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 treatment effects. **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). 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 displayed optimal or close-to-optimal performance across many simulation scenarios with moderate sample size (N=4,250; ~800 events). The RCS-model was optimal for strong non-linear deviations from a constant treatment effect, particularly when sample size was larger (N=17,000). The adaptive approach also required larger sample sizes. These findings were illustrated in the GUSTO-I trial. **Conclusion**: An interaction between baseline risk and treatment assignment should be considered to improve treatment effect predictions.

Keywords: treatment effect heterogeneity absolute benefit prediction models

1. Introduction

Predictive approaches to 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 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 modifier [3,4]. In simulations, modeling treatment-covariate interactions, often led to miscalibrated predictions of absolute benefit, contrary to risk-based methods, despite their weaker discrimination of benefit in the presence of true effect modifiers [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]. This approach identified substantial absolute treatment effect differences between low-risk and high-risk patients in a re-analysis of 32 large trials [7]. However, even though estimates at the risk subgroup level may be accurate, these estimates may not apply to individual patients.

In the current simulation study, we aim to summarize and compare different risk-based models for predicting treatment effects. We simulate different relations between baseline risk and treatment effects and also consider potential harms of treatment. We illustrate the different models by a case study of predicting individualized effects of treatment for acute myocardial infarction (MI) in a large RCT.

2. Methods

2.1. 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 measured covariates. 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 $\big(Y(0), Y(1)\big) \perp\!\!\!\perp Z|X$, as we are in the RCT setting, we can predict CATE from

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

2.2. Simulation scenarios

We simulated a typical RCT, comparing equally-sized treatment and control arms in terms of a binary outcome. For each patient we generated 8 baseline covariates $x_1, \ldots, x_4 \sim N(0, 1)$ and $x_5, \ldots, x_8 \sim B(1, 0.2)$. Outcomes in the control arm were generated from Bernoulli variables with true probabilities following a logistic regression model

including all baseline covariates, i.e. $P(Y(0) = 1 \mid X = x) = \text{expit}(lp_0) = e^{lp_0}/(1+e^{lp_0})$, with $lp_0 = lp_0(x) = x^t\beta$. In the base scenarios coefficient values β were such, that the AUC of the logistic regression model was 0.75 and the event rate in the control arm was 20%.

Outcomes in the treatment arm were first generated using 3 simple scenarios: absent (OR = 1), moderate (OR = 0.8) or strong (OR = 0.5) constant relative treatment effect. We then introduced linear, quadratic and non-monotonic deviations from constant treatment effects using:

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

where lp_1 is the true linear predictor in the treatment arm, so that $P(Y(1) = 1 \mid X = x) = \text{expit}(lp_1)$. Finally, we incorporated constant absolute harms for all treated patients, such that $P(Y(1) = 1|X = x) = \text{expit}(lp_1) + \text{harm}$.

The sample size for the base scenarios was set to 4,250 (80% power for the detection of a marginal OR of 0.8 with the standard alpha of 5%). We evaluated the effect of smaller or larger sample sizes of 1,063 and 17,000, respectively. We also evaluated the effect of risk model discriminative ability, adjusting the baseline covariate coefficients, such that the AUC of the regression model in the control arm was 0.65 and 0.85, respectively.

These settings resulted in a simulation study of 648 scenarios covering the observed HTE in 32 large trials as well as many other potential variations of risk-based treatment effect (Supplement, Sections 2 and 3) [7].

2.3. Individualized risk-based benefit predictions

In each simulation run we internally developed a prediction model on the entire population, using a logistic regression with main effects for all baseline covariates and treatment assignment. Individual risk predictions were derived by setting treatment assignment to 0. Another approach would be to derive the prediction model solely on the control patients; however, it has been shown to lead to biased benefit predictions [5,8,9].

A *stratified HTE method* has been suggested as an alternative to traditional subgroup analyses [3,4]. 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 control and treatment arm patients. We considered this approach as a reference, expecting it to perform worse than the other candidates, as its objective is to provide an illustration of HTE rather than to optimize individualized benefit predictions.

Second, we considered a model which assumes *constant relative treatment effect* (constant odds ratio). Hence, absolute benefit is predicted from $\tau(x;\hat{\beta}) = \text{expit}(\hat{lp}_0) - \text{expit}(\hat{lp}_0 + \delta_0)$, where δ_0 is the log of the assumed constant odds ratio and $\hat{lp}_0 = \hat{lp}_0(x;\hat{\beta}) = x^t\hat{\beta}$ the linear predictor of the estimated baseline risk model.

Third, we considered a logistic regression model including treatment, the prognostic index, and their linear interaction. Absolute benefit is then estimated from $\tau(x;\hat{\beta}) = \text{expit}(\delta_0 + \delta_1\hat{lp}_0) - \text{expit}(\delta_0 + \delta_2 + (\delta_1 + \delta_3)\hat{lp}_0)$

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 [10]. 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. More specifically, we ranked the constant relative treatment effect model, the linear interaction model, and the RCS models with 3, 4, and 5 knots based on their AIC and selected the one with the lowest value. The extra degrees of freedom were 1 (linear interaction), 2, 3 and 4 (RCS models) for these increasingly complex interactions with the treatment effect.

2.4. 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 and the integrated calibration index (ICI) for benefit (Supplement, Section 6) [11].

For each scenario we performed 500 replications, within which all the considered models were fitted. We simulated a super-population of size 500,000 for each scenario within which we calculated RMSE and discrimination and calibration for benefit of all the models in each replication.

2.5. Empirical illustration

We demonstrated the different methods using 30,510 patients with 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 (total of 2,128 events), recorded for all patients. In line with previous analyses [12,13], 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 (Supplement, Section 8).

3. Results

3.1. Simulations

The constant treatment effect approach outperformed other approaches in the base case scenario (N = 4,250; OR = 0.8; AUC= 0.75; no absolute treatment harm) with a true constant treatment effect (median RMSE:

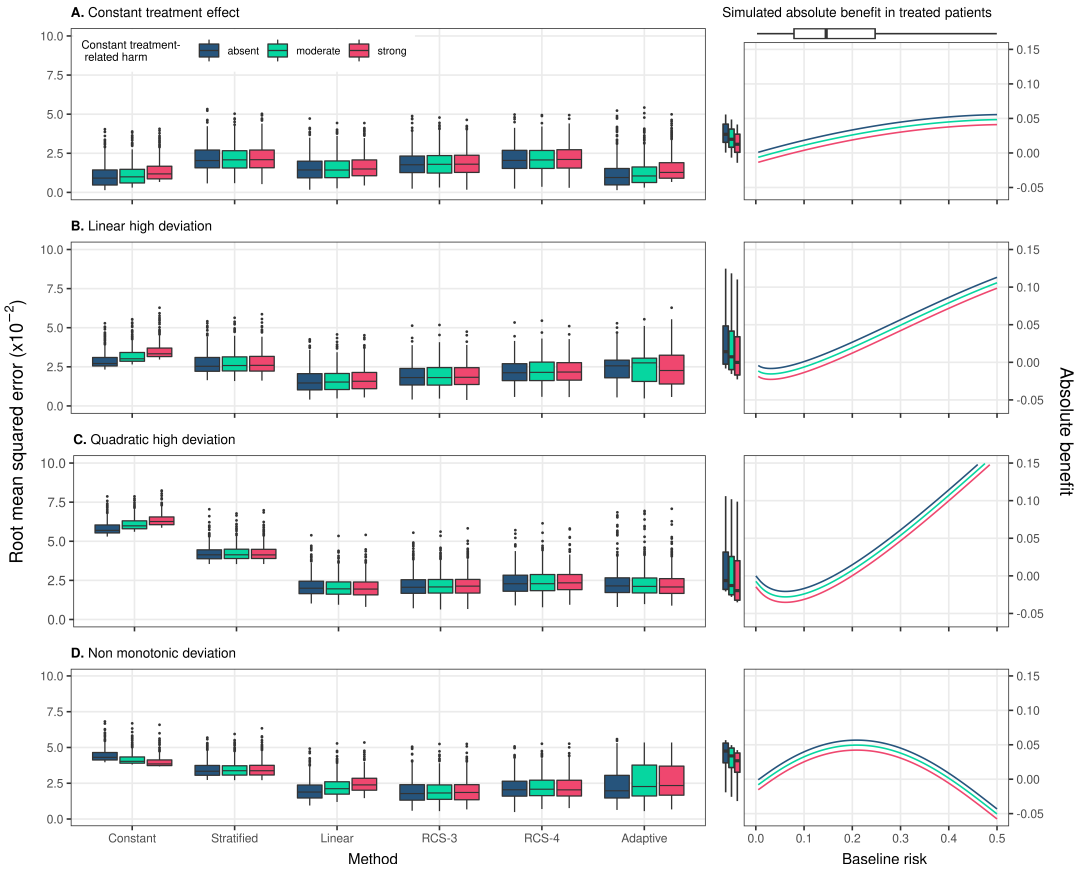

Figure 1: RMSE of the considered methods across 500 replications calculated from a simulated super-population of size 500,000. The scenario with true constant relative treatment effect (panel A) had a true prediction AUC of 0.75 and sample size of 4250. The RMSE is also presented for strong linear (panel 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 on the top. The 2.5, 25, 50, 75, and 97.5 percentiles of the true benefit distributions are expressed by the boxplots on the side of the right-handside panel.

constant treatment effect 0.009; linear interaction 0.014; RCS-3 0.018). The linear interaction model was optimal under true linear deviations (median RMSE: constant treatment effect 0.027; linear interaction 0.015; RCS-3 0.018; Figure 1 panels A-C) and even in the presence of true quadratic deviations (median RMSE: constant treatment effect 0.057; linear interaction 0.020; RCS-3 0.021; Figure 1 panels A-C) from a constant relative treatment effect. With non-monotonic deviations, RCS-3 slightly outperformed the linear interaction model (Median RMSE: linear interaction 0.019; RCS-3 0.018; Figure 1 panel D). With strong treatment-related harms the results were very similar in most scenarios (Figure 1 panels A-C). Under non-monotonic deviations the optimal performance of RCS-3 was more pronounced (Median RMSE: linear interaction 0.024; RCS-3 0.019; Figure 1 panel D). A stronger average treatment effect (OR=0.5) led to larger absolute benefit predictions and consequently to larger RMSE for all approaches, but the relative differences between different approaches were similar to the base case scenario (Supplement, Figure S10).

The adaptive approach had limited loss of performance in terms of the median RMSE to the best-performing

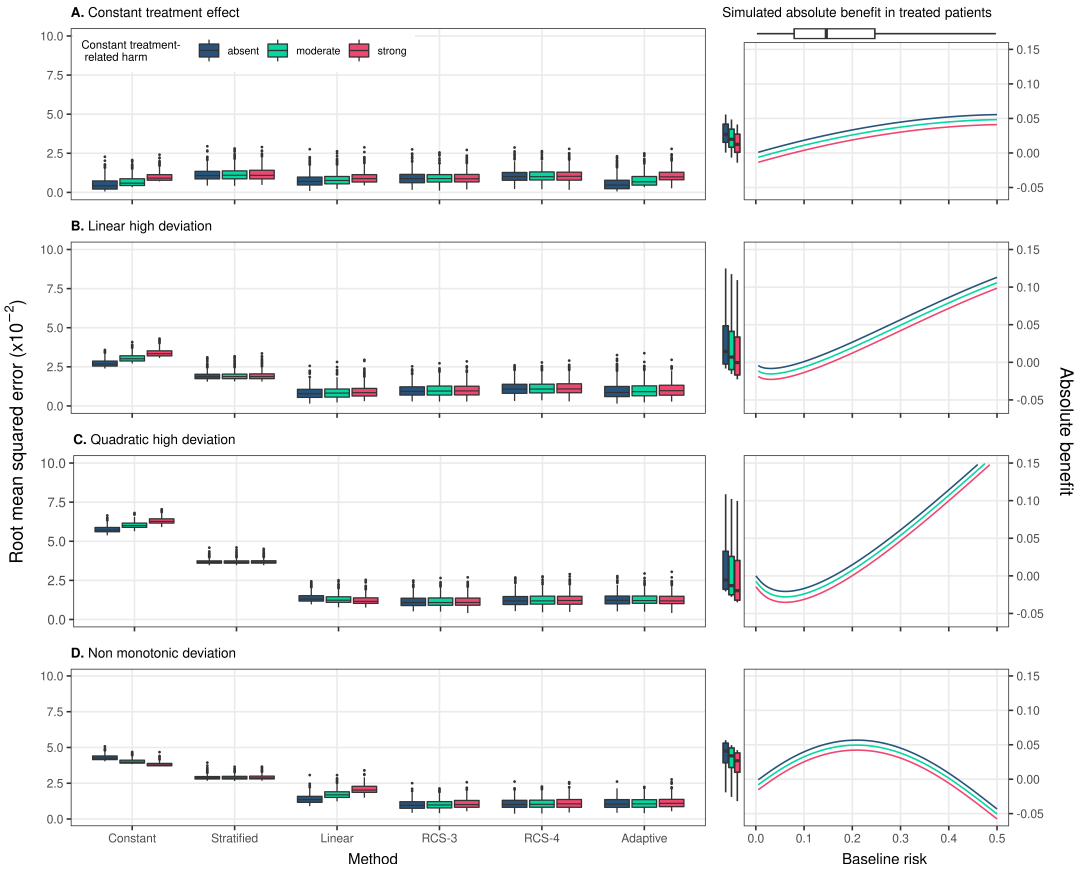

Figure 2: RMSE of the considered methods across 500 replications calculated in simulated samples of size 17,000 rather than 4,250 in Figure 1. RMSE was calculated on a super-population of size 500,000

method in each scenario. However, compared to the best-performing approach, its RMSE was more variable in scenarios with linear and non-monotonic deviations, especially when also including moderate or strong treatment-related harms. On closer inspection, we found that this behavior was caused by selecting the constant treatment effect model in a substantial proportion of the replications (Supplement, Figure S3).

Increasing the sample size to 17,000 favored RCS-3 the most. The difference in performance with the linear interaction approach was more limited in settings with a constant treatment effect (Median RMSE: linear interaction 0.007; RCS-3 0.009) and with a true linear interaction (Median RMSE: linear interaction 0.008; RCS-3 0.009). and more emphasized in settings with strong quadratic deviations (Median RMSE: linear interaction 0.013; RCS-3 0.011) and non-monotonic deviations (Median RMSE: linear interaction 0.014; RCS-3 0.010). Due to the large sample size, the RMSE of the adaptive approach was even more similar to the best-performing method, and the constant relative treatment effect model was less often wrongly selected (Supplement, Figure S4).

Similarly, when we increased the AUC of the true prediction model to 0.85 (OR = 0.8 and N = 4,250), RCS-3 had the lowest RMSE in the case of strong quadratic or non-monotonic deviations and very comparable performance to the – optimal – linear interaction model in the case of strong linear deviations (median RMSE

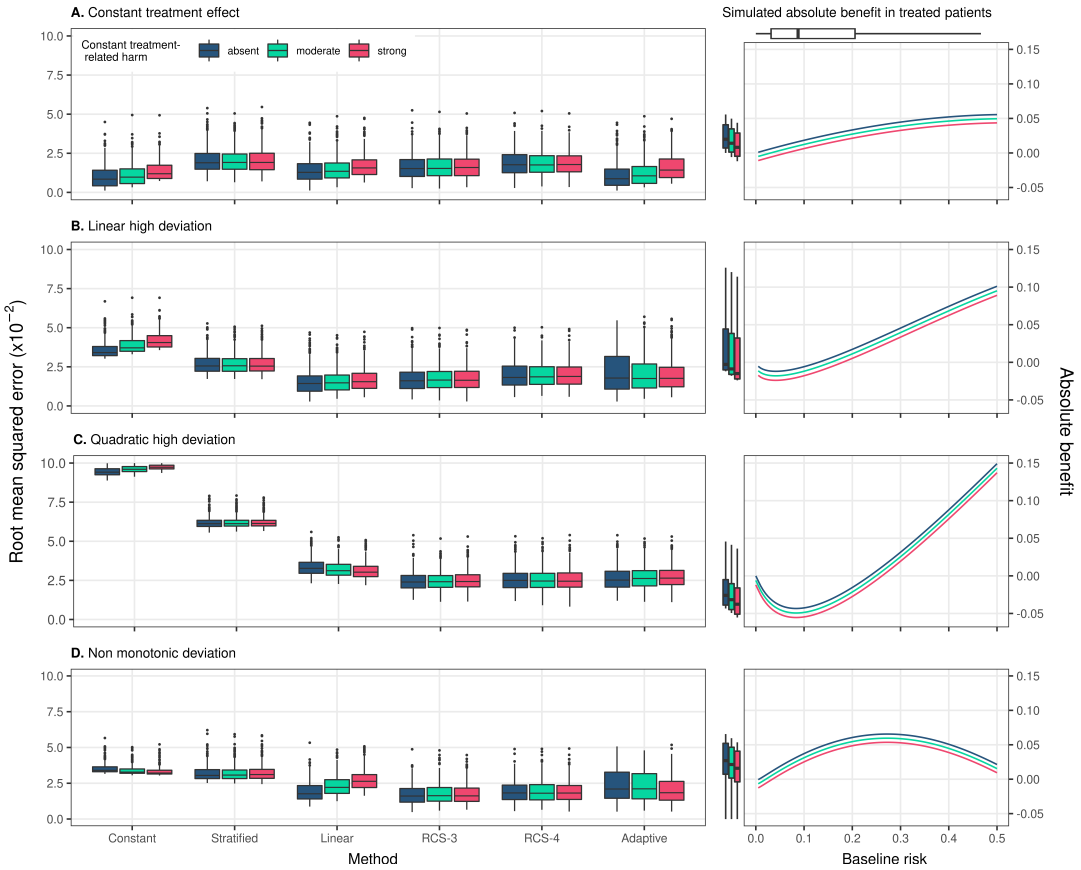

Figure 3: RMSE of the considered methods across 500 replications calculated in simulated samples 4,250. True prediction AUC of 0.85. RMSE was calculated on a super-population of size 500,000

of 0.016 for RCS-3 compared to 0.014 for the linear interaction model). Similar to the base case scenario the adaptive approach wrongly selected the constant treatment effect model (23% and 25% of the replications in the strong linear and non-monotonic deviation scenarios without treatment-related harms, respectively), leading to increased variability of the RMSE (Supplement, Figure S5).

With a true constant relative treatment effect, discrimination for benefit was only slightly lower for the linear interaction model, but substantially lower for the non-linear RCS approaches (Figure 4; panel A). With strong linear or quadratic deviations from a constant relative treatment effect, all methods discriminated quite similarly (Figure 4; panels B-C). With non-monotonic deviations, the constant effect model had much lower discriminative ability compared to all other methods (median AUC of 0.500 for the constant effects model, 0.528 for the linear interaction model and 0.530 Figure 4; panel D). The adaptive approach was unstable in terms of discrimination for benefit, especially with treatment-related harms. With increasing number of RCS knots, we observed decreasing median values and increasing variability of the c-for-benefit in all scenarios. When we increased the sample size to 17,000 we observed similar trends, however the performance of all methods was more stable (Supplement, Figure S6). Finally, when we increased the true prediction AUC to 0.85 the adaptive approach was, again, more

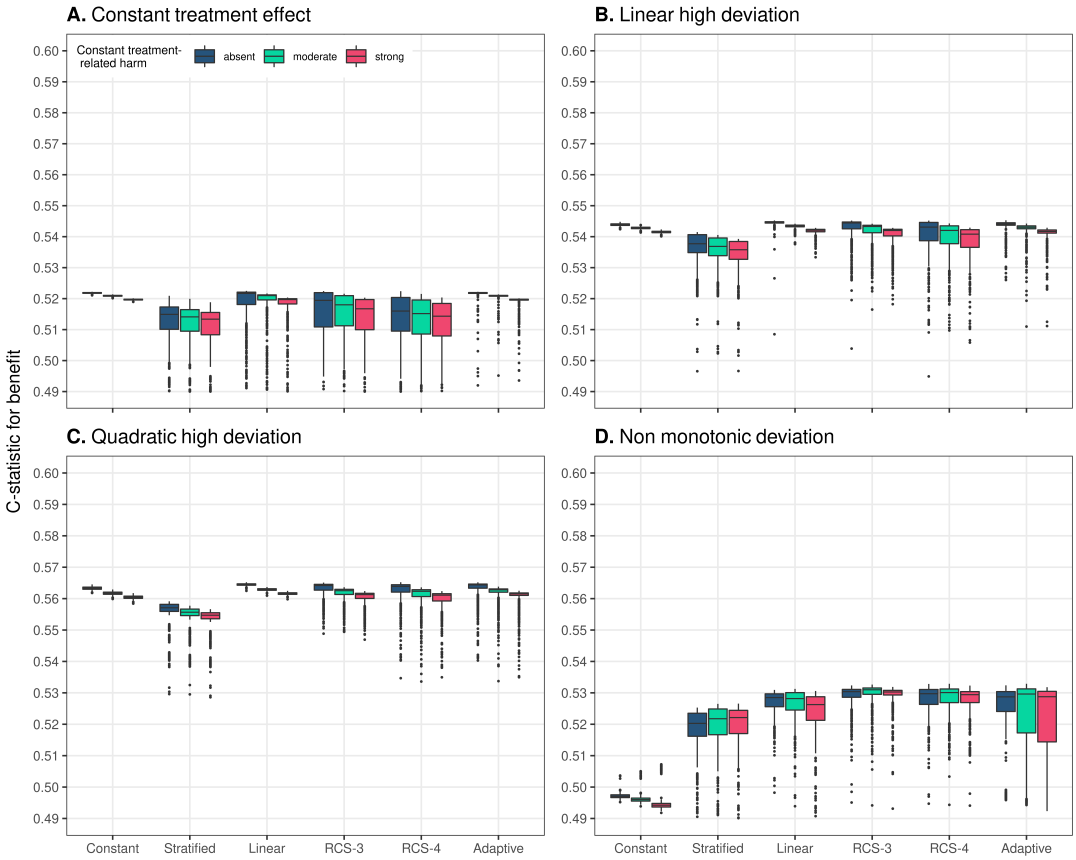

Figure 4: Discrimination for benefit of the considered methods across 500 replications calculated in a simulated samples of size 4,250. True prediction AUC of 0.75.

conservative, especially with non-monotonic deviations and null or moderate treatment-related harms (Supplement, Figure S5).

In terms of calibration for benefit, the constant effects model outperformed all other models in the scenario with 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 was only outperformed by RCS-3 in the case of non-monotonic deviations and treatment-related harms (Figure 5; panel D). The adaptive approach was worse calibrated under strong linear and non-monotonic deviations compared to the linear interaction model and RCS-3. When we increased the sample size to 17,000 (Supplement, Figure S6) or the true prediction AUC to 0.85 (Supplement, Figure S7), RCS-3 was somewhat better calibrated than the linear interaction model with strong quadratic deviations.

The results from all individual scenarios can be explored online at https://arekkas.shinyapps.io/simulation_viewer/. Additionally, all the code for the simulations can be found at https://github.com/rekkasa/arekkas_HteSimulation_XXXX_2021

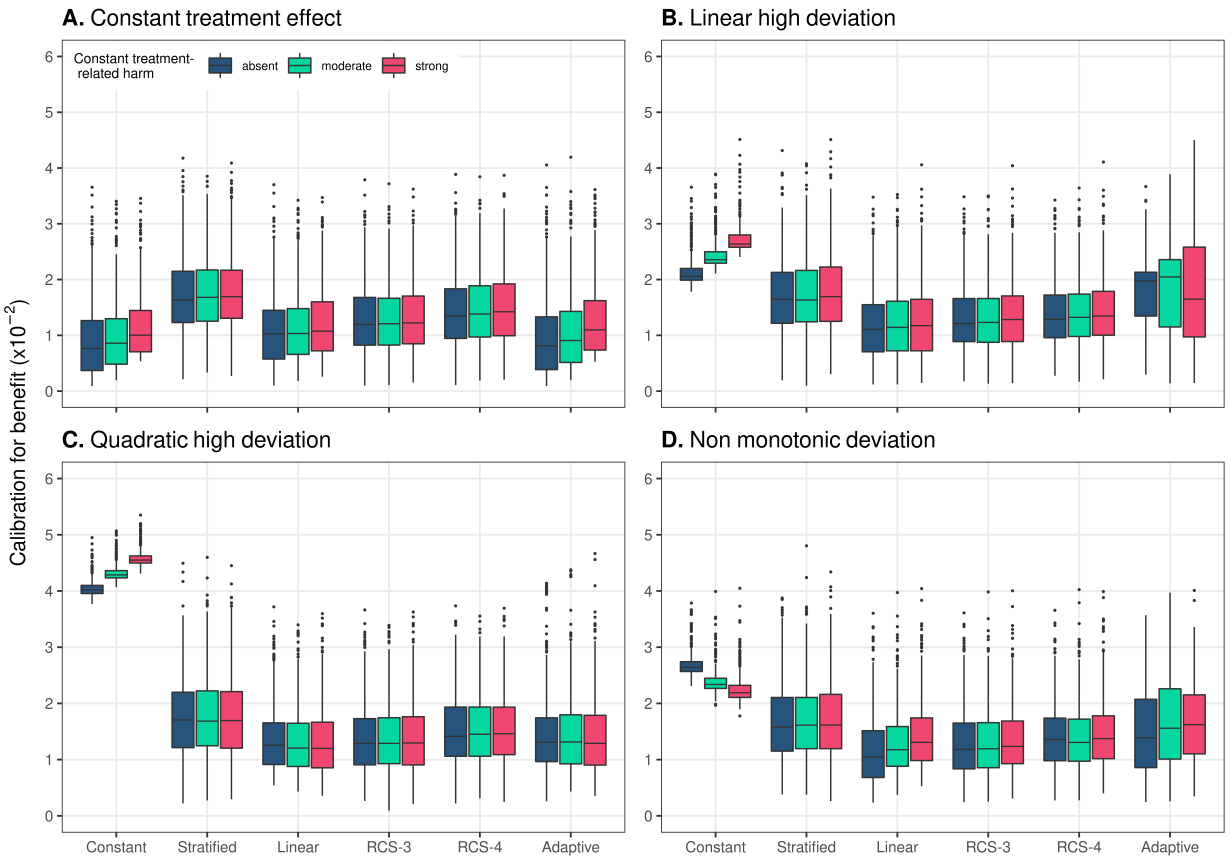

Figure 5: Calibration for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.75 and sample size of 4,250.

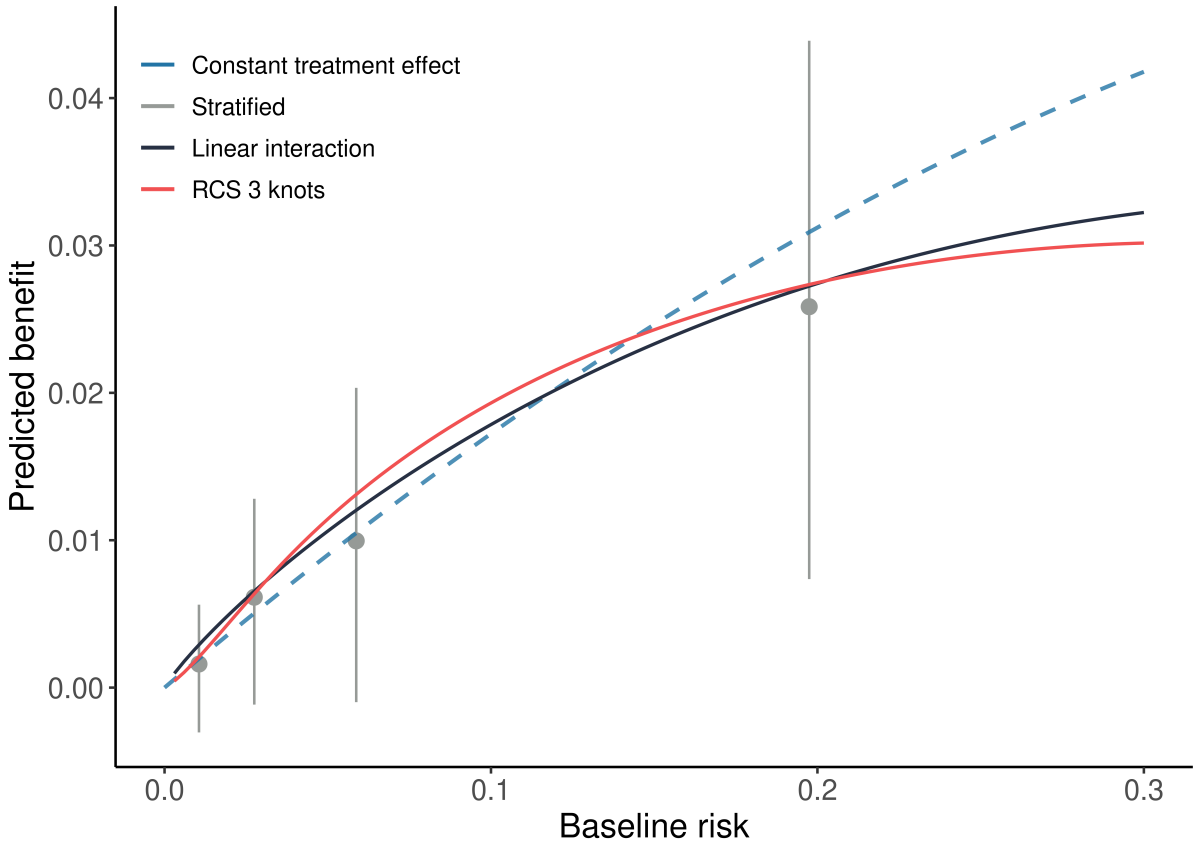

Figure 6: Individualized absolute benefit predictions based on baseline risk when using a constant treatment effect approach, a linear interaction approach and RCS smoothing using 3 knots. Risk stratified estimates of absolute benefit are presented within quartiles of baseline risk as reference.

3.2. Empirical illustration

We used the derived prognostic index to fit a constant treatment effect, a linear interaction and an RCS-3 model individualizing absolute benefit predictions. Following our simulation results, RCS-4 and RCS-5 models were excluded. Finally, an adaptive approach with the 3 candidate models was applied.

All considered methods provided similar fits, predicting increasing benefits for patients with higher baseline risk predictions, and followed the evolution of the stratified estimates closely (Figure 6). The constant treatment effect model had somewhat lower AIC compared to the linear interaction model (AIC: 9,336 versus 9,342), equal cross-validated discrimination (c-for-benefit: 0.525), and slightly better cross-validated calibration (ICI-for benefit: 0.010 versus 0.012). In conclusion, although the sample size (30,510 patients; 2,128 events) allowed for flexible modeling approaches, a simpler constant treatment effect model is adequate for predicting absolute 30-day mortality benefits of treatment with tPA in patients with acute MI.

4. Discussion

The linear interaction and the RCS-3 models displayed very good performance under many of the considered simulation scenarios. The linear interaction model was optimal in cases with moderate sample sizes (4.250 patients; ~785 events) and moderately performing baseline risk prediction models, that is, it had lower RMSE, was better calibrated for benefit and had better discrimination for benefit, even in scenarios with strong quadratic deviations. In scenarios with true non-monotonic deviations, the linear interaction model was outperformed by RCS-3, especially in the presence of treatment-related harms. Increasing the sample size or the prediction model's discriminative ability favored RCS-3, especially in scenarios with strong non-linear deviations from a constant treatment effect.

Our simulation results clearly express the trade-off between the advantages of flexibly modeling the relationship between baseline risk and treatment effect and the disadvantages of overfitting this relationship to the sample at hand. With infinite sample size, the more flexible approach (here RCS) will be optimal, but in practice, with limited sample size, parsimonious models may be preferable. Even with the substantial sample size of our base case scenario, the (less flexible) linear interaction model performed better than the (more flexible) RCS approach for most simulation settings. The even less flexible constant treatment effect model, however, was only optimal when the treatment effect was truly constant. Moreover, the assumption of a constant treatment effect may often be too strong [7,14]. For example, infants at lower risk of bronchopulmonary dysplasia benefit relatively more from vitamin A therapy than infants at higher risk [15]; higher risk prediabetic patients benefit relatively more from metformin than lower risk patients [16]. Hence, a linear interaction between baseline risk and the effect of treatment may be the most sensible approach with moderate sample sizes.

RCS-4 and RCS-5 were too flexible in all considered scenarios, 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, these more flexible methods did not outperform the simpler RCS-3 approach. Higher flexibility may only be helpful under more extreme patterns of HTE compared to the quadratic deviations considered here. Considering interactions in RCS-3 models as the most complex approach often may be reasonable.

Increasing the discriminative ability of the risk model reduced RMSE for all methods. Higher discrimination translates in higher variability of predicted risks, which, in turn, allows the considered methods to better capture absolute treatment benefits. As a consequence, better risk discrimination also led to higher discrimination between those with low or high benefit (as reflected in values of c-for-benefit).

The adaptive approach had adequate median performance, following the "true" model in most scenarios. With smaller sample sizes it tended to miss the treatment-baseline risk interaction and selected simpler models (Supplement Section 4). This conservative behavior resulted in increased RMSE variability in these scenarios, especially with true strong linear or non-monotonic deviations. Therefore, with smaller sample sizes the simpler

linear interaction model may be a safer choice for predicting absolute benefits, especially in the presence of any suspected treatment-related harms.

One limitation is that we assumed treatment benefit to be a function of baseline risk in the majority of the simulation scenarios. We attempted to expand our scenarios by considering constant moderate and strong treatment-related harms, applied on the absolute scale, in line with previous work [17]. In a limited set of scenarios with true interactions between treatment assignment and covariates, our conclusions remained unchanged (Supplement, Section 7). Even though the average error rates increased for all the considered methods, due to the miss-specification of the outcome model, the linear interaction model had the lowest error rates. RCS-3 had very comparable performance. The constant treatment effect model was often biased, especially with moderate or strong treatment-related harms. Future simulation studies could explore the effect of more extensive deviations from risk-based treatment effects.

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 [18–21]. Many of these methods rely on incorporating treatment-covariate interactions when predicting benefit. An important caveat of such approaches is their sensitivity to overfitting, which may exaggerate the magnitude of predicted benefits. In a wide range of simulation settings, a simpler risk modeling approach was consistently better calibrated for benefit compared to more complex treatment effect modelling approaches [5]. Similarly, when SYNTAX score II, a model developed for identifying patients with complex coronary artery disease that benefit more from percutaneous coronary intervention or from coronary artery bypass grafting was redeveloped using fewer treatment-covariate interactions had better external performance compared to its predecessor [22,23]. However, whether this remains the case in a range of empirical settings still needs to be explored.

In conclusion, the linear interaction approach is a viable option with moderate sample sizes and/or moderately performing risk prediction models, assuming a non-constant relative treatment effect plausible. RCS-3 is a better option with more abundant sample size and when non-monotonic deviations from a constant relative treatment effect and/or substantial treatment-related harms are anticipated. Increasing the complexity of the RCS models by increasing the number of knots does not improve benefit prediction. Using AIC for model selection is attractive with larger sample size.

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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Rothwell PM. Can overall results of clinical trials be applied to all patients? The Lancet 1995;345:1616–9. https://doi.org/10.1016/s0140-6736(95)90120-5.

[15] Rysavy MA, Li L, Tyson JE, Jensen EA, Das A, Ambalavanan N, et al. Should vitamin a injections to prevent bronchopulmonary dysplasia or death be reserved for high-risk infants? Reanalysis of the national institute of child health and human development neonatal research network randomized trial. The Journal of Pediatrics 2021;236:78–85.e5. https://doi.org/10.1016/j.jpeds.2021.05.022.

[16] Sussman JB, Kent DM, Nelson JP, Hayward RA. Improving diabetes prevention with benefit based tailored treatment: Risk based reanalysis of diabetes prevention program. BMJ 2015;350:h454–4. https://doi.org/10.1136/bmj.h454.

[17] Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. BMJ 1995;311:1356–9. https://doi.org/10.1136/bmj.311.7016.1356.

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

[19] 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.

[20] 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.

[21] Powers S, Qian J, Jung K, Schuler A, Shah NH, Hastie T, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. Statistics in Medicine 2018;37:1767–87.

[22] Farooq V, Van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, et al. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: Development and validation of syntax score ii. The Lancet 2013;381:639–50.

[23] Takahashi K, Serruys PW, Fuster V, Farkouh ME, Spertus JA, Cohen DJ, et al. Redevelopment and validation of the syntax score ii to individualise decision making between percutaneous and surgical revascularisation in patients with complex coronary artery disease: Secondary analysis of the multicentre randomised controlled

syntaxes trial with external cohort validation. The Lancet 2020;396:1399–412.