Linear interaction of treatment with baseline risk was sufficient for predicting individualized benefit

Alexandros Rekkas
Department of Medical Informatics
Erasmus Medical Center
Rotterdam, The Netherlands

Peter R. Rijnbeek
Department of Medical Informatics
Erasmus Medical Center
Rotterdam, The Netherlands

Ewout W. Steyerberg
Department of Biomedical Data Sciences
Leiden University Medical Center
Leiden, The Netherlands

David van Klaveren
Department of Public Health
Erasmus Medical Center
Rotterdam, The Netherlands

Abstract

Objective: To compare different risk-based methods predicting individualized treatment effects in RCTs. **Study Design and Setting:** We simulated data using diverse assumptions for 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. 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. Starting from a base scenario (sample size 4,250, treatment odds ratio 0.8, AUC of the PI 0.75), we varied the sample size, the treatment effect size, and the PI's AUC. **Results:** Models including a PI by treatment interaction performed best with smaller sample sizes and/or lower AUC of the PI. Otherwise, RCS (3 knots) models proved more robust for the majority of the deviation settings. The adaptive approach was unstable with smaller sample sizes. **Conclusion:** Depending on the setting, the linear interaction or the RCS (3 knots) model should be preferred.

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]. 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, i.e. 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 may be accurate, these estimates do not apply to individual patients, especially for patients with predicted risk at the boundaries of the risk intervals. Hence, the risk-stratified approach is useful for exploring and presenting HTE, but is not useful for supporting treatment decisions for individual patients.

To individualize treatment effects, the recent PATH statement suggested various risk-based models including a prognostic index of baseline risk (PI) and treatment assignment [3,4]. We aimed to summarize and compare different risk-based models for predicting individualized treatment effects. We simulated RCT settings to compare the performance of these models under different assumptions of the relationship between baseline risk and treatment. We illustrated the different models by a case study of predicting individualized effects of tissue plasminogen activator (tPA) versus streptokinase treatment in patients with an acute myocardial infarction (MI).

2 Methods

2.1 Simulation scenarios

For each patient we generated 8 baseline covariates $x_1, \ldots, x_4 \sim N(0,1)$ and $x_5, \ldots, x_8 \sim B(1, 0.2)$. Treatment was allocated using a 50:50 split. Outcomes for patients in the control arm were generated from a logistic regression model including all baseline covariates. In the base scenarios coefficient values were such, that the AUC of the logistic regression model was 0.75 and the event rate in the control arm was 20%. Binary outcomes in the control arm were generated 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 high 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:

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

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 simulated scenarios where a constant absolute harm is applied across all treated patients. In this case we have $P(y = 1|X, t_x = 1) = \text{expit}(lp_1) + \text{harm}$.

The sample size for the base scenarios was set to 4,250 (80% power for the detection of a marginal OR of 0.8). We evaluated the effect of smaller or larger sample sizes of 1,063 and 17,000, respectively. We also evaluated the effect of worse or better discriminative ability for risk, adjusting the baseline covariate coefficients, such that the AUC of the regression model in the control arm was 0.65 and 0.85 respectively.

Combining all these settings resulted in a simulation study of 648 scenarios (exact settings in the supplementary material).

2.2 Individualized risk-based benefit predictions

All methods assume that a risk prediction model is available to assign risk predictions to individual patients. For the simulations we developed a prediction model internally, using logistic regression including main effects for all baseline covariates and treatment assignment. Risk predictions for individual patients were based on treatment assignment to the control arm, that is setting treatment assignment to 0.

A *stratified HTE method* has been suggested as an alternative to traditional subgroup analyses. Patients are stratified into equally-sized risk strata—in this case based on risk quartiles. Absolute treatment effects within risk strata are estimated by the difference in event rate between patients in the control arm and patients in the treated arm. We considered this approach as a reference, expecting it to perform worse than the other candidates, as its objective is not to individualize benefit prediction.

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

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

Fourth, we used *restricted cubic splines* (RCS) to relax the linearity assumption on the effect of the linear predictor [7]. We considered splines with 3 (RCS-3), 4 (RCS-4) and 5 (RCS-5) knots to compare models with different levels of flexibility.

Finally, we considered an adaptive approach using Akaike's Information Criterion (AIC) for model selection. The candidate models were: a constant treatment effect model, a model with a linear interaction with treatment and RCS models with 3, 4 and 5 knots.

2.3 Evaluation metrics

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

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

We compared the discriminative ability of the methods under study using c-for-benefit [8]. 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 [9]. 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 0 represent better calibration.

3 Results

3.1 Simulations

The linear interaction model outperformed all RCS methods in terms of RMSE in scenarios with true constant relative treatment effect (OR = 0.8, N = 4,250 and AUC = 0.75), strong linear and even strong quadratic deviations (Figure 1). However, with non-monotonic deviations errors of the linear interaction model increased substantially, especially in the presence of treatment-related harms. In these scenarios, RCS-3 proved to be quite robust outperforming all other methods. The constant treatment effect approach had overall best performance under true constant treatment effect settings, but was sensitive to all considered deviations, resulting in increased error rates. Finally, the adaptive approach had comparable performance to the best-performing method in each scenario. However, we observed increased error variability in the case of linear and non-monotonic deviations, especially for moderate or strong treatment-related harms.

In the case of true constant relative treatment effects, the adaptive approach selected the constnant effect model the majority of the time, even in the presence of treatment-related harms (96% of the time with no harms to

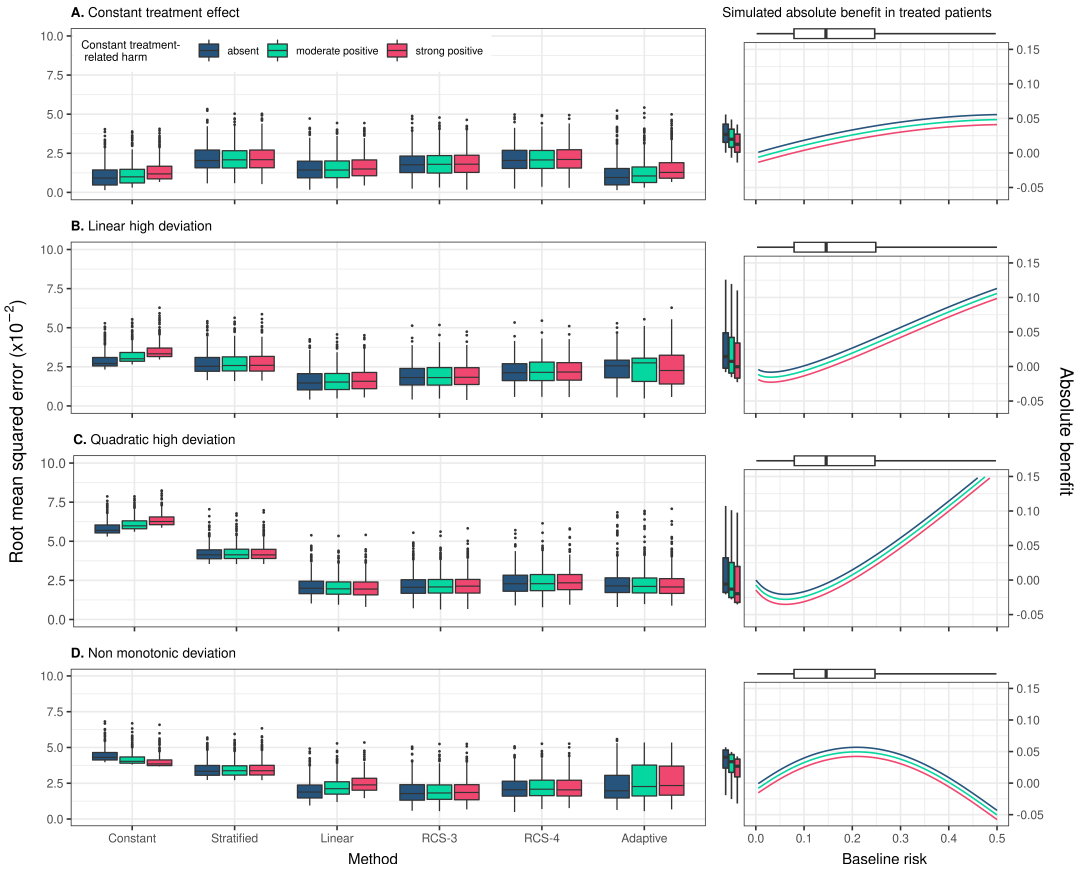

Figure 1: RMSE of the considered methods across 500 replications calculated in a simulated super-population of size 500,000. The scenario with true constant relative treatment effect had a true prediction AUC of 0.75 and sample size of 4250. Results are presented under moderate linear, strong linear, and strong quadratic deviations from constant relative treatment effects.

88% with strong harms). With strong linear or quadratic deviations and increasing treatment-related harms the adaptive approach increasingly favored the linear interaction model. In the case of true non-monotonic deviations, we observed increasing selection frequencies for RCS-3 with increasing treatment harms (from 32% with no harms to 53% with strong harms), whereas selection frequencies for the linear interaction model dropped from 40% with no harms to 7% with strong harms. Finally, in the case of strong linear and non-monotonic deviations selection frequencies of the constant treatment effect model remained high despite increasing treatment harms (Supplement, Table XX).

Increasing the sample size to 17,000 favored RCS-3 the most, as it achieved lowest or close to lowest RMSE across all scenarios (Figure 2). Especially in cases of strong quadratic and non-monotonic deviations RCS-3 had lower error rates (median RMSE 0.011 for strong quadratic deviations and 0.010 for non-monotonic deviations with no treatment-related harms) compared to the linear interaction approach (median RMSE 0.013 and 0.014, respectively), regardless of treatment-related harms strength. The issues with the large error variability of the adaptive approach improved with larger sample sizes.

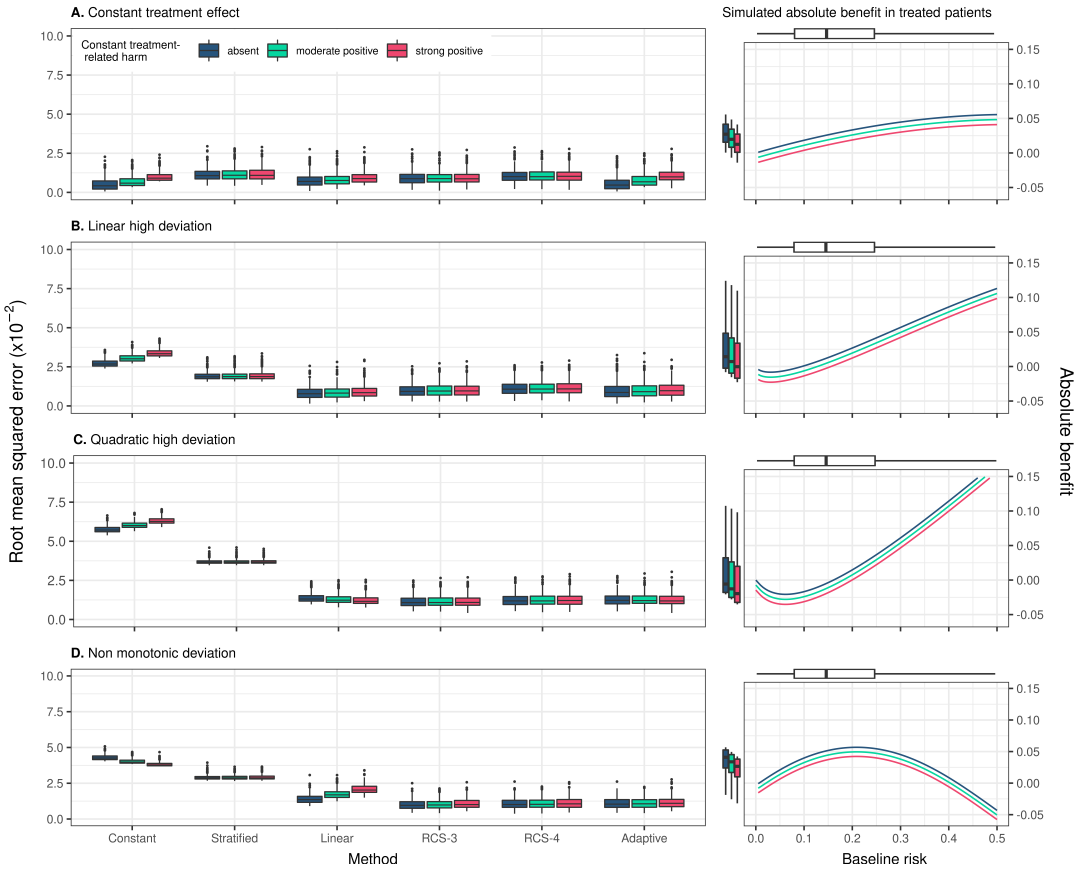

Figure 2: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000. Sample size 17,000 rather than 4250 in Figure 1

The adaptive approach tended to increasingly favor smoother methods (especially RCS-3) with increasing treatment-related harms (see Supplement, Table XX). However, in the case of true strong quadratic deviations the opposite was observed: selection frequency of the linear interaction model increased from 31% (no harms) to 50% (strong harms) whereas for RCS-3 decreased from 52% (no harms) to 34% (strong harms).

When we increased the AUC of the true prediction model to 0.85 (OR = 0.8 and N = 4,250) the constant effect model outperformed all other methods in the case of true constant treatment effects, but proved to be the least robust to deviations. Again, RCS-3 had the lowest error rates in the case of strong quadratic or monotonic deviations and very comparable performance to the best-performing linear interaction model in the case of strong linear deviations (median RMSE 0.016 for RCS-3 compared to 0.014 for the linear interaction model). The adaptive approach, though it performed similar to the best performing method in each scenario, on average, had increased variability in error rates in the case of strong linear and non-monotonic deviations. In these scenarios, the adaptive approach often selected the constant treatment effect model (23% and 25% in the strong linear and non-monotonic deviation scenarios without treatment-related harms, respectively).

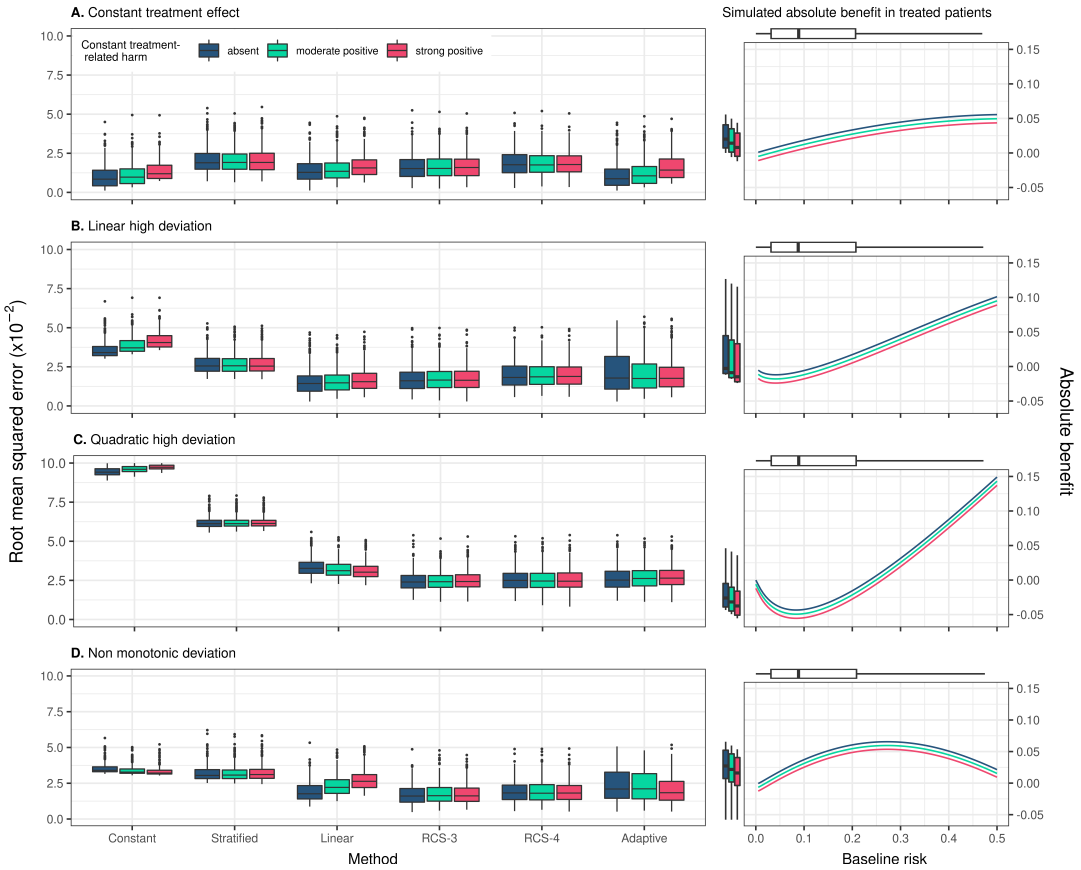

Figure 3: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.85 and sample size of 4,250.

In terms of discrimination for benefit (OR = 0.8, N = 4,250 and AUC = 0.75), all methods performed similarly, on average (Figure 4). In the case of non-monotonic deviations, the constant effect model had much lower discriminative performance compared to the rest of the methods (median AUC of 0.04 for the constant effects model compared to the best-performing RCS-3 with 0.02). The linear interaction model was the most stable compared to the other methods in terms of error variability. With increasing number of RCS knots, we observed decreasing median values and increasing variability of the c-for-benefit in all scenarios.

In terms of calibration for benefit, the constant effects model outperformed all other models in the case of true constant treatment effects, but was miscalibrated for all deviation scenarios (Figure 5). The linear interaction model showed best or close to best calibration across all scenarios and only showed worse calibration compared to RCS-3 in the case of non-monotonic deviations and treatment-related harms. The adaptive approach was worse calibrated in scenarios with strong linear and non-monotonic deviations compared to the linear interaction model and RCS-3.

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

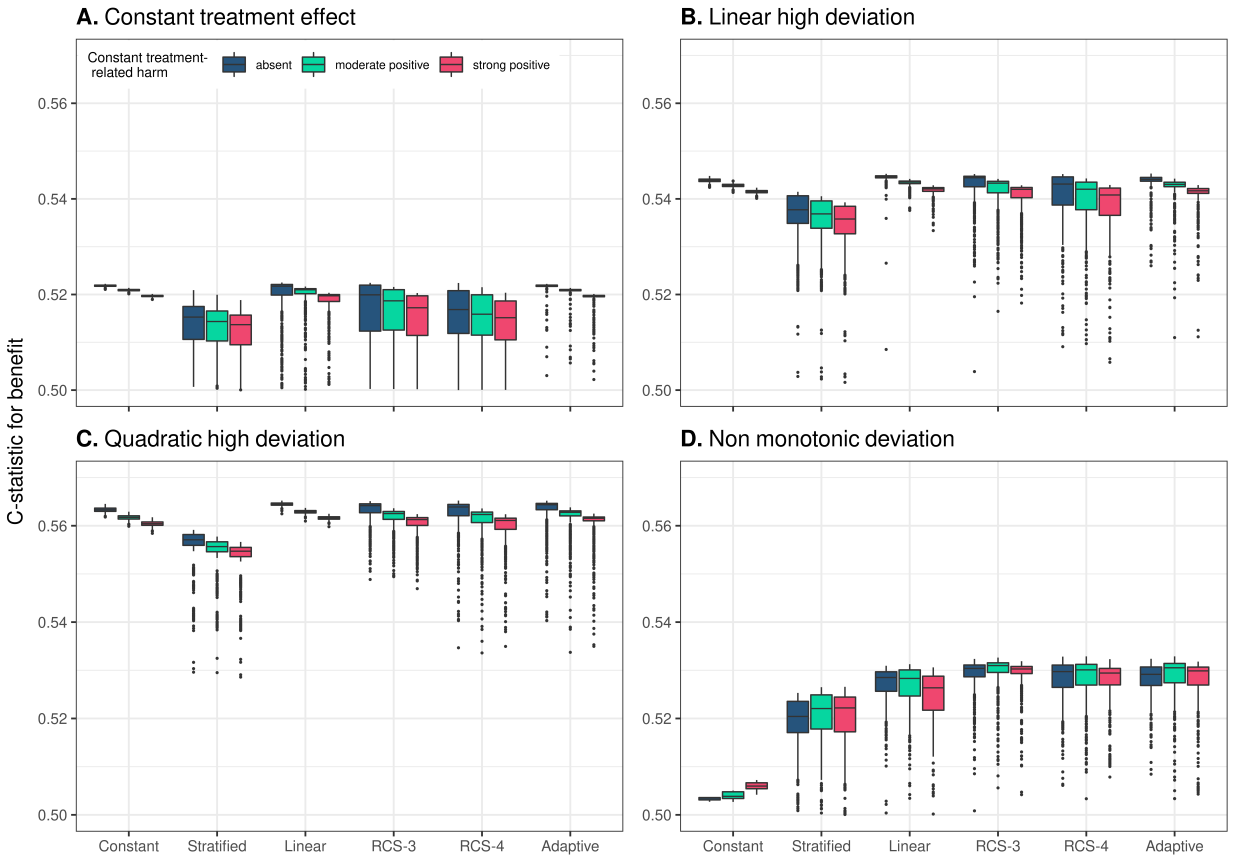

Figure 4: Discrimination for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.75 and sample size of 4,250.

3.2 Case study

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 proposed methods under study for individualizing absolute benefit predictions. All methods predicted increasing benefits for patients with higher baseline risk predictions, but the fitted patterns were clearly different. The adaptive approach selected the model with RCS smoothing with 4 knots. However, for very low baseline risk this model predicted decreasing benefit with increasing risk may be

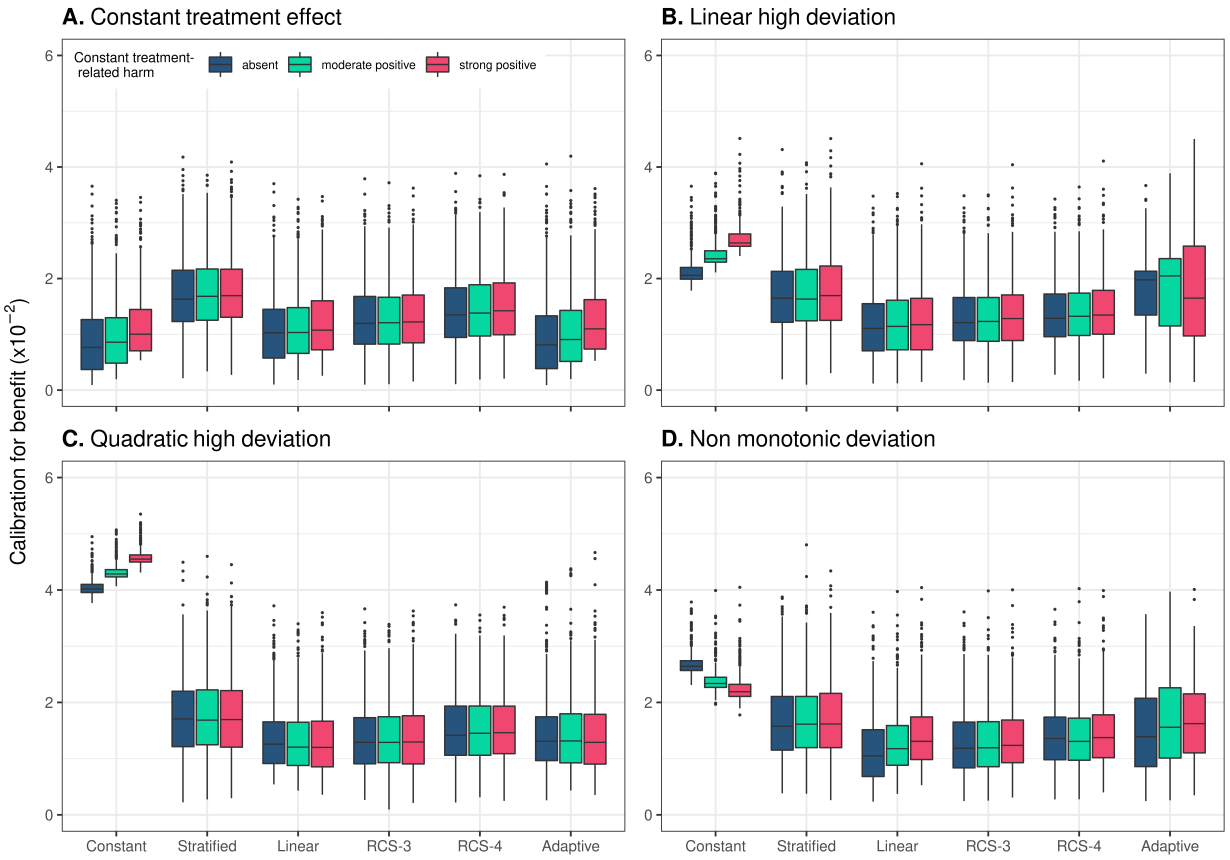

Figure 5: Calibration for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.75 and sample size of 4,250.

somewhat too flexible. The more robust models, the linear interaction model or the model with RCS smoothing (3 knots), gave very similar benefit predictions, followed the evolution of the stratified estimates very closely and may therefore be preferable for use in clinical practice. The linear interaction model had somewhat lower AIC compared to the model with RCS smoothing (3 knots), slightly better cross-validated discrimination (c-for-benefit 0.526 vs 0.525) and quite similar cross-validated calibration (ICI-for benefit 0.0115 vs 0.0117).

4 Discussion

The linear interaction model displayed very good performance overall under many of the considered simulation scenarios. Especially in cases with smaller sample sizes and moderately performing baseline risk prediction models it had lower RMSE, was better calibrated for benefit and had better discrimination for benefit, even in scenarios with strong quadratic deviations. However, in scenarios with true non-monotonic deviations, the linear interaction model was outperformed by RCS-3, especially in the presence of true treatment-related harms. Increasing the sample size or the prediction model's discrimination favored RCS-3 which had better or very comparable performance to the linear interaction model, but was more robust to non-monotonic deviations and the presence of treatment-related harms.

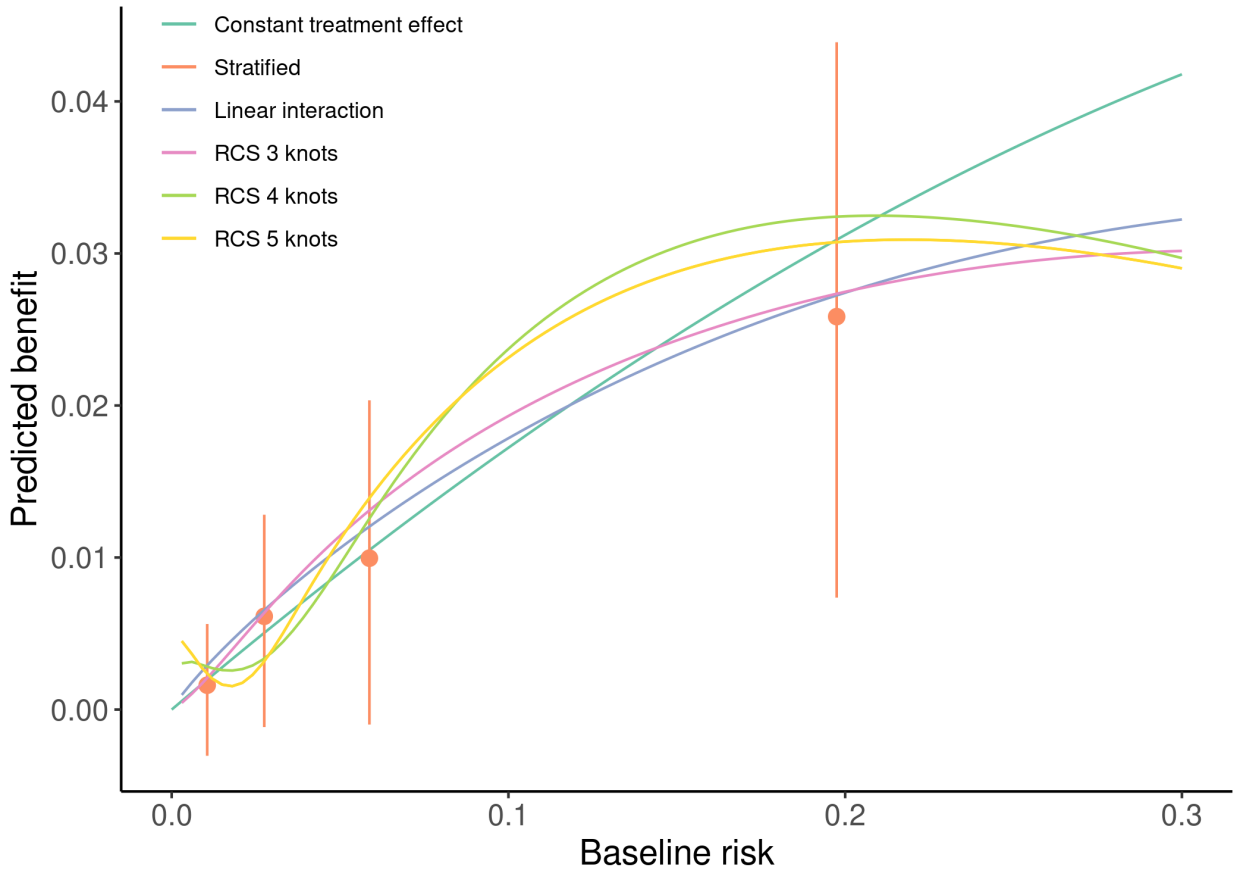

Figure 6: Individualized absolute benefit predictions based on baseline risk when using a constant treatment effect approach, a linear interaction approach and RCS smoothing using 3,4 and 5 knots. Risk stratified estimates of absolute benefit are presented within quartiles of baseline risk as reference.

RCS-4 and RCS-5 proved to be too flexible, as indicated by higher RMSE, increased variability of discrimination for benefit and worse calibration of benefit predictions. Even with larger sample sizes and strong quadratic or non-monotonic deviations from the base case scenario of constant relative treatment effects, these more flexible restricted cubic splines did not outperform the simpler RCS-3 These approaches may only be helpful if we expect more extreme patterns of heterogeneous treatment effects compared to the quadratic deviations considered here.

The constant treatment effect model, despite having adequate performance in the presence of weak treatment effect heterogeneity on the relative scale, quickly broke down with stronger deviations from constant relative treatment effects. In these cases, the stratified approach generally had lower error rates compared to the constant treatment effect model. Stepwise treatment benefit estimates are very useful for demonstrating treatment effect heterogeneity–because estimating treatment effect requires groups of patients rather than individual patients–but are not helpful for making individualized absolute benefit predictions.

Increasing the discriminative ability of the risk model–by increasing the predictor coefficients of the true risk model–reduced RMSE for all methods. This increase in discriminative ability translates in higher variability of

predicted risks, which, in turn, allows the considered methods to better capture absolute treatment benefits. As a consequence, the increase in discriminative ability of the risk model also led to higher values of c-for-benefit. Even though risk model performance is very important for the ability of risk-based methods to predict treatment benefit, prediction model development was outside the scope of this work and has already been studied extensively [5,12,13].

The adaptive approach had adequate performance, following closely on average the performance of the "true" model in most scenarios. However, with smaller sample sizes it tended to "miss" the treatment-risk interactions and selected simpler models (Supplementary Table S7). This resulted in increased RMSE variability in these scenarios, especially in the case of true strong linear or non-monotonic deviations from the base case scenario. Therefore, in the case of smaller sample sizes the simpler linear interaction model is a safer choice for predicting absolute benefits.

Risk-based approaches to predictive HTE estimate treatment benefit as a function of baseline risk. A limitation of our study is that we assumed treatment benefit to be a function of baseline risk in the majority of the simulation scenarios. We attempted to address that by introducing constant moderate and strong treatment-related harms, applied on the absolute scale. Also, we considered a small number of scenarios with true treatment-covariate interactions, in which our main conclusions remained the same (Supplement, XX). Future simulation studies could explore the effect of more extensive deviations from risk-based treatment effects.

Recent years have seen an increased interest in predictive HTE approaches focusing on individualized benefit predictions. In our simulations we only focused on risk-based methods, using baseline risk as a reference in a two-stage approach to individualizing benefit predictions. However, there is a plethora of different methods, ranging from treatment effect modeling to tree-based approaches available in more recent literature [14–16]. Simulations are also needed to assess relative performance and define the settings where these break down or outperform each other.

In conclusion, the best option for predicting individualized treatment benefit using a risk-based approach depends on the setting. With smaller sample sizes and/or moderately performing risk prediction models the linear interaction approach is a viable option. When those constraints are not present or when we anticipate non-negligible treatment-related harms, RCS-3 is a better option in terms of error rates, discrimination and calibration for benefit. With larger sample size, an adaptive approach based on AIC can also be considered as a more automated alternative.

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.

[12] Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circulation: Cardiovascular Quality and Outcomes 2014;7:163–9. https://doi.org/10.1161/circoutcomes.113.000497.

[13] Abadie A, Chingos MM, West MR. Endogenous stratification in randomized experiments. The Review of Economics and Statistics 2018;100:567–80. https://doi.org/10.1162/rest_a_00732.

[14] Athey S, Tibshirani J, Wager S. Generalized random forests. The Annals of Statistics 2019;47. https://doi.org/10.1214/18-aos1709.

[15] Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating individual treatment effect in observational data using random forest methods. Journal of Computational and Graphical Statistics 2018;27:209–19. https://doi.org/10.1080/10618600.2017.1356325.

[16] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 2018;113:1228–42. https://doi.org/10.1080/01621459.2017.1319839.