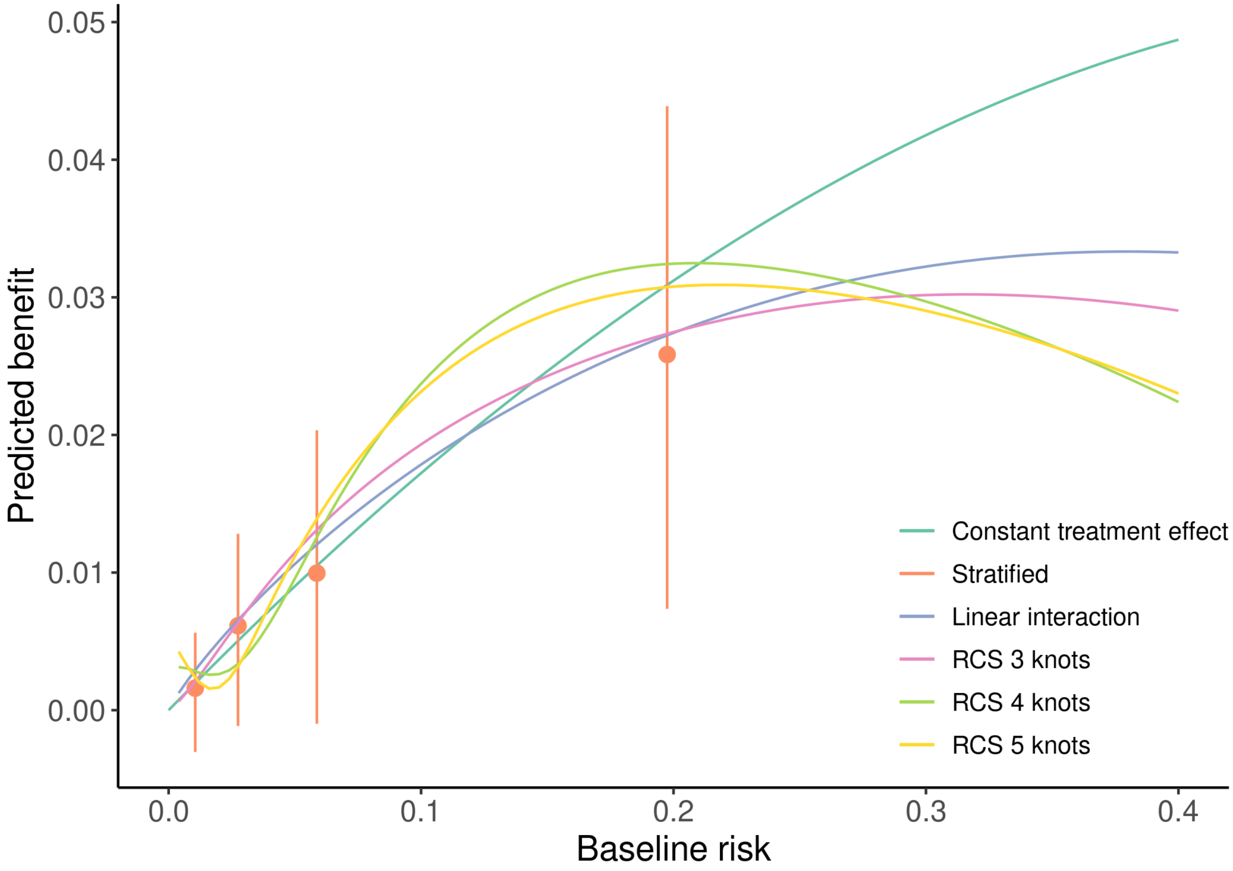

Figure 6: Caption

104 # 4 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.