

# Supplementary material

## Contents

<b>1 Notation</b>	<b>2</b>
<b>2 Simulation settings</b>	<b>2</b>
2.1 Base-case scenario . . . . .	3
2.2 Deviations from base-case . . . . .	3
<b>3 Plausible scenario settings</b>	<b>19</b>
<b>4 Approaches to individualize benefit predictions</b>	<b>19</b>
4.1 Risk modeling . . . . .	19
4.2 Risk stratification . . . . .	19
4.3 Constant treatment effect . . . . .	20
4.4 Linear interaction . . . . .	20
4.5 Restricted cubic splines . . . . .	20
<b>5 Adaptive model selection frequencies</b>	<b>22</b>
<b>6 Discrimination and calibration for benefit</b>	<b>25</b>
<b>7 Strong relative treatment effect</b>	<b>29</b>
<b>8 Treatment interactions</b>	<b>32</b>
<b>9 Empirical illustration</b>	<b>37</b>
<b>10 References</b>	<b>38</b>

## 1 Notation

We observe RCT data  $(Z, X, Y)$ , where for each patient  $Z_i = 0, 1$  is the treatment status,  $Y_i = 0, 1$  is the observed outcome and  $X_i$  is a set of covariates measured. Let  $\{Y_i(z), z = 0, 1\}$  denote the unobservable potential outcomes. We observe  $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ . We are interested in predicting the conditional average treatment effect (CATE),

$$\tau(x) = E\{Y(0) - Y(1) | X = x\}$$

Assuming that  $(Z, X, Y)$  is a random sample from the target population and that  $(Y(0), Y(1)) \perp\!\!\!\perp Z | X$ , as we are in the RCT setting, we can predict CATE from

$$\begin{aligned}\tau(x) &= E\{Y(0) | X = x\} - E\{Y(1) | X = x\} \\ &= E\{Y | X = x, Z = 0\} - E\{Y | X = x, Z = 1\}\end{aligned}$$

Based on an estimate of baseline risk

$$E\{Y | X = x, Z = 0\} = g(\hat{lp}(x))$$

with  $\hat{u} = \hat{lp}(x) = x^t \hat{\beta}$  the linear predictor and  $g$  the link function, we predict CATE from

$$\hat{\tau}(x) = g(f(\hat{u}, 0)) - g(f(\hat{u}, 1))$$

where  $f(u, z)$  describes interactions of the baseline risk linear predictor with treatment.

## 2 Simulation settings

For all patients we observe covariates  $X_1, \dots, X_8$ , of which 4 are continuous and 4 are binary. More specifically,

$$\begin{aligned}X_1, \dots, X_4 &\sim N(0, 1) \\ X_5, \dots, X_8 &\sim B(1, 0.2)\end{aligned}$$

We first, generate the binary outcomes  $Y$  for the untreated patients ( $Z = 0$ ), based on

$$P(Y = 1 | X = x, Z = 0) = g(\beta_0 + \beta_1 x_1 + \dots + \beta_8 x_8) = g(lp_0), \quad (1)$$

where

$$g(x) = \frac{e^x}{1 + e^x}$$

For treated patients, outcomes are generated from:

$$P(Y = 1 | X = x, Z = 1) = g(lp_1) \quad (2)$$

where

$$lp_1 = \gamma_2(lp_0 - c)^2 + \gamma_1(lp_0 - c) + \gamma_0$$

## 2.1 Base-case scenario

The base-case scenario assumes a constant odds ratio of 0.8 in favor of treatment. The simulated datasets are of size  $n = 4250$ , where treatment is allocated at random using a 50/50 split (80% power for the detection of an unadjusted OR of 0.8, assuming an event rate of 20% in the untreated arm). Outcome incidence in the untreated population is set at 20%. For the development of the prediction model we use the model defined in (1) including a constant treatment effect. When doing predictions,  $Z$  is set to 0. The value of the true  $\beta$  is such that the above prediction model has an AUC of 0.75.

The previously defined targets are achieved when  $\beta = (-2.08, 0.49, \dots, 0.49)^t$ . For the derivations in the treatment arm we use  $\gamma = (\log(0.8), 1, 0)^t$ .

## 2.2 Deviations from base-case

We deviate from the base-case scenario in two ways. First, we alter the overall target settings of sample size, overall treatment effect and prediction model AUC. In a second stage, we consider settings that violate the assumption of a constant relative treatment effect, using a model-based approach.

For the first part, we consider:

- Sample size:
  - $n = 1064$
  - $n = 17000$
- Overall treatment effect:
  - $OR = 0.5$
  - $OR = 1$
- Prediction performance:
  - $AUC = 0.65$
  - $AUC = 0.85$

We set the true risk model coefficients to be  $\beta = (-1.63, 0.26, \dots, 0.26)^t$  for  $AUC = 0.65$  and  $\beta = (-2.7, 0.82, \dots, 0.82)^t$  for  $AUC = 0.85$ . In both cases,  $\beta_0$  is selected so that an event rate of 20% is maintained in the control arm.

For the second part linear, quadratic and non-monotonic deviations from the assumption of constant relative effect are considered. We also consider different intensity levels of these deviations. Finally, constant absolute treatment-related harms are introduced, i.e. positive ( $0.25 \times$  true average benefit), strong positive ( $0.50 \times$  true average benefit) and negative ( $-0.25 \times$  true average benefit; i.e. constant absolute treatment-related benefit). In case of true absent treatment effects, treatment-related harms are set to 1%, 2% and -1% for positive, strong positive and negative setting, respectively. The settings for these deviations are defined in Table S1.





























627	high	1,063	0.75		strong-positive	-2.08	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	-0.084	2.035	0.210	0	0.079	0.040
628	high	1,063	0.75		negative	-2.08	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	-0.084	2.035	0.210	0	0.079	0.099
629	high	1,063	0.65		absent	-1.63	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.786	2.762	0.321	0	0.089	0.089
630	high	1,063	0.65		moderate-positive	-1.63	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.786	2.762	0.321	0	0.089	0.067
631	high	1,063	0.65		strong-positive	-1.63	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.786	2.762	0.321	0	0.089	0.044
632	high	1,063	0.65		negative	-1.63	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.786	2.762	0.321	0	0.089	0.111
633	high	1,063	0.85		absent	-2.70	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	-0.621	1.566	0.138	0	0.069	0.069
634	high	1,063	0.85		moderate-positive	-2.70	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	-0.621	1.566	0.138	0	0.069	0.052
635	high	1,063	0.85		strong-positive	-2.70	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	-0.621	1.566	0.138	0	0.069	0.034
636	high	1,063	0.85		negative	-2.70	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	-0.621	1.566	0.138	0	0.069	0.086
637	high	17,000	0.75		absent	-2.08	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	-0.084	2.035	0.210	0	0.079	0.079
638	high	17,000	0.75		moderate-positive	-2.08	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	-0.084	2.035	0.210	0	0.079	0.059
639	high	17,000	0.75		strong-positive	-2.08	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	-0.084	2.035	0.210	0	0.079	0.040
640	high	17,000	0.75		negative	-2.08	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	-0.084	2.035	0.210	0	0.079	0.099
641	high	17,000	0.65		absent	-1.63	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.786	2.762	0.321	0	0.089	0.089
642	high	17,000	0.65		moderate-positive	-1.63	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.786	2.762	0.321	0	0.089	0.067
643	high	17,000	0.65		strong-positive	-1.63	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.786	2.762	0.321	0	0.089	0.044
644	high	17,000	0.65		negative	-1.63	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.786	2.762	0.321	0	0.089	0.111
645	high	17,000	0.85		absent	-2.70	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	-0.621	1.566	0.138	0	0.069	0.069
646	high	17,000	0.85		moderate-positive	-2.70	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	-0.621	1.566	0.138	0	0.069	0.052
647	high	17,000	0.85		strong-positive	-2.70	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	-0.621	1.566	0.138	0	0.069	0.034
648	high	17,000	0.85		negative	-2.70	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	-0.621	1.566	0.138	0	0.069	0.086

### 3 Plausible scenario settings

In this section we present specific scenarios from our simulation settings in which evolution of benefit followed similar patterns to [1]. In this case patients were stratified into risk quarters based on their true baseline risk. Within each risk quarter we constructed boxplots of true benefit.

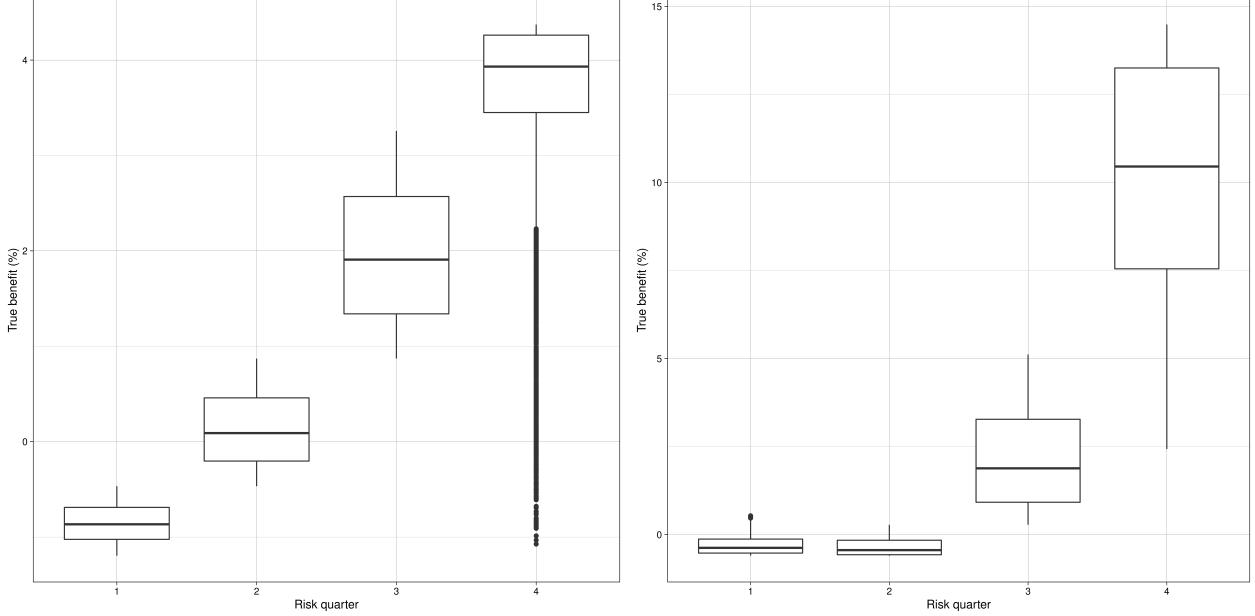


Figure S1: Simulation scenarios that closely follow trials. In this case, we see increasing absolute benefits with increasing baseline risk.

### 4 Approaches to individualize benefit predictions

#### 4.1 Risk modeling

Merging treatment arms, we develop prediction models including a constant relative treatment effect:

$$P(Y = 1 \mid X = x, Z = z) = g(x^t \beta + \gamma z) \quad (3)$$

We derive baseline risk predictions for patients by setting  $Z = 0$  in (4.1). All methods for individualizing benefit predictions are 2-stage methods, that start by fitting a model for predicting baseline risk. The estimated linear predictor of this model is

$$\hat{lp} = lp(x; \hat{\beta}) = x^t \hat{\beta}$$

#### 4.2 Risk stratification

Derive a prediction model using the same approach as above and divide the population in equally sized risk-based subgroups. Estimate subgroup-specific absolute benefit from the observed absolute differences.

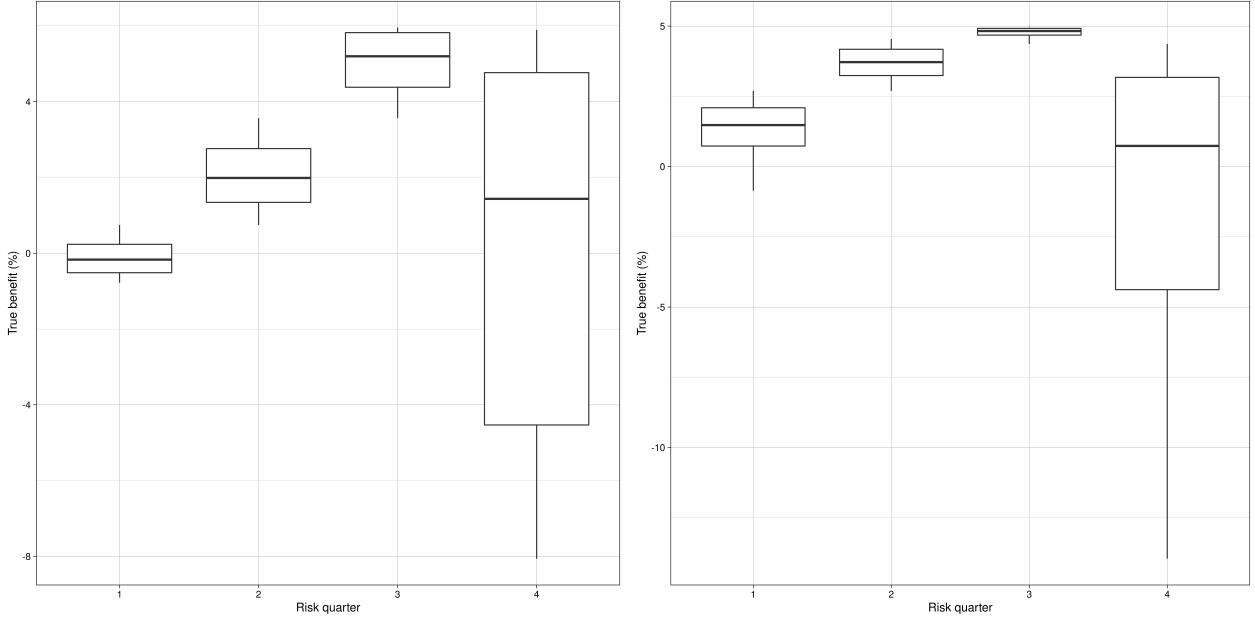


Figure S2: Simulation scenarios that closely follow trials. In this case, we see increasing absolute benefits with increasing baseline risk up to the third risk quarter. In the fourth risk quarter this trend is interrupted and benefits are diminished.

Subject-specific benefit predictions are made by attributing to individuals their corresponding subgroup-specific estimate.

### 4.3 Constant treatment effect

Assuming a constant relative treatment effect, fit the adjusted model in (4.1). Then, predict absolute benefit using

$$\hat{\tau}(x; \hat{\beta}, \hat{\gamma}) = g(f(\hat{lp}, 0)) - g(f(\hat{lp}, 1)), \quad (4)$$

where  $f(\hat{lp}, z) = \hat{lp} + \hat{\gamma}z$ , with  $\hat{\gamma}$  the estimated relative treatment effect (log odds ratio).

### 4.4 Linear interaction

We relax the assumption of a constant relative treatment effect in (4) by setting

$$f(\hat{lp}, z) = \gamma_0 + \gamma_1 z + \gamma_2 \hat{lp} + \gamma_3 z \hat{lp}$$

### 4.5 Restricted cubic splines

Finally, we drop the linearity assumption and predict absolute benefit using smoothing with restricted cubic splines with  $k = 3, 4$  and  $5$  knots. More specifically, we set:

$$f(\hat{lp}, z) = \gamma_0 + \gamma_1 z + zs(\hat{lp})$$

where

$$s(x) = \alpha_0 + \alpha_1 h_1(x) + \alpha_2 h_2(x) + \cdots + \alpha_{k-1} h_{k-1}(x)$$

with  $h_1(x) = x$  and for  $j = 2, \dots, k-2$

$$h_{j+1}(x) = (x - t_j)^3 - (x - t_{k-1})^3_+ \frac{t_k - t_j}{t_k - t_{k-1}} + (x - t_k)^3_+ \frac{t_{k-1} - t_j}{t_k - t_{k-1}}$$

where  $t_1, \dots, t_k$  are the selected knots [2].

## 5 Adaptive model selection frequencies

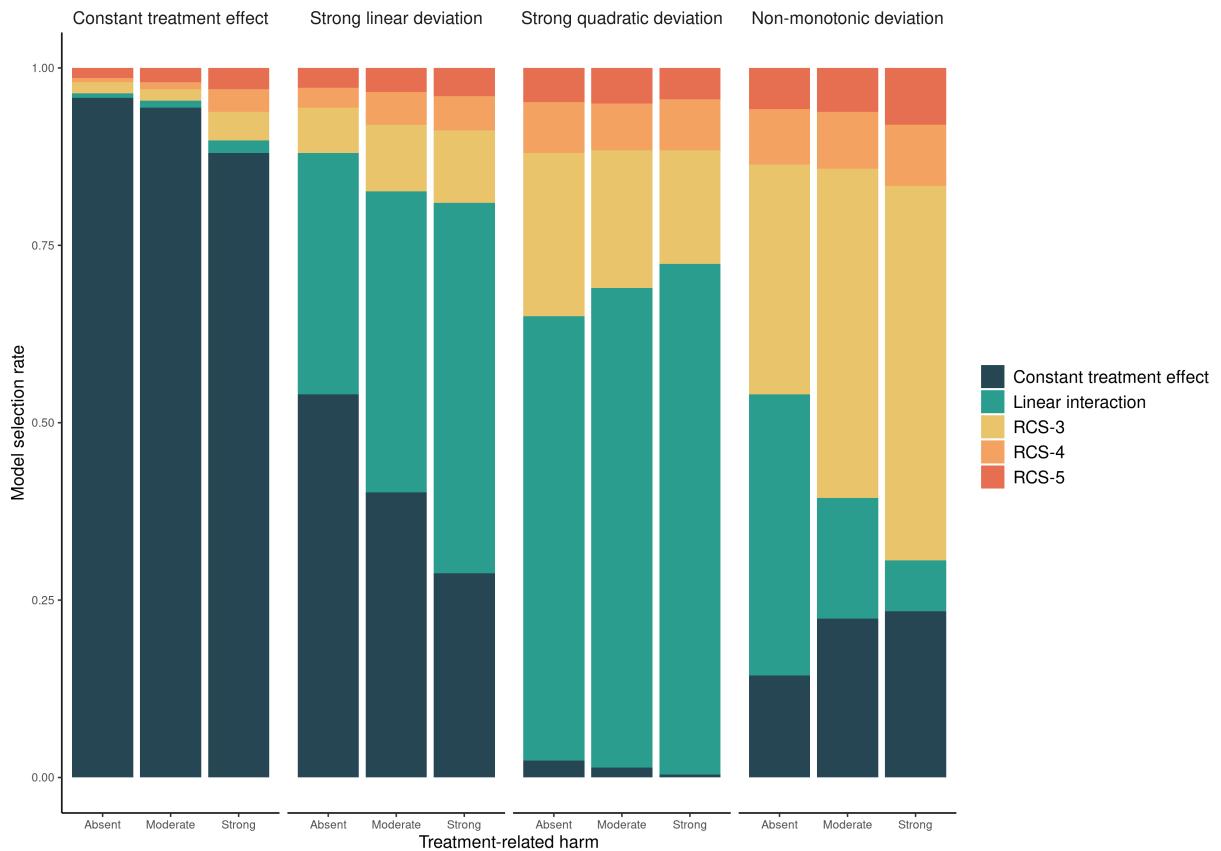


Figure S3: Model selection frequencies of the adaptive approach based on Akaike's Information Criterion across 500 replications. The scenario with the true constant relative treatment effect (first panel) had a true prediction AUC of 0.75 and sample size of 4,250.

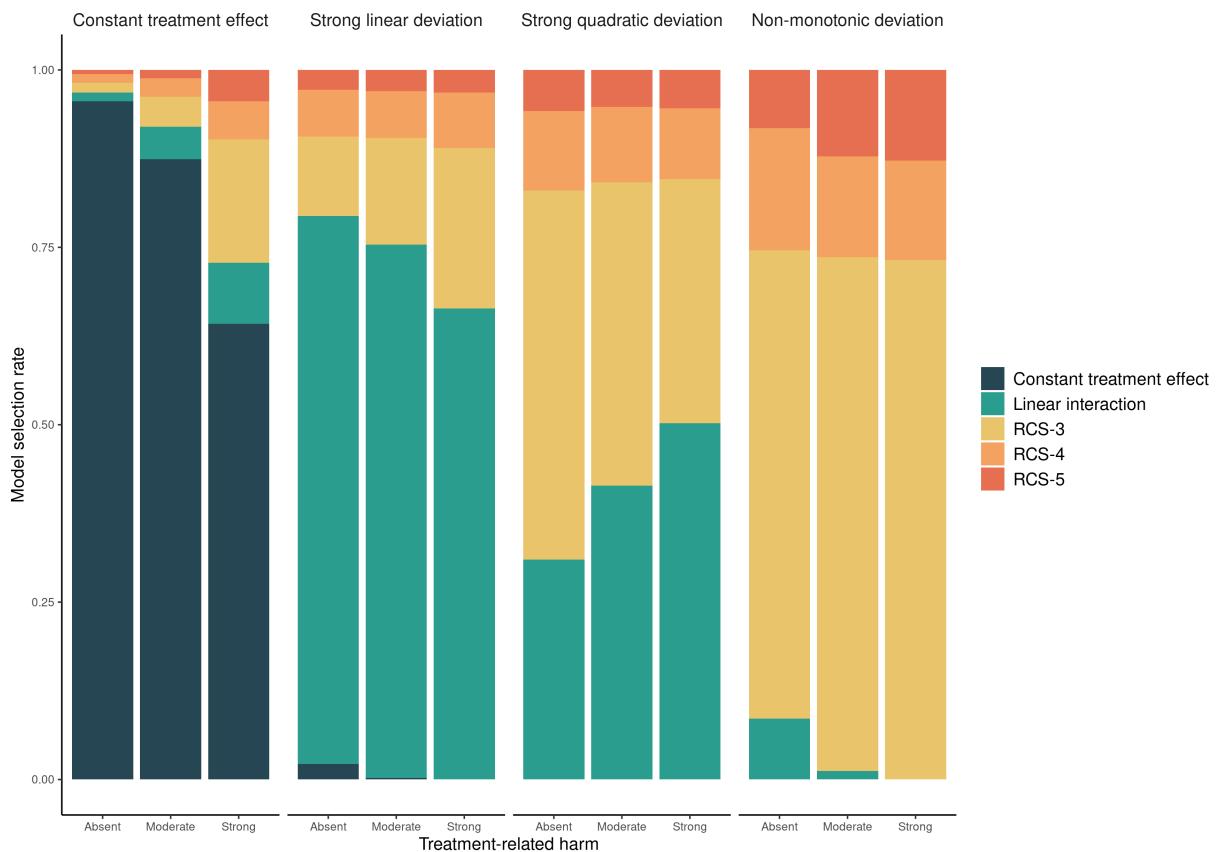


Figure S4: Model selection frequencies of the adaptive approach based on Akaike's Information Criterion across 500 replications. Sample size is 17,000 rather than 4,250 in Figure S3

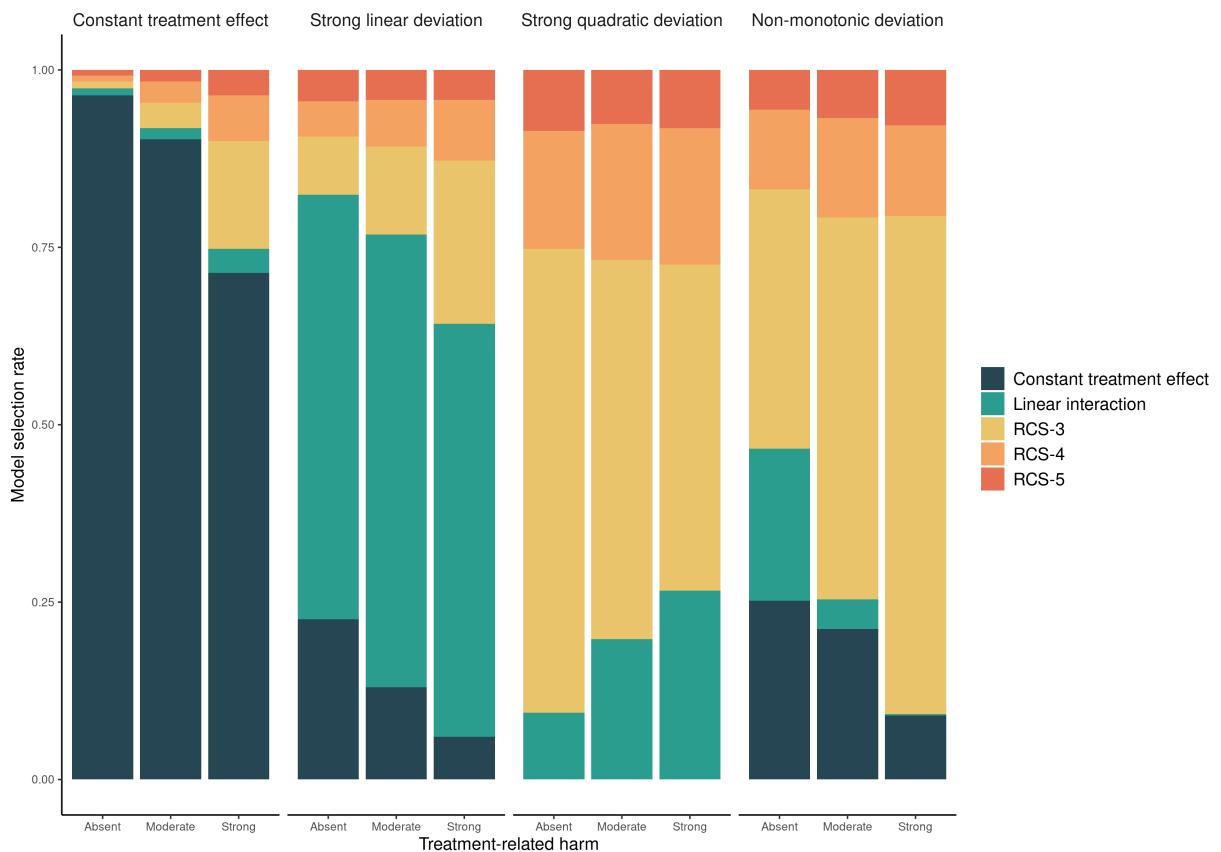


Figure S5: Model selection frequencies of the adaptive approach based on Akaike's Information Criterion across 500 replications. AUC is 0.85 rather than 0.75 in Figure S3

## 6 Discrimination and calibration for benefit

The c-for-benefit represents 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. To be able to calculate observed benefit, patients in each treatment arm are ranked based on their predicted benefit and then matched 1:1 across treatment arms. Observed treatment benefit is defined as the difference of observed outcomes between the untreated and the treated patient of each matched patient pair. Predicted benefit is defined as the average of predicted benefit within each matched patient pair.

We evaluated calibration in a similar manner, using the integrated calibration index (ICI) for benefit [3]. The observed benefits are regressed on the predicted benefits using a locally weighted scatterplot smoother (loess). The ICI-for-benefit is the average absolute difference between predicted and smooth observed benefit. Values closer to represent better calibration.

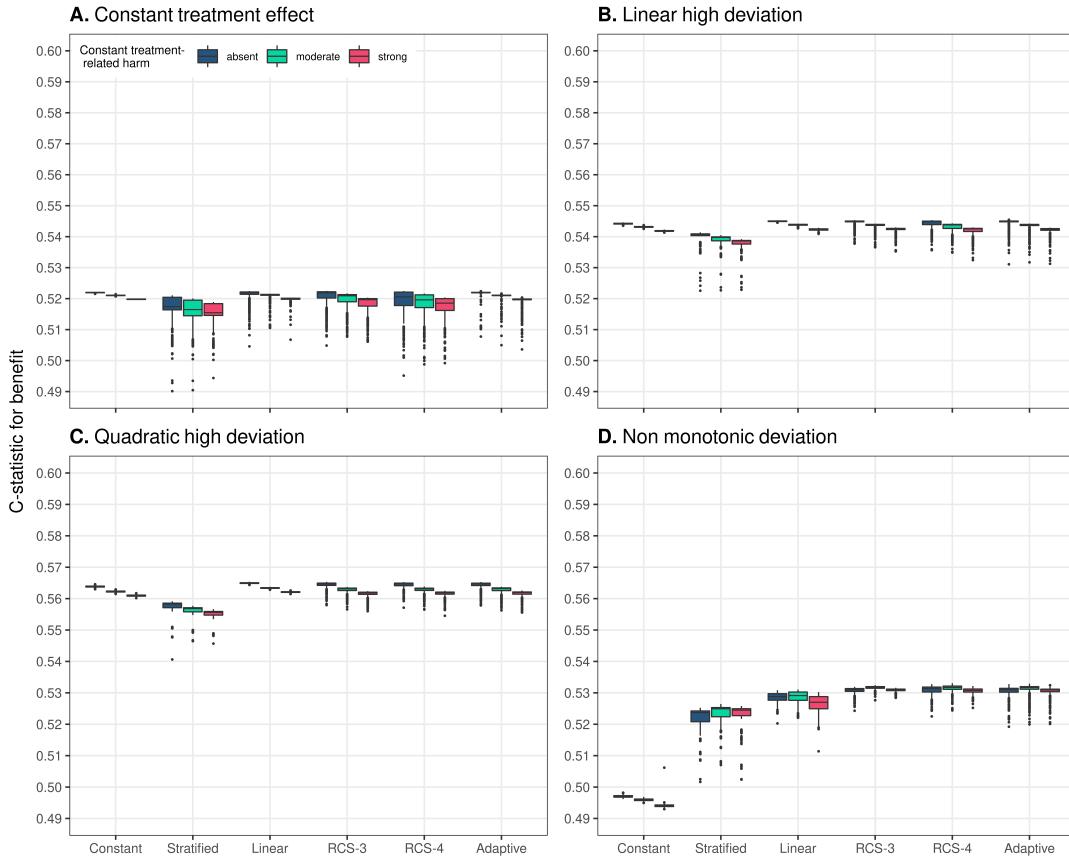


Figure S6: Discrimination for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.75 and sample size of 17,000

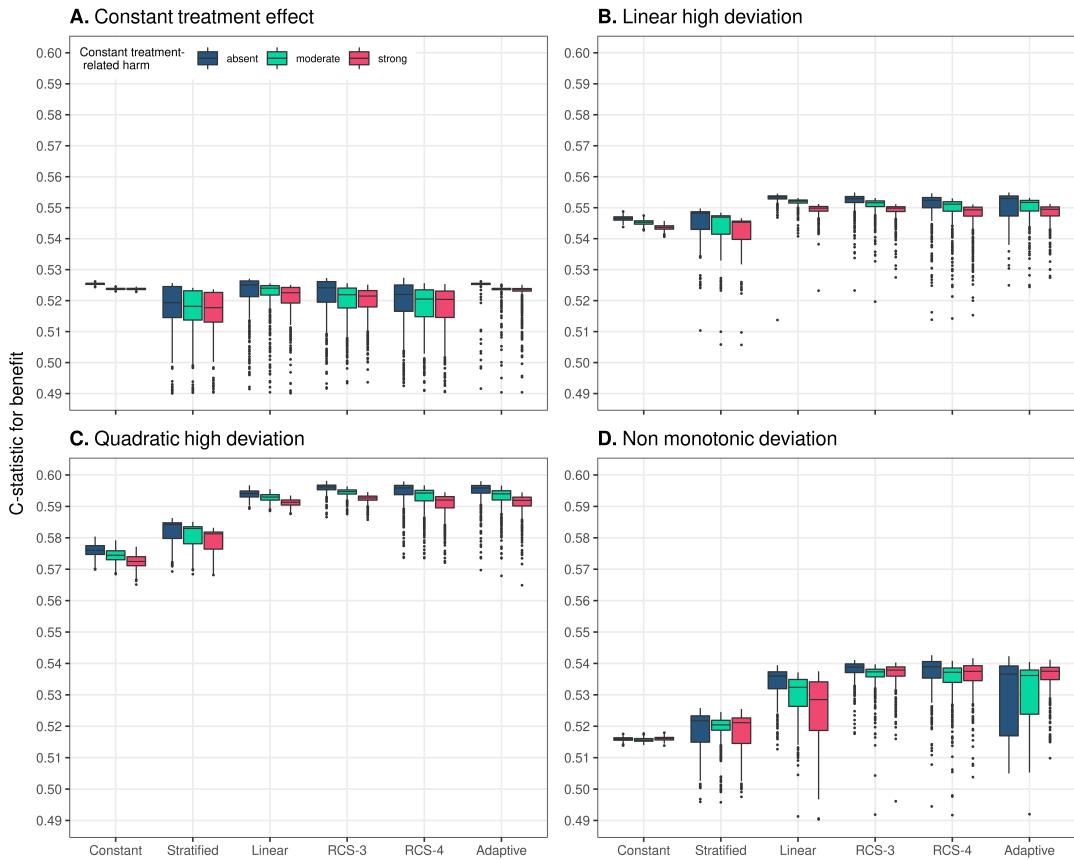


Figure S7: Discrimination for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.85 and sample size of 4,250

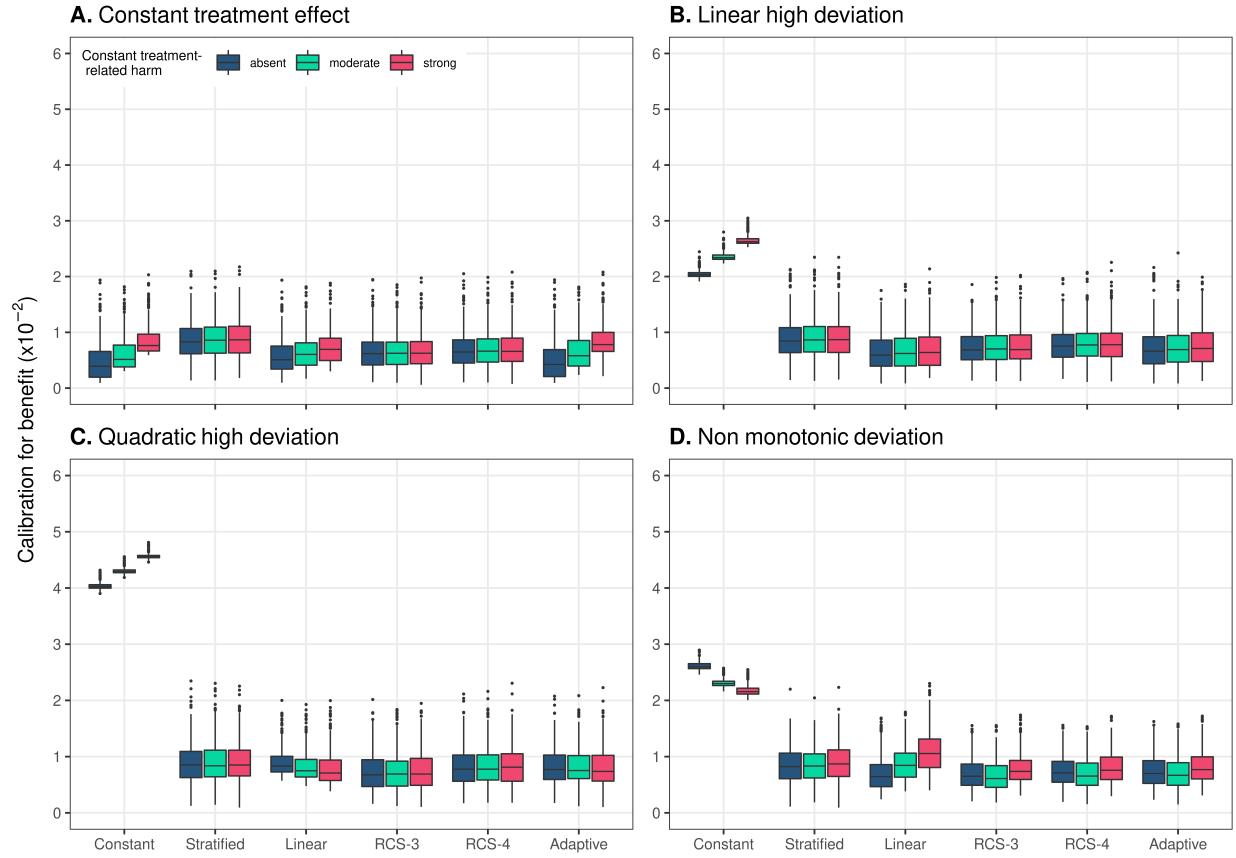


Figure S8: Calibration for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.75 and sample size of 17,000

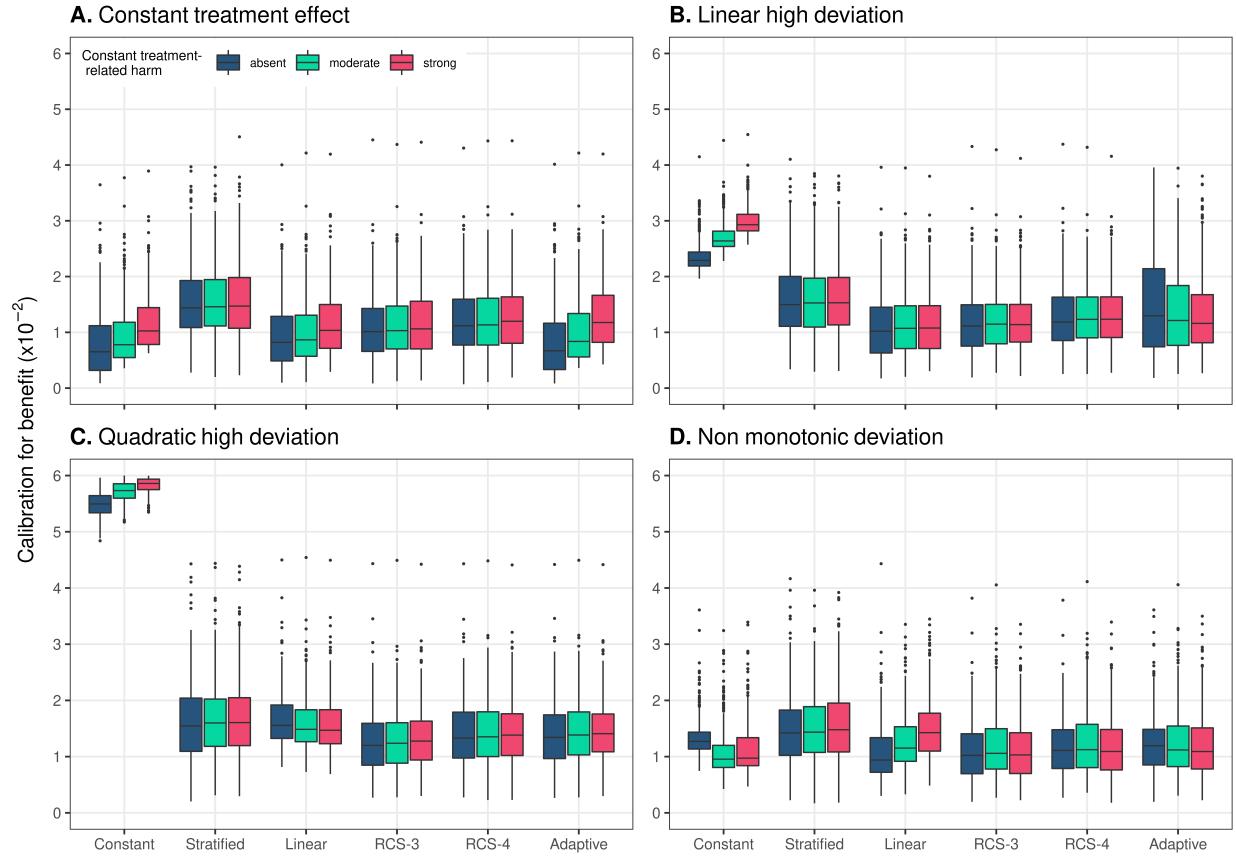


Figure S9: Calibration for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.85 and sample size of 4,250

## 7 Strong relative treatment effect

Here we present the root mean squared error of the considered methods using strong constant relative treatment effect ( $OR = 0.5$ ) as the reference. Again, the same sample size and prediction performance settings were considered along with the same settings for linear, quadratic and non-monotonic deviations from the base case scenario of constant relative treatment effects are considered. All results can be found at [https://arekkas.shinyapps.io/simulation\\_viewer/](https://arekkas.shinyapps.io/simulation_viewer/).

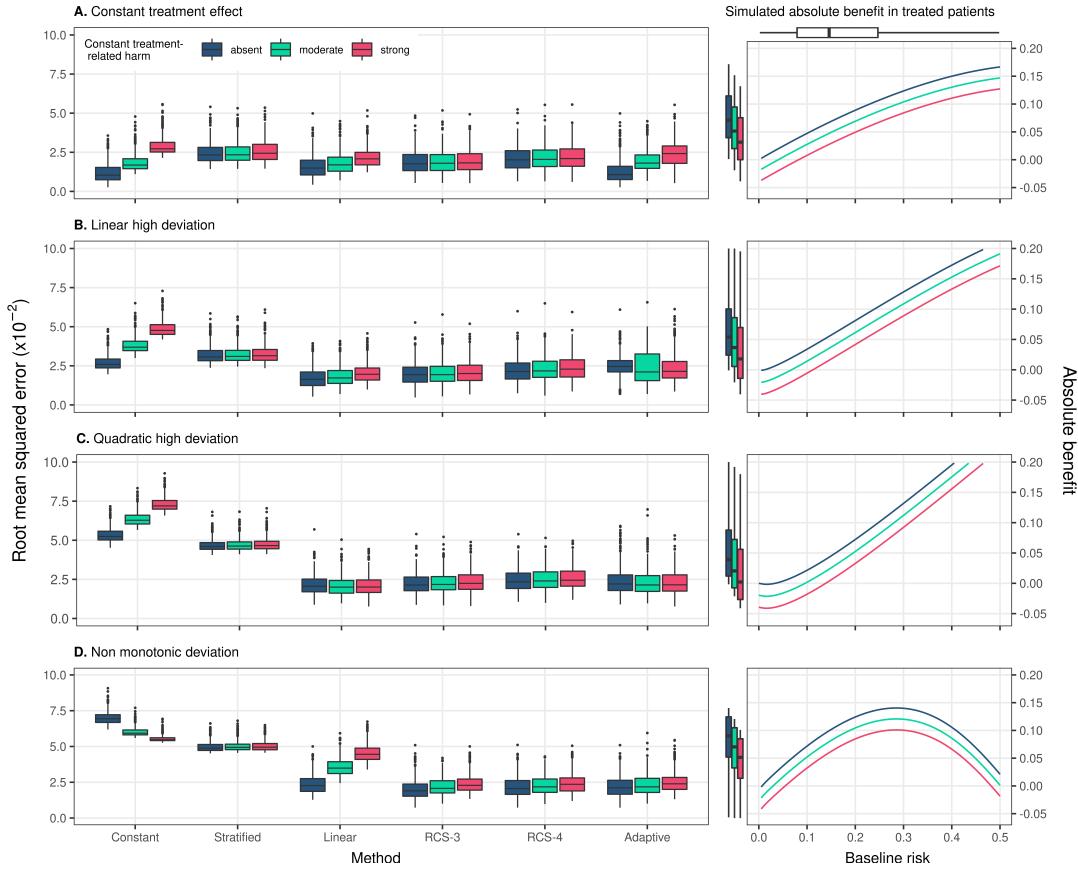


Figure S10: RMSE of the considered methods across 500 replications calculated in a simulated super-population of size 500,000. The scenario with true constant relative treatment effect (panel A) had a true prediction AUC of 0.75 and sample size of 4,250. The RMSE is also presented for strong linear (panel B), strong quadratic (panel C), and non-monotonic (panel D) deviations from constant relative treatment effects. Panels on the right side present the true relationship between baseline risk (x-axis) and absolute treatment benefit (y-axis). The 2.5, 25, 75 and 97.5 percentiles of the risk distribution are expressed in the boxplot.

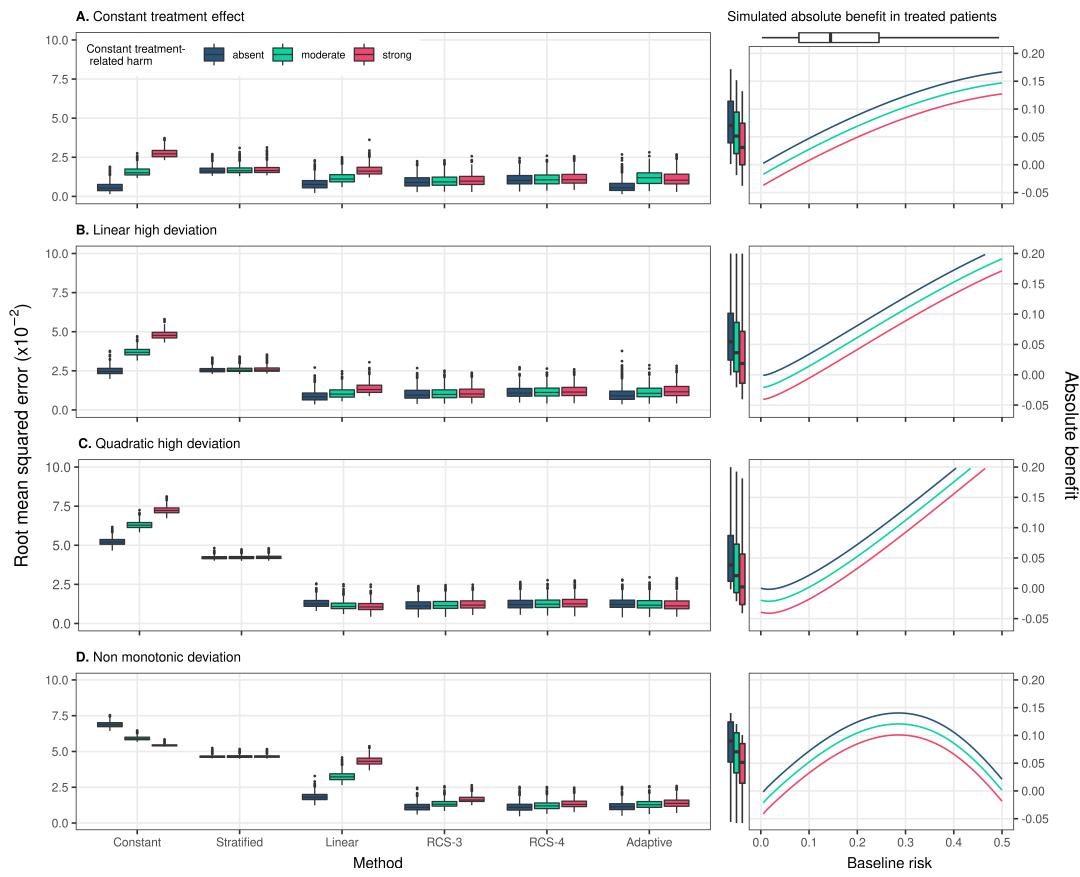


Figure S11: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000. Sample size is 17,000 rather than 4,250 in Figure S10.

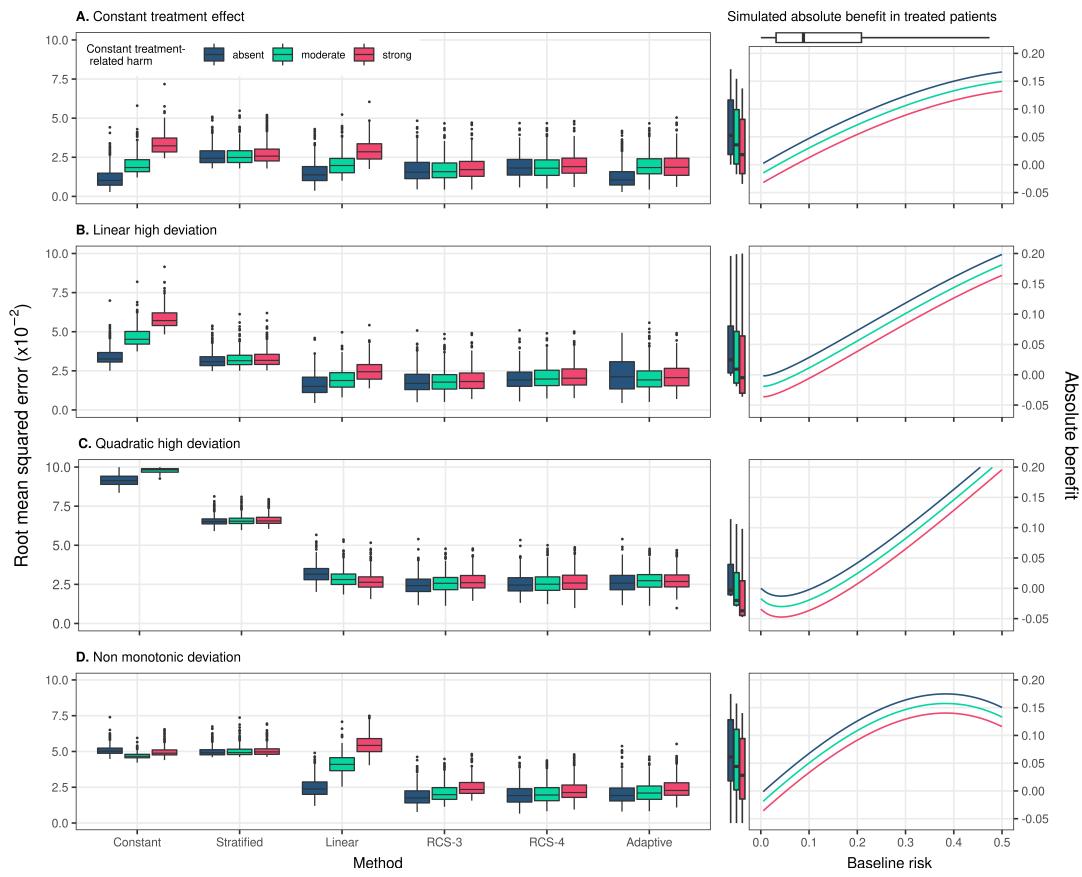


Figure S12: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000. AUC is 0.85 rather than in Figure S10.

## 8 Treatment interactions

We carried out a smaller set of simulations, in which we assumed true treatment-covariate interactions. Sample size was set to 4,250 and the AUC of the true prediction model was set to 0.75. The following scenarios were considered: 1) 4 true weak positive interactions ( $OR_{Z=1}/OR_{Z=0} = 0.83$ ); 2) 4 strong positive interactions ( $OR_{Z=1}/OR_{Z=0} = 0.61$ ); 3) 2 weak and 2 strong positive interactions; 4) 4 weak negative interactions ( $OR_{Z=1}/OR_{Z=0} = 1.17$ ); 5) 4 strong negative interactions ( $OR_{Z=1}/OR_{Z=0} = 1.39$ ); 6) 2 weak and 2 strong negative interactions; 7) combined positive and negative strong interactions. We also considered constant treatment-related harms applied on the absolute scale to all treated patients. The exact settings were: 1) absent treatment-related harms; 2) moderate treatment-related harms, defined as 25% of the average true benefit of the scenario without treatment-related harms; 3) strong treatment-related harms defined as 50% of the true average benefit of the scenario without treatment-related harms; 4) negative treatment-related harms (benefit), defined as an absolute risk reduction for treated patients of 50% of the true average benefit of the scenario without treatment-related harms. The exact settings can be found in Table S2.



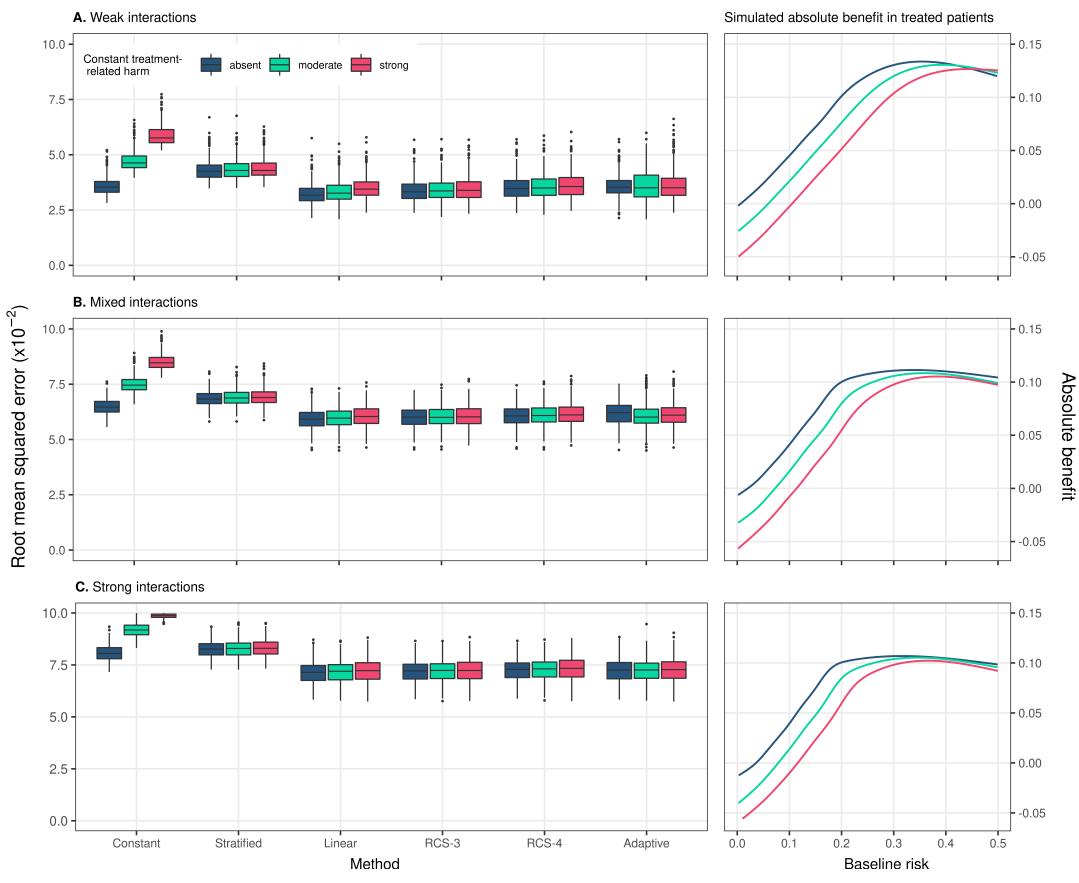


Figure S13: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000 where treatment-covariate interactions all favoring treatment were considered.

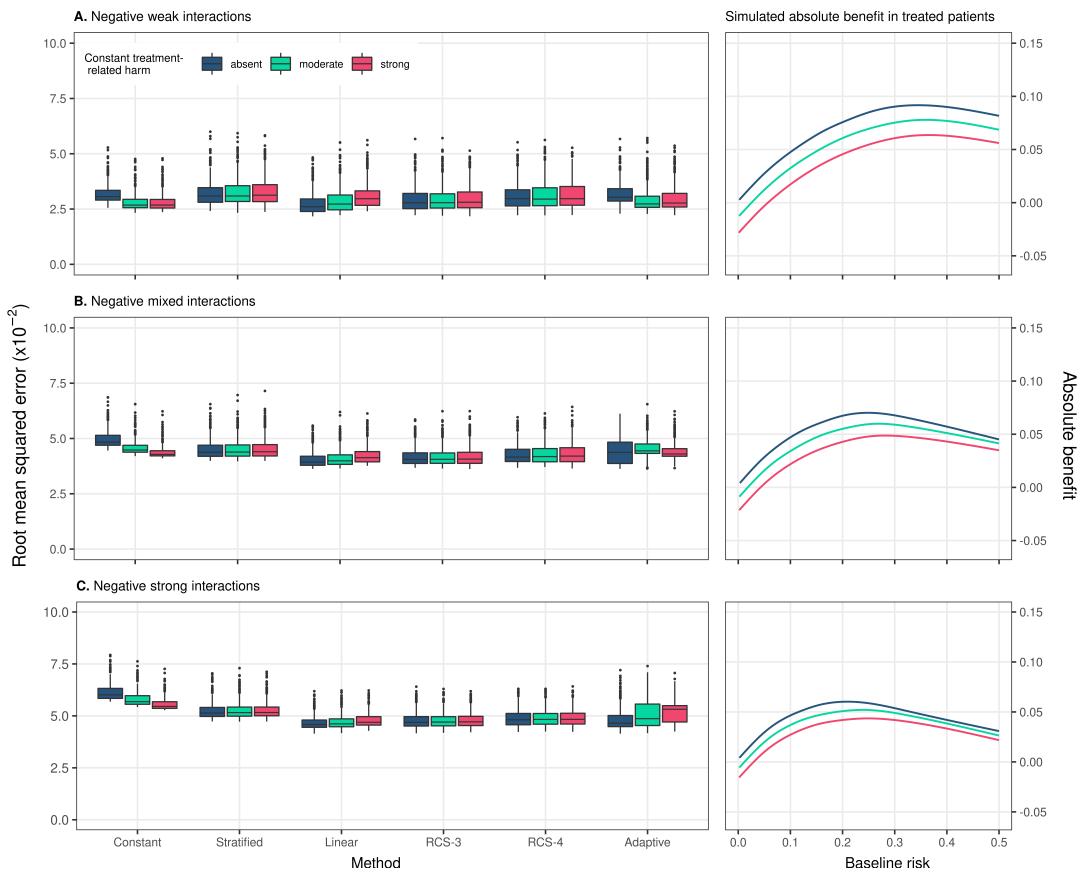


Figure S14: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000 where treatment-covariate interactions all favoring the control were considered.

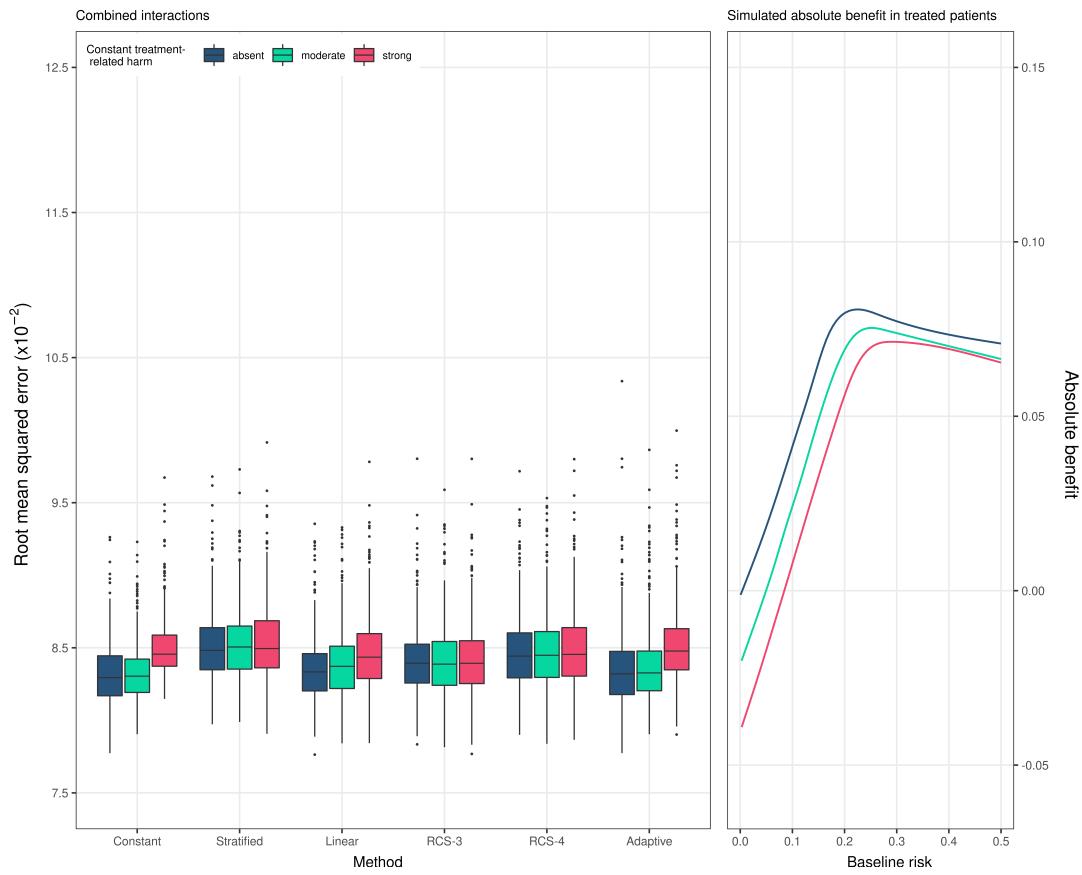


Figure S15: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000 where treatment-covariate interactions 2 favoring treatment and 2 favoring the control were considered.

## 9 Empirical illustration

For predicting baseline risk of 30-day mortality we fitted a logistic regression model with age, Killip class (*Killip*), systolic blood pressure (*sysbp*), pulse rate (*pulse*), prior myocardial infarction (*pmi*), location of myocardial infarction (*miloc*) and treatment as the covariates. Baseline predictions were made setting treatment to 0.

$$P(\text{outcome} = 1 | X = x) = \text{expit}(lp(x)),$$

where

$$\begin{aligned} lp(x) = & \beta_0 + \beta_1 \text{age} + \beta_2 I(\text{Killip} = II) + \beta_3 I(\text{Killip} = III) + \\ & \beta_4 I(\text{Killip} = IV) + \beta_5 \min(\text{sysbp}, 120) + \beta_6 \text{pulse} + \\ & \beta_7 \max(\text{pulse} - 50, 0) + \beta_8 I(\text{pmi} = yes) + \\ & \beta_9 I(\text{miloc} = \text{Anterior}) + \beta_{10} I(\text{miloc} = \text{Other}) + \\ & \gamma \times \text{treatment} \end{aligned}$$

and  $\text{expit}(x) = \frac{e^x}{1+e^x}$

Table S3: Coefficients of the prediction model for 30-day mortality, based on the data from GUSTO-I trial.

Variable	Estimate	stderror	zvalue	pvalue
Intercept	-3.020	0.797	-3.788	0.000
Age	-0.208	0.053	-3.935	0.000
Killip class = II	0.077	0.002	31.280	0.000
Killip class = III	0.614	0.059	10.423	0.000
Killip class = IV	1.161	0.121	9.566	0.000
Systolic blood pressure	1.921	0.162	11.872	0.000
Pulse rate (1)	-0.039	0.002	-20.332	0.000
Pulse rate (2)	-0.024	0.016	-1.521	0.128
Previous MI (yes)	0.043	0.016	2.675	0.007
MI location (Other)	0.447	0.056	7.964	0.000
MI location (Anterior)	0.286	0.135	2.126	0.033
Treatment	0.543	0.051	10.625	0.000

## 10 References

- [1] Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: Re-analysis of individual participant data from 32 large clinical trials. International Journal of Epidemiology 2016;dyw118. <https://doi.org/10.1093/ije/dyw118>.
- [2] Harrell FE. Regression modeling strategies. vol. 330. Springer; 2017.
- [3] 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>.