

Individualized treatment effect was predicted best by modeling baseline risk in interaction with treatment assignment

Alexandros Rekkas^a, Peter R. Rijnbeek^a, David M. Kent^b, Ewout W. Steyerberg^c, David van Klaveren^d

^a*Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands*

^b*Predictive Analytics and Comparative Effectiveness Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, USA*

^c*Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands*

^d*Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands*

Abstract

Objective: To compare different risk-based methods for optimal prediction of individualized treatment effects from RCTs. **Study Design and Setting:** We simulated RCT data using diverse assumptions for the average treatment effect, a baseline prognostic index of risk (PI), the shape of its interaction with treatment (none, linear, quadratic or non-monotonic) and the magnitude of treatment-related harms (none or constant independent of the PI). In each sample we predicted absolute benefit using: models with a constant relative treatment effect; stratification in quarters of the PI; models including a linear interaction of treatment with the PI; models including an interaction of treatment with a restricted cubic spline (RCS) transformation of the PI; an adaptive approach using Akaike's Information Criterion. We evaluated predictive performance using root mean squared error and measures of discrimination and calibration for benefit. **Results:** The linear-interaction model and the RCS-interaction model outperformed the constant treatment effect model in many simulation scenarios. The RCS-model was optimal when quadratic or non-monotonic deviations from a constant treatment effect were stronger, and when sample size was larger. Larger sample size also supported the adaptive approach. Illustrations in the GUSTO-I trial confirmed these findings. **Conclusion:** An interaction between baseline risk and treatment assignment should be considered to improve treatment effect predictions. Non-linear RCS interactions should be considered only in larger sample sizes.

Keywords: treatment effect heterogeneity absolute benefit prediction models

1. Introduction

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]. First, risk-modeling approaches use prediction of baseline risk as the reference; second, treatment effect modeling approaches also

7 model treatment-covariate interactions, in addition to risk factors; third, optimal treatment regime approaches
8 focus on developing treatment assignment rules and rely heavily on modeling treatment effect modifiers. A key
9 difference between these approaches is their parsimony in dealing the treatment effect modifiers, with no interaction
10 considered (risk modeling), a limited number of interactions (effect modeling), or a larger set of interactions
11 (optimal treatment regime approaches).

12 Risk-modeling approaches to predictive HTE analyses provide a viable option in the absence of well-established
13 treatment effect modifiers [3,4]. In simulations, modeling of effect modifiers, i.e. treatment-covariate interactions,
14 often led to miscalibrated predictions of benefit, while risk-based methods proved quite robust [5]. Most often,
15 risk-modeling approaches are carried out in two steps: first a risk prediction model is developed externally or
16 internally on the entire RCT population, “blinded” to treatment; then the RCT population is stratified using this
17 prediction model to evaluate risk-based treatment effect variation [6]. This two-step approach identified substantial
18 absolute treatment effect differences between low-risk and high-risk patients in a re-analysis of 32 large trials [7].
19 However, even though estimates at the risk subgroup level may be accurate, these estimates may need further
20 refinement for individual patients, especially for patients with predicted risk at the boundaries of the risk intervals.
21 Hence, the risk-stratified approach is useful for exploring and presenting HTE, but is not sufficient for supporting
22 treatment decisions for individual patients.

23 To individualize treatment effects, the recent PATH statement suggested various risk-based models including
24 a prognostic index of baseline risk (PI) and treatment assignment [3,4]. We aimed to summarize and compare
25 different risk-based models for predicting individualized treatment effects. We simulated RCT settings to compare
26 the performance of these models under different assumptions of the relation between baseline risk and treatment.
27 We illustrated the different models by a case study of predicting individualized effects of treatments for acute
28 myocardial infarction (MI) in a large randomized controlled trial (RCT).

29 **2. Methods**

30 *2.1. Simulation scenarios*

31 We simulated a typical RCT that is undertaken to compare a binary outcome (e.g. death) between a group of
32 patients in the treatment arm and a group of untreated patients in the control arm. For each patient we generated
33 8 baseline covariates $x_1, \dots, x_4 \sim N(0, 1)$ and $x_5, \dots, x_8 \sim B(1, 0.2)$. Treatment was allocated using a 50:50
34 split. Outcomes for patients in the control arm were generated from a logistic regression model including all
35 baseline covariates. In the base scenarios coefficient values were such, that the AUC of the logistic regression
36 model was 0.75 and the event rate in the control arm was 20%. Binary outcomes in the control arm were generated
37 from Bernoulli variables with true probabilities $P(y = 1|X, t_x = 0) = \text{expit}(PI) = \frac{e^{PI}}{1+e^{PI}}$.

Outcomes in the treatment arm were generated using 3 base scenarios: absent treatment effect ($OR = 1$), moderate treatment effect ($OR = 0.8$) and strong treatment effect ($OR = 0.5$). We started with simulating outcomes based on true constant relative treatment effects for the 3 base scenarios. We then simulated linear, quadratic and non-monotonic deviations from constant treatment effects using:

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

38 where lp_1 is the true linear predictor in the treatment arm, so that $P(y = 1|X, t_x = 1) = \text{expit}(lp_1)$. Finally, we
39 simulated scenarios where a constant absolute harm is applied across all treated patients. In this case we have
40 $P(y = 1|X, t_x = 1) = \text{expit}(lp_1) + \text{harm}$.

41 The sample size for the base scenarios was set to 4,250, since this sample size provides 80% power for the
42 detection of a marginal OR of 0.8 with the standard alpha of 0.5%. We evaluated the effect of smaller or larger
43 sample sizes of 1,063 (4,250 divided by 4) and 17,000 (4250 multiplied by 4), respectively. We also evaluated the
44 effect of worse or better discriminative ability for risk, adjusting the baseline covariate coefficients, such that the
45 AUC of the regression model in the control arm was 0.65 and 0.85 respectively.

46 Combining all these settings resulted in a simulation study of 648 scenarios (exact settings in the supplementary
47 material). With these scenarios we were able to cover the observed treatment effect heterogeneity in 32 large trials
48 as well as many other potential variations of risk-based treatment effect [7].

49 *2.2. Individualized risk-based benefit predictions*

50 All risk-based methods assume that a risk prediction model is available to assign risk predictions to individual
51 patients. For the simulations we developed a prediction model internally on the entire population, using a logistic
52 regression model with main effects for all baseline covariates and treatment assignment. Risk predictions for
53 individual patients were derived by setting treatment assignment to 0. Another common approach is to derive the
54 prediction model solely on the control patients, however this approach has been shown to lead to biased benefit
55 predictions [5,8,9].

56 A *stratified HTE method* has been suggested as an alternative to traditional subgroup analyses. Patients are
57 stratified into equally-sized risk strata—in this case based on risk quartiles. Absolute treatment effects within risk
58 strata are estimated by the difference in event rate between patients in the control arm and patients in the treated
59 arm. We considered this approach as a reference, expecting it to perform worse than the other candidates, as its
60 objective is to provide an illustration of HTE rather than to optimize individualized benefit predictions.

61 Second, we considered a model which assumes *constant relative treatment effect* (constant odds ratio). Hence,
62 absolute benefit is predicted from $\hat{\pi}(\mathbf{x}) = \text{expit}(PI + \log(OR))$.

63 Third, we considered a logistic regression model including treatment, the prognostic index, and their linear
64 interaction. Absolute benefit is then estimated from $\hat{\tau}(\mathbf{x}) = \text{expit}(\beta_0 + \beta_{PI}PI) - \text{expit}(\beta_0 + \beta_{tx} + (\beta_{PI} + \beta_*)PI)$.
65 We will refer to this method as the *linear interaction* approach.

66 Fourth, we used *restricted cubic splines* (RCS) to relax the linearity assumption on the effect of the linear
67 predictor [10]. We considered splines with 3 (RCS-3), 4 (RCS-4) and 5 (RCS-5) knots to compare models with
68 different levels of flexibility.

69 Finally, we considered an *adaptive approach* using Akaike's Information Criterion (AIC) for model selection.
70 More specifically, for the adaptive approach we ranked the constant relative treatment effect model, the linear
71 interaction model, and the RCS models with 3, 4, and 5 knots based on their AIC and selected the one with
72 the lowest value. The extra degrees of freedom were 1 (linear interaction), 2, 3 and 4 (RCS models) for these
73 increasingly complex interactions with the treatment effect.

74 2.3. Evaluation metrics

75 We evaluated the predictive accuracy of the considered methods by the root mean squared error (RMSE):

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

76 We compared the discriminative ability of the methods under study using c-for-benefit [11]. The c-for-benefit
77 represents the probability that from two randomly chosen matched patient pairs with unequal observed benefit,
78 the pair with greater observed benefit also has a higher predicted benefit. To be able to calculate observed benefit,
79 patients in each treatment arm are ranked based on their predicted benefit and then matched 1:1 across treatment
80 arms. *Observed* treatment benefit is defined as the difference of observed outcomes between the untreated and
81 the treated patient of each matched patient pair. *Predicted* benefit is defined as the average of predicted benefit
82 within each matched patient pair.

83 We evaluated calibration in a similar manner, using the integrated calibration index (ICI) for benefit [12]. The
84 observed benefits are regressed on the predicted benefits using a locally weighted scatterplot smoother (loess).
85 The ICI-for-benefit is the average absolute difference between predicted and smooth observed benefit. Values
86 closer to 0 represent better calibration.

87 2.4. Empirical illustration

88 We demonstrate the different methods for individualizing treatment benefits using data from 30,510 patients
89 with acute myocardial infarction (MI) included in the GUSTO-I trial. 10,348 patients were randomized to tissue
90 plasminogen activator (tPA) treatment and 20,162 were randomized to streptokinase. The outcome of interest
91 was 30-day mortality (total of 2,128 events), recorded for all patients.

92 **3. Results**

93 *3.1. Simulations*

94 The linear interaction model outperformed all RCS methods in terms of RMSE in scenarios with true constant
95 relative treatment effect ($OR = 0.8$, $N = 4,250$ and $AUC = 0.75$), strong linear and even strong quadratic deviations
96 from a constant relative treatment effect (Figure 1; panels A-C). However, with non-monotonic deviations from a
97 constant relative treatment effect, the RMSE of the linear interaction model increased substantially, especially in the
98 presence of treatment-related harms (Figure 1; panel D). In these scenarios, RCS-3 outperformed all other methods
99 in terms of RMSE. As might be expected the constant treatment effect approach had overall best performance
100 under true constant treatment effect settings. It was sensitive to all considered deviations, resulting in increased
101 RMSE. Finally, the adaptive approach had comparable performance to the best-performing method in each scenario.
102 However, in comparison with the best-performing approach, its RMSE was more variable in the scenarios with
103 linear and non-monotonic deviations, especially when also including moderate or strong treatment-related harms.
104 On closer inspection, we found that this behavior was caused by wrongly selecting the constant treatment effect
105 model in a substantial proportion of the replications (Supplement, Figure S1). This problematic behavior was less
106 with larger sample sizes (see below).

107 Increasing the sample size to 17,000 favored RCS-3 the most, It achieved lowest or close to lowest RMSE
108 across all scenarios (Figure 2). Especially in cases of strong quadratic and non-monotonic deviations RCS-3
109 had lower RMSE (median 0.011 for strong quadratic deviations and 0.010 for non-monotonic deviations with no
110 treatment-related harms) compared to the linear interaction approach (median 0.013 and 0.014, respectively),
111 regardless of the strength of treatment-related harms. Due to the large sample size, the RMSE of the adaptive
112 approach was even more similar to the best-performing method, and the constant relative treatment effect model
113 was less often wrongly selected (Supplement, Figure S2).

114 When we increased the AUC of the true prediction model to 0.85 ($OR = 0.8$ and $N = 4,250$). RCS-3 had the
115 lowest RMSE in the case of strong quadratic or non-monotonic deviations and very comparable performance to
116 the – optimal – linear interaction model in the case of strong linear deviations (median RMSE 0.016 for RCS-3
117 compared to 0.014 for the linear interaction model). As observed in the base case scenario the adaptive approach
118 wrongly selected the constant treatment effect model (23% and 25% of the replications in the strong linear and
119 non-monotonic deviation scenarios without treatment-related harms, respectively), leading to more variability of
120 the RMSE (Supplement, Figure S3).

121 In comparison with the true approach, dsicrimination for benefit in the scenario with a constant relative
122 treatment effect was only slightly lower for the linear interaction model, but substantially lower for the non-linear
123 RCS approaches (Figure 4; panel A). With strong linear or quadratic deviations from a constant relative treatment

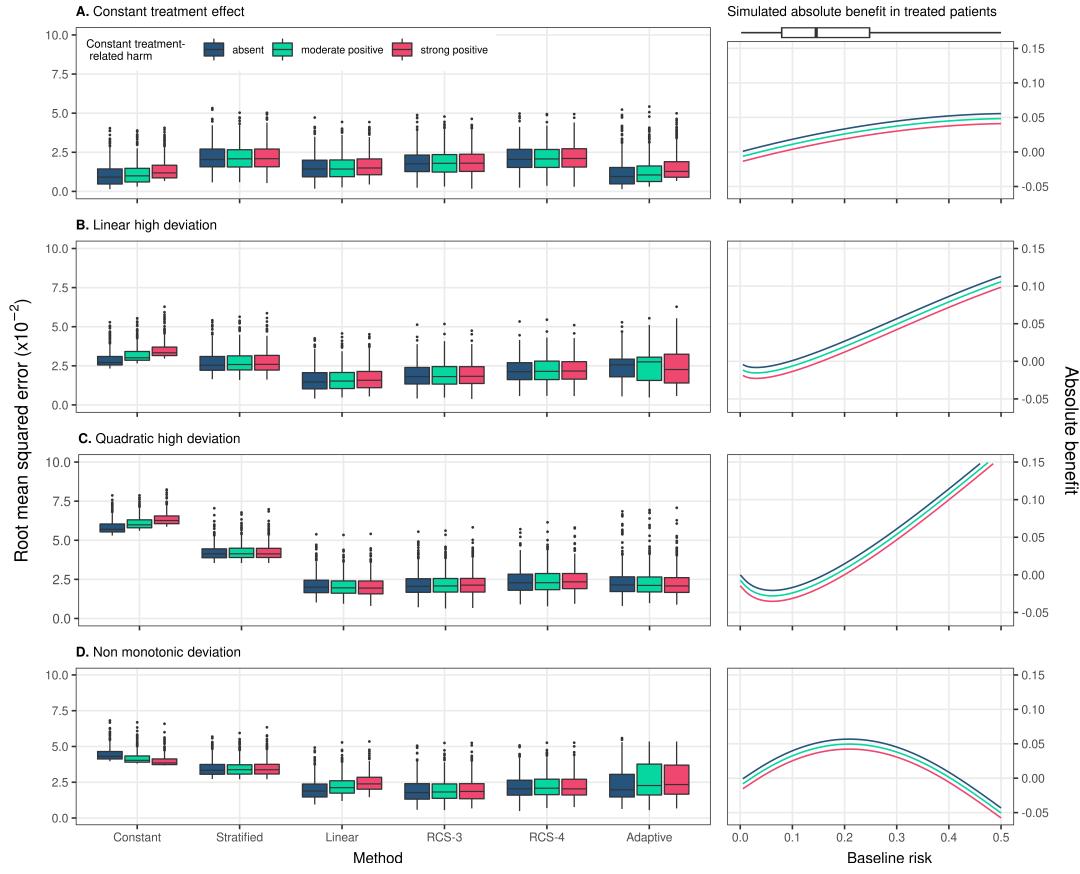


Figure 1: RMSE of the considered methods across 500 replications calculated from a simulated super-population of size 500,000. The scenario with true constant relative treatment effect (panel A) had a true prediction AUC of 0.75 and sample size of 4250. The RMSE is also presented for strong linear (panel B), strong quadratic (panel C), and non-monotonic (panel D) from constant relative treatment effects. Panels on the right side present the true relations between baseline risk (x-axis) and absolute treatment benefit (y-axis). The 2.5, 25, 50, 75, and 97.5 percentiles of the risk distribution are expressed by the boxplot.

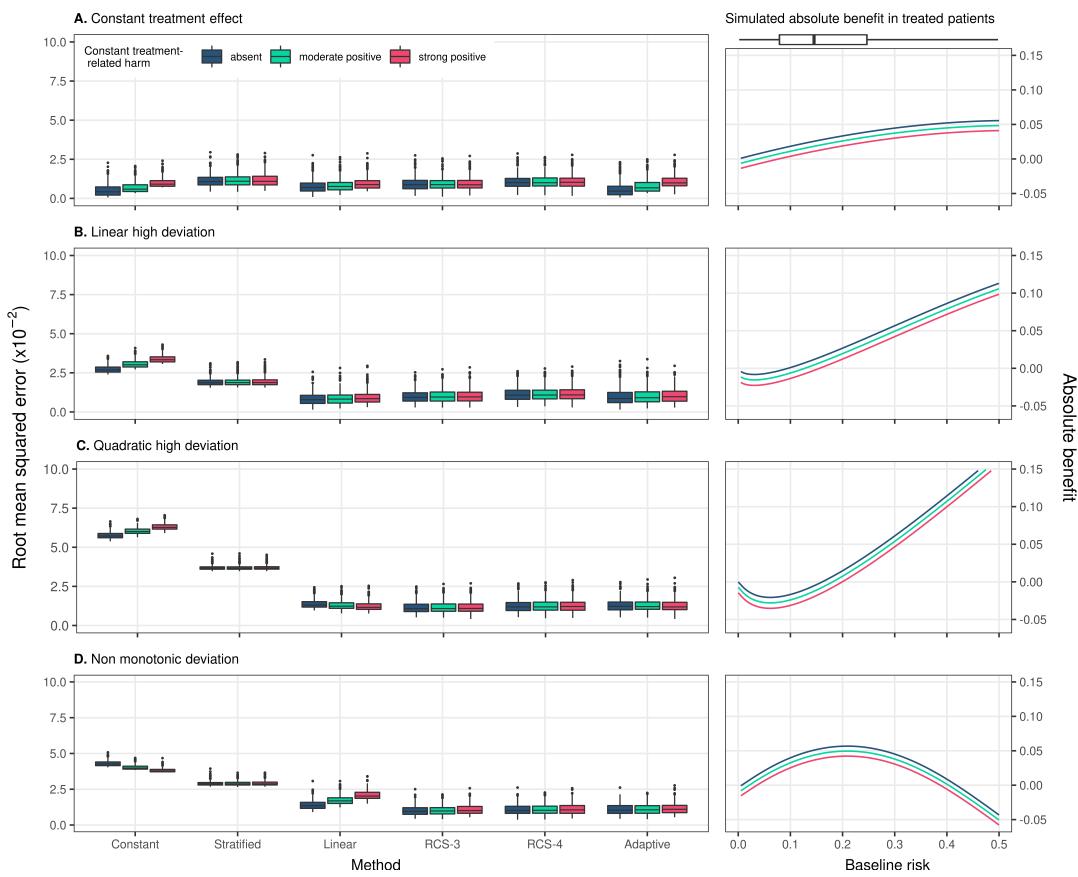


Figure 2: RMSE of the considered methods across 500 replications calculated in simulated samples of size 17,000 rather than 4,250 in Figure 1

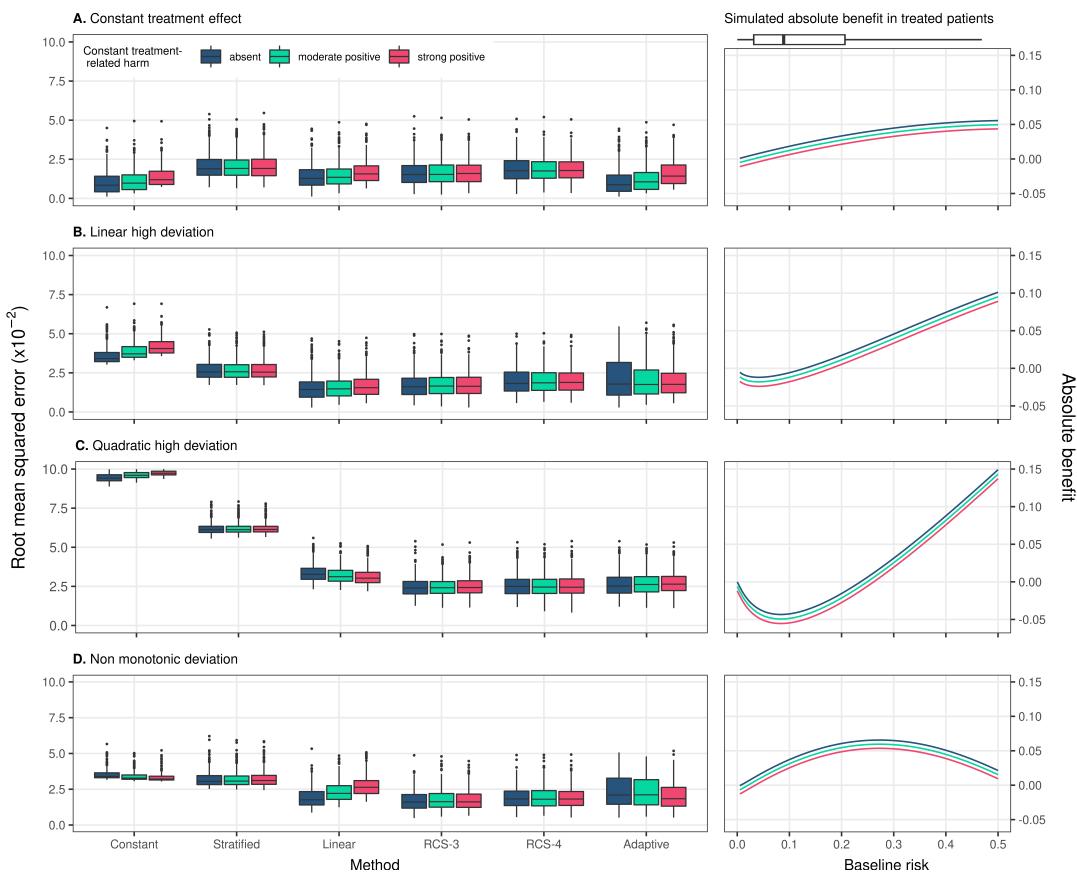


Figure 3: RMSE of the considered methods across 500 replications calculated in a simulated samples 4,250. True prediction AUC of 0.85.

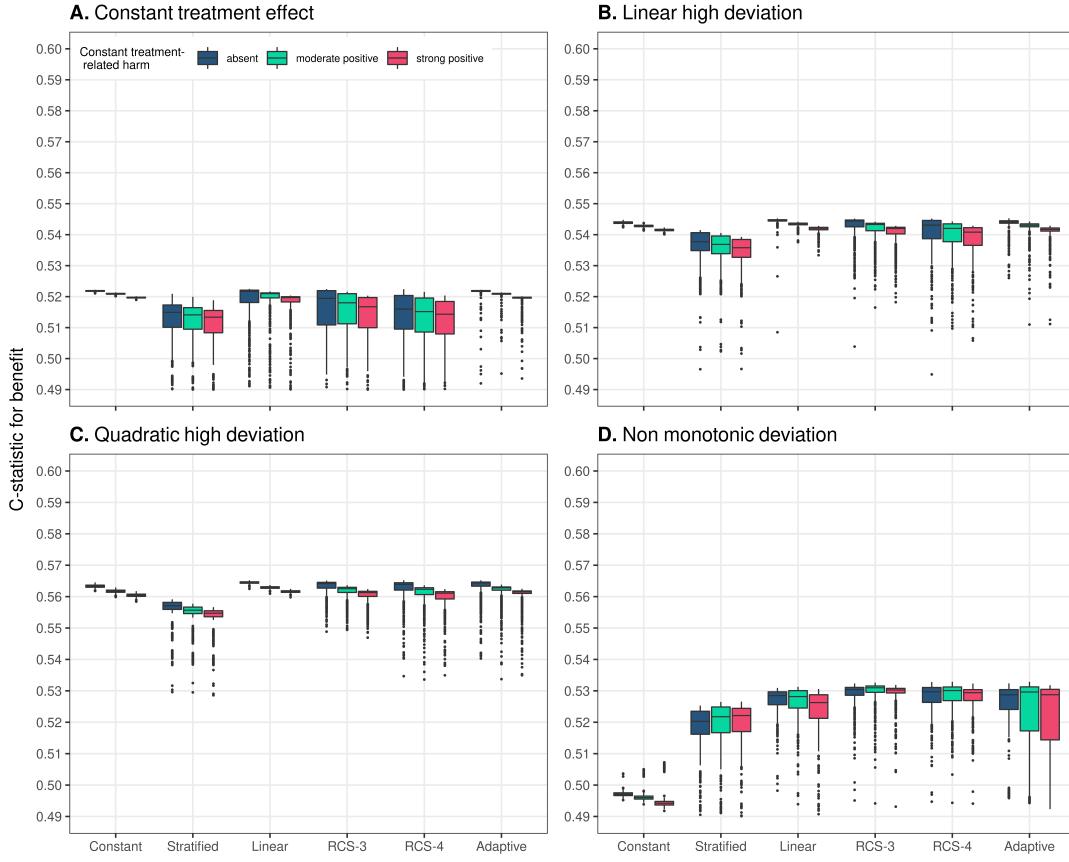


Figure 4: Discrimination for benefit of the considered methods across 500 replications calculated in a simulated samples of size 4,250. True prediction AUC of 0.75.

124 effect, all methods discriminated quite similarly (Figure 4; panels B-C). In the scenario with non-monotonic
 125 deviations, the constant effect model had much lower discriminative ability compared to all other methods
 126 (median AUC of 0.4971 for the constant effects model, 0.5285 for the linear interaction model and 0.5304 for
 127 the best-performing RCS-3; Figure 4; panel D). The adaptive approach was unstable in terms of discrimination
 128 for benefit, especially in the presence of treatment-related harms. With increasing number of RCS knots, we
 129 observed decreasing median values and increasing variability of the c-for-benefit in all scenarios. When we increased
 130 the sample size to 17,000 we observed similar trends, however the performance of all methods was more stable
 131 (Supplement, Figure S4). Finally, when we increased the true prediction AUC to 0.85 the adaptive approach in the
 132 case of non-monotonic deviations was,
 133 again, quite unstable, especially with null or moderate treatment-related harms (Supplement, Figure S5). In
 134 these scenarios, the adaptive approach tended to select more often the inferior constant treatment effect method
 135 (Supplement, Figure S3)
 136 In terms of calibration for benefit, the constant effects model outperformed all other models in the scenario with
 137 true constant treatment effects, but was miscalibrated for all deviation scenarios (Figure 5). The linear interaction

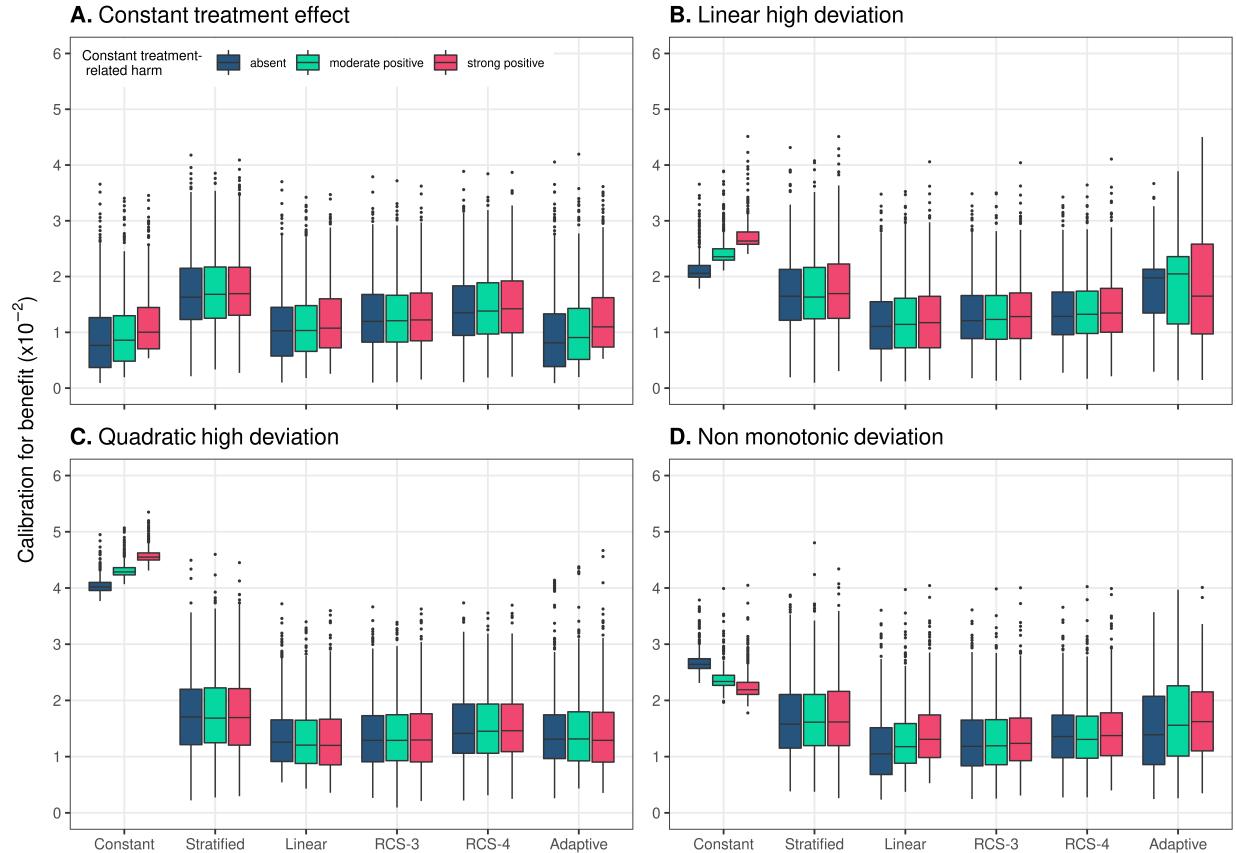


Figure 5: Calibration for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.75 and sample size of 4,250.

model showed best or close to best calibration across all scenarios and only showed worse calibration compared to RCS-3 in case of non-monotonic deviations and treatment-related harms (Figure 5; panel D). The adaptive approach was worse calibrated in scenarios with strong linear and non-monotonic deviations compared to the linear interaction model and RCS-3. When we increased sample size to 17,000 similar conclusions on calibration for benefit could be drawn. As expected, all methods displayed more stable calibration performance due to the larger number of patients (Supplement, Figure S6). When we increased the true prediction AUC to 0.85, the linear interaction model was worse calibrated, on average, than RCS-3 in the case of strong quadratic deviations from constant relative treatment effects (Supplement, Figure S7).

The results from all individual scenarios can be explored online at https://arekkas.shinyapps.io/simulation_viewer/.

3.2. Empirical illustration

In line with previous analyses [13,14], 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

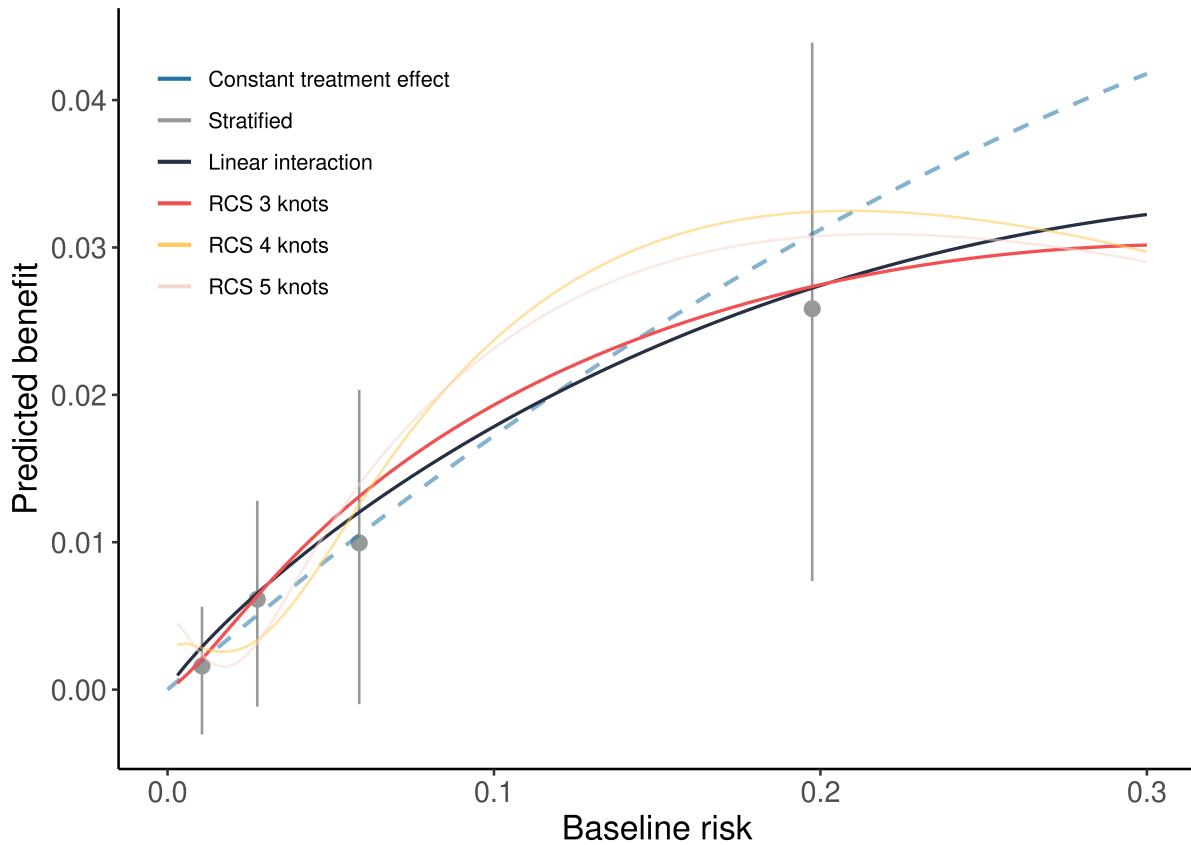


Figure 6: Individualized absolute benefit predictions based on baseline risk when using a constant treatment effect approach, a linear interaction approach and RCS smoothing using 3,4 and 5 knots. Risk stratified estimates of absolute benefit are presented within quartiles of baseline risk as reference.

151 30-day mortality risk. A constant effect of treatment was included in the model. When deriving risk predictions
 152 for individuals we set the treatment indicator to 0. More information on model development can be found in the
 153 supplement.

154 We used the risk linear predictor to fit the proposed methods under study for individualizing absolute benefit
 155 predictions. All methods predicted increasing benefits for patients with higher baseline risk predictions, but the
 156 fitted patterns were somewhat different. The adaptive approach selected the model with RCS smoothing with 4
 157 knots. However, for very low baseline risk the decreasing predicted benefit with increasing risk may be somewhat
 158 too flexible. The more robust models, such as the linear interaction model and the model with RCS smoothing (3
 159 knots), gave very similar benefit predictions. These models followed the evolution of the stratified estimates very
 160 closely and may therefore be preferable for use in clinical practice. The linear interaction model had somewhat
 161 lower AIC compared to the model with RCS smoothing (3 knots), slightly better cross-validated discrimination
 162 (c-for-benefit 0.526 vs 0.525) and quite similar cross-validated calibration (ICI-for benefit 0.0115 vs 0.0117).

163 **4. Discussion**

164 The linear interaction model and the RCS-3 model both displayed very good performance under many of
165 the considered simulation scenarios, in contrast with the constant relative treatment effect model. The linear
166 interaction model was optimal in cases with smaller sample sizes and moderately performing baseline risk prediction
167 models, that is, it had lower RMSE, was better calibrated for benefit and had better discrimination for benefit, even
168 in scenarios with strong quadratic deviations. In scenarios with true non-monotonic deviations, the linear interaction
169 model was outperformed by RCS-3, especially in the presence of true treatment-related harms. Increasing the sample
170 size or the prediction model's discriminative ability favored RCS-3, especially in scenarios with non-monotonic
171 deviations and in the presence of treatment-related harms.

172 RCS-4 and RCS-5 proved to be too flexible in all considered scenarios, as indicated by higher RMSE, increased
173 variability of discrimination for benefit and worse calibration of benefit predictions. Even with larger sample sizes
174 and strong quadratic or non-monotonic deviations from the base case scenario of constant relative treatment
175 effects, these more flexible restricted cubic splines did not outperform the simpler RCS-3. These approaches may
176 only be helpful if we expect more extreme patterns of heterogeneous treatment effects compared to the quadratic
177 deviations considered here. Considering interactions in RCS-3 models as the most complex approach often may be
178 reasonable.

179 The constant treatment effect model, despite having adequate performance in the presence of weak treatment
180 effect heterogeneity on the relative scale, quickly broke down with stronger deviations from constant relative
181 treatment effects. In these cases, the stratified approach generally had lower error rates compared to the constant
182 treatment effect model. Such stepwise treatment benefit estimates are useful for visually demonstrating treatment
183 effect heterogeneity but may be considered insufficient for making individualized benefit predictions.

184 Increasing the discriminative ability of the risk model—by increasing the predictor coefficients of the true risk
185 model—reduced RMSE for all methods. This increase in discriminative ability translates in higher variability of
186 predicted risks, which, in turn, allows the considered methods to better capture absolute treatment benefits. As
187 a consequence, the increase in discriminative ability of the risk model also led to higher discrimination between
188 those with low or high benefit (as reflected in values of c-for-benefit). Even though risk model performance is very
189 important for the ability of risk-based methods to predict treatment benefit, prediction model development was
190 outside the scope of this work and has already been studied extensively [5,8,9].

191 The adaptive approach had adequate performance, following closely on average the performance of the “true”
192 model in most scenarios. With smaller sample sizes it tended to miss the treatment-risk interactions and selected
193 simpler models (Supplementary Table S7). This conservative behavior resulted in increased RMSE variability
194 in these scenarios, especially in the case of true strong linear or non-monotonic deviations from the base case

195 scenario. Therefore, in the case of smaller sample sizes the simpler linear interaction model may be a safer choice
196 for predicting absolute benefits, if a non-constant treatment effect is suspected.

197 Risk-based approaches to predictive HTE estimate treatment benefit as a function of baseline risk. A limitation
198 of our study is that we assumed treatment benefit to be a function of baseline risk in the majority of the simulation
199 scenarios. We also considered constant moderate and strong treatment-related harms, applied on the absolute scale
200 to expand the range of scenarios in line with previous work [??]. Also, we considered a small number of scenarios
201 with true treatment-covariate interactions, in which our main conclusions remained the same (Supplement, XX).
202 Future simulation studies could explore the effect of more extensive deviations from risk-based treatment effects.

203 In our simulations we only focused on risk-based methods, using baseline risk as a reference in a two-stage
204 approach to individualizing benefit predictions. However, there is a plethora of different methods, ranging from
205 treatment effect modeling to tree-based approaches available in more recent literature [15–18]. Many of these
206 methods rely on incorporating treatment-covariate interactions in the prediction of benefit. An important caveat of
207 such approaches is that they may be prone to overfitting, thus exaggerating the magnitude of the predicted benefits.
208 In a wide range of simulation settings, a simpler risk modeling approach was consistently better calibrated for
209 benefit compared to more complex treatment effect modelling approaches [5]. Similarly, when SYNTAX score II, a
210 model developed for identifying patients with complex coronary artery disease that benefit more from percutaneous
211 coronary intervention or from coronary artery bypass grafting was redeveloped using fewer treatment-covariate
212 interactions had better external performance compared to its predecessor[19,20].

213 In conclusion, the linear interaction approach is a viable option with smaller sample sizes and/or moderately
214 performing risk prediction models if we consider a non-constant relative treatment effect plausible. RCS-3 is
215 a better option when non-monotonic deviations from a constant relative treatment effect and/or substantial
216 treatment-related harms are anticipated. With larger sample size, an adaptive approach – selecting the method
217 with optimal AIC – can also be considered as an automated alternative.

218 **5. References**

- 219 [1] Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of
220 treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology* 2013;66:818–25.
221 <https://doi.org/10.1016/j.jclinepi.2013.02.009>.
- 222 [2] Rekkas A, Paulus JK, Raman G, Wong JB, Steyerberg EW, Rijnbeek PR, et al. Predictive approaches
223 to heterogeneous treatment effects: A scoping review. *BMC Medical Research Methodology* 2020;20.
224 <https://doi.org/10.1186/s12874-020-01145-1>.
- 225 [3] Kent DM, Paulus JK, Klaveren D van, D'Agostino R, Goodman S, Hayward R, et al. The predictive
226 approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine* 2019;172:35.
227 <https://doi.org/10.7326/m18-3667>.
- 228 [4] Kent DM, Klaveren D van, Paulus JK, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches
229 to treatment effect heterogeneity (PATH) statement: Explanation and elaboration. *Annals of Internal Medicine*
230 2019;172:W1. <https://doi.org/10.7326/m18-3668>.
- 231 [5] Klaveren D van, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of
232 treatment effects and were prone to treatment mistargeting. *Journal of Clinical Epidemiology* 2019;114:72–83.
233 <https://doi.org/10.1016/j.jclinepi.2019.05.029>.
- 234 [6] Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in
235 treatment effects in clinical trials: A proposal. *Trials* 2010;11. <https://doi.org/10.1186/1745-6215-11-85>.
- 236 [7] Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect
237 heterogeneity: Re-analysis of individual participant data from 32 large clinical trials. *International Journal of
238 Epidemiology* 2016;dyw118. <https://doi.org/10.1093/ije/dyw118>.
- 239 [8] Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity
240 in treatment effects in clinical trials. *Circulation: Cardiovascular Quality and Outcomes* 2014;7:163–9.
241 <https://doi.org/10.1161/circoutcomes.113.000497>.
- 242 [9] Abadie A, Chingos MM, West MR. Endogenous stratification in randomized experiments. *The Review of
243 Economics and Statistics* 2018;100:567–80. https://doi.org/10.1162/rest_a_00732.
- 244 [10] Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: Determining relationships between
245 predictors and response. *JNCI Journal of the National Cancer Institute* 1988;80:1198–202. <https://doi.org/10.1093/jnci/80.15.1198>.
- 246 [11] Klaveren D van, Steyerberg EW, Serruys PW, Kent DM. The proposed “concordance-statistic for benefit”
247 provided a useful metric when modeling heterogeneous treatment effects. *Journal of Clinical Epidemiology*
248 2018;94:59–68. <https://doi.org/10.1016/j.jclinepi.2017.10.021>.

- 250 [12] Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the
251 calibration of logistic regression models. *Statistics in Medicine* 2019;38:4051–65. <https://doi.org/10.1002/sim.8281>.
- 253 [13] Califf RM, Woodlief LH, Harrell FE, Lee KL, White HD, Guerci A, et al. Selection of thrombolytic
254 therapy for individual patients: Development of a clinical model. *American Heart Journal* 1997;133:630–9.
255 [https://doi.org/10.1016/s0002-8703\(97\)70164-9](https://doi.org/10.1016/s0002-8703(97)70164-9).
- 256 [14] Steyerberg EW, Bossuyt PMM, Lee KL. Clinical trials in acute myocardial infarction: Should we adjust
257 for baseline characteristics? *American Heart Journal* 2000;139:745–51. [https://doi.org/10.1016/s0002-8703\(00\)90001-2](https://doi.org/10.1016/s0002-8703(00)90001-2).
- 259 [15] Athey S, Tibshirani J, Wager S. Generalized random forests. *The Annals of Statistics* 2019;47. <https://doi.org/10.1214/18-aos1709>.
- 261 [16] Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating individual treatment effect in observational data
262 using random forest methods. *Journal of Computational and Graphical Statistics* 2018;27:209–19. <https://doi.org/10.1080/10618600.2017.1356325>.
- 264 [17] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal*
265 *of the American Statistical Association* 2018;113:1228–42. <https://doi.org/10.1080/01621459.2017.1319839>.
- 266 [18] Powers S, Qian J, Jung K, Schuler A, Shah NH, Hastie T, et al. Some methods for heterogeneous treatment
267 effect estimation in high dimensions. *Statistics in Medicine* 2018;37:1767–87.
- 268 [19] Farooq V, Van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, et al. Anatomical and
269 clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous
270 coronary intervention for individual patients: Development and validation of syntax score ii. *The Lancet*
271 2013;381:639–50.
- 272 [20] Takahashi K, Serruys PW, Fuster V, Farkouh ME, Spertus JA, Cohen DJ, et al. Redevelopment and validation
273 of the syntax score ii to individualise decision making between percutaneous and surgical revascularisation in
274 patients with complex coronary artery disease: Secondary analysis of the multicentre randomised controlled
275 syntaxes trial with external cohort validation. *The Lancet* 2020;396:1399–412.