Journal of Clinical Epidemiology 114 (2019) 72−83

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting

David van Klaveren[a,b,*,1], Theodor A. Balan[c,1], Ewout W. Steyerberg[a,c], David M. Kent[b]

[a]*Department of Public Health, Erasmus University Medical Center, Rotterdam, the Netherlands*

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

[c]*Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands*

Accepted 24 May 2019; Published online 10 June 2019

Abstract

Objectives: We aimed to compare the performance of different regression modeling approaches for the prediction of heterogeneous treatment effects.

Study Design and Setting: We simulated trial samples (n = 3,600; 80% power for a treatment odds ratio of 0.8) from a superpopulation (N = 1,000,000) with 12 binary risk predictors, both without and with six true treatment interactions. We assessed predictions of treatment benefit for four regression models: a ''risk model'' (with a constant effect of treatment assignment) and three ''effect models'' (including interactions of risk predictors with treatment assignment). Three novel performance measures were evaluated: calibration for benefit (i.e., observed vs. predicted risk difference in treated vs. untreated), discrimination for benefit, and prediction error for benefit.

Results: The risk modeling approach was well-calibrated for benefit, whereas effect models were consistently overfit, even with doubled sample sizes. Penalized regression reduced miscalibration of the effect models considerably. In terms of discrimination and prediction error, the risk modeling approach was superior in the absence of true treatment effect interactions, whereas penalized regression was optimal in the presence of true treatment interactions.

Conclusion: A risk modeling approach yields models consistently well calibrated for benefit. Effect modeling may improve discrimination for benefit in the presence of true interactions but is prone to overfitting. Hence, effect models—including only plausible interactions—should be fitted using penalized regression.

Keywords: Personalized medicine; Heterogeneity of treatment effect; Treatment benefit; Prediction models; Regression analysis; Penalized regression analysis

1. Background

The main goal of predictive analyses of heterogeneous treatment effects is to develop models that can be used to predict which of two or more treatments will be better for a particular individual [1]. In prior work, we discuss two broad approaches to regression-based predictive heterogeneity of treatment effects (HTE) analysis: (1) outcome risk modeling (or ''risk modeling'') and (2) treatment effect modeling (or ''effect modeling'') [2]. Risk modeling is performed ''blinded'' to treatment assignment, where the treatment effect is examined across risk strata; treatment effect interactions with candidate relative effect modifiers are not explored. This conservative approach is designed to produce honest estimates of treatment effects within groups that differ by their risk of the outcome of interest and therefore by their risk difference (i.e., treatment effect on the clinically important absolute scale). Because risk is a mathematical determinant of treatment effect, it can be used to identify risk groups who differ greatly in their degree of benefit (particularly on the clinically important absolute scale) [3,4]. Treatment effect modeling, on the other hand, is a more ''aggressive'' data-driven approach that seeks to explore and/or include potential relative effect modifiers

Funding sources: This work was partially supported through two Patient-Centered Outcomes Research Institute (PCORI) Awards: the Predictive Analytics Resource Center (PARC) (SA.Tufts.PARC.OSCO.2018.01.25) and the PCORI Methods Award (ME-1606-35555).

Disclosure: All statements in this work, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute, its Board of Governors, or Methodology Committee.

[1] These authors contributed equally to this article.

* Corresponding author. Tel.: +31 10 704 35 02; fax: +31 10 703 84 75.

E-mail address: d.vanklaveren@erasmusmc.nl (D. van Klaveren).

https://doi.org/10.1016/j.jclinepi.2019.05.029

What is new?

Key findings

- The risk modeling approach was well calibrated for benefit in contrast with effect models, which were consistently miscalibrated (''overfit'').
- Lasso regression considerably reduced but did not nullify benefit miscalibration of the effect models.
- In terms of discrimination and prediction error, the risk modeling approach was superior in the absence, whereas penalized regression was optimal in the presence of true treatment interactions, respectively.

What this adds to what was known?

- Effect models overestimate the heterogeneity of treatment effects considerably, even for large sample sizes.
- Penalized regression prevents overfitting of risk but still suffers from overfitting of benefit even when the events per variable are high.

What is the implication and what should change now?

- Effect models, including treatment interactions, may lead to treatment mistargeting.
- We recommend the risk modeling approach for the analysis of heterogeneity of treatment effects.
- Effect models should only be considered when treatment interactions are plausible and should always be fitted with penalized regression.

within a predictive model. This approach has the potential to better segregate those patients who benefit from those who do not. Particularly when treatment is associated with some treatment-related harm or burden, these approaches may be useful in targeting treatment to those most likely to benefit.

We performed simulations to compare the various approaches to predictive HTE analysis and illustrate and quantify some of the problems with overfitting for models used to predict benefit that are not widely appreciated. For risk modeling, we sought to illustrate problems that arise when an endogenous model is developed on the control arm only. Although this problem has been shown in previous simulations [5,6], we sought to use data generated with a substantially different set of assumptions to test the robustness of prior observations, especially those indicating that modeling on the whole trial eliminated the bias induced by modeling only on the control arm.

We also sought to examine how risk modeling compared with several different effect modeling approaches, using both measures of calibration for benefit and discrimination for benefit—where benefit is defined as the risk difference. These simulations permitted us to explore how serious the problems of overfitting were when using ''conventional'' nonpenalized maximum likelihood regression for effect modeling in settings that emulate the statistical power likely to be available to explore heterogeneous effects in large clinical trials. Finally, we also sought to examine the ability of penalized regression (i.e., Lasso and Ridge regression) to correct for the overfitting anticipated when ''conventional'' nonpenalized maximum likelihood regression was used for effect modeling.

2. Methods

2.1. Population

The population consists of 1 million patients equally divided into a control arm and a treatment arm. For each patient, 12 binary baseline characteristics are independently generated with a prevalence of 20% each. The outcomes in the control arm are generated with an average event rate of approximately 25% from a logistic regression model with associations between baseline characteristics and outcomes varying from an odds ratio (OR) of 1−2 (Table 1; ''Control arm''). The outcomes for patients in the treatment arm were first generated from the same logistic regression model (Table 1; ''Treatment arm, without interactions''), but including a main treatment effect OR of either 1 (''null''), 0.8 (''moderate''; we considered this the ''base case''), or 0.5 (''strong''), respectively. Alternatively, the outcomes for patients in the treatment arm were generated from a logistic regression model in which 6 of the 12 associations between baseline characteristics and outcomes differed from the control arm (Table 1; ''Treatment arm, with interactions''), with relative ORs varying from 0.67 to 1.33. The prevalence of true interactions among all possible interactions (50%) is designed to emulate the ''prior probability'' for confirmatory subgroup analysis, whereas the prior probability for exploratory subgroup analysis would be considerably lower [7−9]. To analyze the sensitivity of the results to the use of continuous rather than binary baseline characteristics, we repeated the analysis with 12 normally distributed baseline characteristics. We chose the same mean (0.2) and variance (0.16) as for the binary baseline characteristics to obtain similar precision when estimating the effects of baseline characteristics and the effects of their interaction with treatment. To analyze the sensitivity of the results to correlation between baseline characteristics, we repeated the analysis with 12 correlated normally distributed baseline characteristics (correlation coefficients of 0.5).

Table 1. Associations between the outcome and population (N = 1,000,000) characteristics x_1 to x_{12}

	Control arm	Treatment arm, without interactions		Treatment arm, with interactions	
Variable	OR_C	OR_{TR}	OR_{TR}/OR_C	OR_{TR}	OR_{TR}/OR_C
x_1	1	1	1	1	1
x_2	1	1	1	1	1
x_3	1	1	1	1	1
x_4	1.2	1.2	1	1.4	**1.17**
x_5	1.2	1.2	1	1.2	1
x_6	1.2	1.2	1	1	**0.83**
x_7	1.5	1.5	1	2	**1.33**
x_8	1.5	1.5	1	1.5	1
x_9	1.5	1.5	1	1	**0.67**
x_{10}	2	2	1	2.5	**1.25**
x_{11}	2	2	1	2	1
x_{12}	2	2	1	1.5	**0.75**

Abbreviation: OR, odds ratio.

Outcomes are simulated from a logistic regression model.

For each patient, 12 binary baseline characteristics were independently generated with a prevalence of 20% each. The outcomes of the 500,000 patients in the control arm (average event rate of 25%) are generated from a logistic regression model with associations between baseline characteristics and outcomes according to column 2 (''control arm''). The outcomes for patients in the treatment were first generated from the same logistic regression model (column 3; ''Treatment arm, without interactions''), but including a main treatment effect odds ratio of either 1 (''null''), 0.8 (''moderate''), or 0.5 (''strong''), respectively. Alternatively the outcomes for patients in the treatment arm were generated from a different logistic regression model (column 5; ''Treatment arm, with interactions'').

Treatment interactions are in bold font.

2.2. Clinical example

We illustrate the simulated population with the clinical example of choosing between treatment of patients with complex coronary artery disease with either Coronary Artery Bypass Graft (CABG) surgery (''treatment'') or Percutaneous Coronary Intervention (PCI; ''control''). Evidence suggests that, on average, treatment with CABG leads to lower mortality than treatment with PCI. Baseline characteristics that are associated with high mortality risk are high SYNTAX score (high anatomic disease complexity), high age, low creatinine clearance, low left ventricular ejection fraction, presence of unprotected left main coronary artery disease, peripheral vascular disease, female sex, chronic obstructive pulmonary disease (COPD), and diabetes mellitus. The simulation scenario without true treatment interaction assumes that these baseline characteristics increase the odds of mortality equally after both treatments, for example, COPD doubles the odds of mortality, regardless of treatment. The simulation scenario with true treatment interaction assumes that some of the associations between the baseline characteristics and the odds of mortality are different after treatment with CABG and treatment with PCI, for example, COPD increases the odds of mortality after PCI with 50%, but doubles the odds of mortality after CABG.

2.3. Samples

A simulation consists of 500 random samples of 3,600 patients (80% power to detect a treatment vs. control OR of 0.8) from the population. To study the impact of sample size, we alternatively sampled 1,800 patients and 7,200 patients per sample, respectively.

2.4. Risk models

In each sample of patients, we fitted a risk model consisting of all of the 12 baseline characteristics. This model was fitted in the patients in the control arm only, as well as in the whole sample, blinded to treatment. For each of these risk models, the predicted risk was calculated for each patient in the sample and the population. The sample and the population were stratified into quartiles of predicted risk (according to the sample-derived risk model) and the observed absolute treatment benefit (i.e., the event rate of control patients minus the event rate of treated patients) was determined in each risk quartile of the sample and the population. To compare these approaches (control arm only vs. whole trial), within-strata event rates in the population were used to represent ''ground truth.'' We assessed the ''calibration for benefit'' of these approaches by examining the observed absolute treatment benefit (i.e., risk difference) in quartiles of predicted risk in the sample vs. that in the population. We present box plots describing the distribution of these values across the 500 samples.

2.5. Treatment effect models

In each sample of patients, we first fitted:

1. a model consisting of all of the 12 baseline characteristics and the treatment, assuming a constant relative treatment effect. This approach is equivalent to a risk modeling approach, but with a constant relative effect imposed across risk groups; this was done to generate values for predicted benefit—predicted outcome with vs. without treatment—so that the approach can be compared with the other effect modeling approaches below.

Then we fitted different treatment effect models in each sample of patients consisting of all of the 12 baseline characteristics and

2. two prespecified treatment interactions with baseline characteristics x_{10} and x_{12}, corresponding to a parsimonious and non−data-driven approach including only of two ''established'' relative effect modifiers
3. treatment interactions with all 12 baseline characteristics
4. treatment interactions with all 12 baseline characteristics, using backward selection based on Akaike information criterion (AIC; $P = 0.157$).
2. treatment interactions with all 12 baseline characteristics, using Lasso regression with minimum mean cross-validated error (referred to as Standard Lasso).
3. treatment interactions with all 12 baseline characteristics, using Ridge regression with minimum mean cross-validated error (referred to as Standard Ridge).
4. treatment interactions with all 12 baseline characteristics, using Lasso regression with maximum penalization such that the error is within one standard error of the minimum (referred to as Strong Lasso).
5. treatment interactions with all 12 baseline characteristics, using Ridge regression with maximum penalization such that the error is within one standard error of the minimum (referred to as Strong Ridge).

For each effect model that was fitted in each consecutive sample, the predicted absolute treatment benefit (i.e., predicted risk difference) in the population was calculated as the model's outcome prediction conditional on assignment to the control arm minus the outcome prediction conditional on assignment to the treatment arm. The population was stratified into quartiles of predicted absolute treatment benefit. In each quartile of the population, we calculated the mean predicted absolute treatment benefit and the observed absolute treatment benefit, that is, the event rate of control patients minus the event rate of treated patients.

2.7. Metrics of model performance

2.7.1. Calibration for benefit

A well-calibrated sample-based model fit would have good agreement in the population between predicted and observed benefit in quartiles of predicted benefit. We assessed the calibration for the benefit of each model fit by the difference between predicted absolute treatment benefit and observed absolute treatment benefit in quartiles of predicted absolute benefit of the population. We present box plots of the observed vs. predicted treatment benefit in quartiles of predicted benefit for each model describing the distribution of these values across the 500 samples. Because calibration in the second and third quartiles of predicted benefit is always excellent, we numerically express calibration by the mean difference between the predicted treatment benefit and the observed treatment benefit in the two extreme predicted benefit quartiles (i.e., lowest and highest quartiles) of the population.

2.7.2. Discrimination for benefit

A well-discriminating sample-based model fit would have a large difference between the observed benefit in the first and fourth predicted benefit quartiles of the population. We assessed the discrimination of each model fit by the extreme quartile difference (EQD) of benefit in the population, that is, the arithmetic difference between the observed absolute benefit in the fourth quartile and the observed absolute benefit in the first quartile of the population. A model that shows a higher EQD in the population is better able to discriminate patients for treatment benefit, that is, to separate patients with low treatment benefit from those with high treatment benefit.

2.7.3. Prediction error for benefit

The simulation settings (Table 1) allow the calculation of true individual treatment benefit for each individual patient in the population, which is unidentifiable in real-world settings. We were thus able to assess the accuracy of each model's benefit predictions by the root mean squared error (rMSE), that is, the root of the mean of the square differences between predicted absolute benefit and true absolute benefit in the population.

3. Results

3.1. Simulated population

The discriminative ability of the true risk model in the control arm of the population was moderate with an area under the receiver operating characteristic curve of 0.66. The event rates in the control arm were 14.7%, 21.2%, 28.9%, and 43.3% in true risk quartiles 1−4, respectively. With a main treatment effect OR of 0.8 (the ''base case''), these event rates were reduced by absolute treatment benefits of 2.4%, 3.3%, 4.2%, and 5.1%, respectively. With a main treatment effect OR of 0.5, these event rates were reduced by absolute treatment benefits of 6.8%, 9.2%, 12.0%, and 15.4%, respectively.

3.2. Risk models

When the risk model was fitted in the whole sample, the benefit in quartiles of predicted risk in the sample was well calibrated to the ''true'' benefit in quartiles of predicted risk in the population (Fig. 1A). This was true regardless of the sample size (Fig. A.1A), the treatment effect magnitude, and whether the data were generated by a regression model with or without treatment interactions (Figs. A.1A and A.2A). In contrast, when the risk model was fitted in only the control patients of the sample, the benefit in quartiles of predicted risk was more heterogeneous in the sample compared with the population, that is, the benefit in the first risk quartile was substantially lower in the sample than in the population, whereas the benefit in the fourth risk quartile was substantially higher in the sample than in the population, reflecting differential fit (i.e., overfitting) of the model on the sample control arm compared with the treatment arm. For example, in the base case scenario (treatment OR of 0.8; sample size 3,600, Fig. 1B), the median benefit in the highest risk fourth quartile (6.5%) of the sample was 30% higher than the median benefit in this quartile (5.1%) of the population. Similarly, the median benefit in the lowest risk first quartile (1.7%) of the sample was 30% lower than the median benefit in this quartile (2.4%) of the population. Even where there was no treatment effect whatsoever (treatment OR of 1; Figs. A.1B and A.2B), modeling in the control arm induced a spurious risk-by-treatment effect interaction, which was more often statistically significant (in 9.6% of the samples) than expected under the null condition of no treatment effect interaction (5%). This effect attenuated but persisted even when the sample was very large (n = 7,200; Figs A.1B and A.2B), with an event-per-variable (EPV) ratio of 81.

3.3. Treatment effect models

3.3.1. General

Using backward selection in the base case scenario (treatment OR of 0.8; sample size 3,600), 90% of the true prognostic factors and 48% of the true interactions were kept in the model (true positive), whereas 29% of the null prognostic factors and 17% of the null interactions were kept in the model (false positive; Table A.1). With Lasso regression, 95% of the true prognostic factor effects and 75% of the true interaction effects were kept in the model (true positive), whereas 55% of the null prognostic factor effects and 54% of the null interaction effects were kept in the model (false positive). As expected, for backward selection, the true positive rate increased with sample size with a stable false positive rate. For Lasso regression, both the true and the false positive rate increased with sample size, but the effect sizes of the retained variables were shrunk.

3.3.2. Calibration for benefit

In the absence of true treatment interactions, the model with a constant relative treatment effect was well calibrated (Fig. 2A), whereas all the models that included treatment interactions overfitted treatment benefit (Fig. 2B–D;

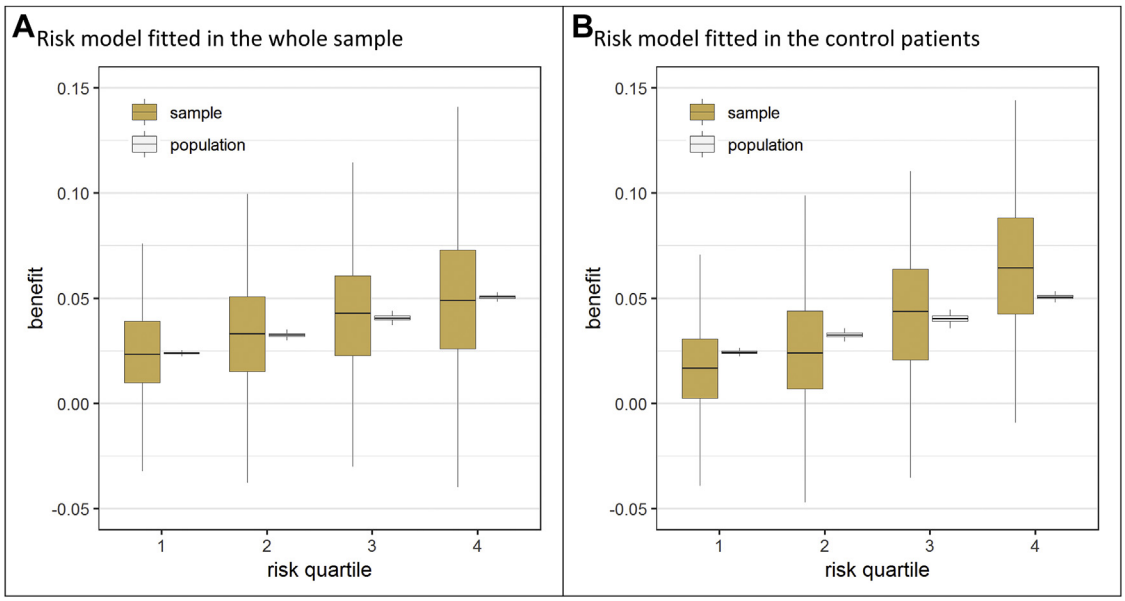

Fig. 1. Risk modeling on only the control patients led to an exaggeration of the treatment effect heterogeneity across predicted risk quartiles. In the base case simulation scenario without true interaction, the risk model was fitted in either the whole sample or in the control patients of the sample. When the model fitted on the whole sample was used for stratification in risk quartiles (panel A), the observed benefit in the sample (brown bars) is an unbiased estimate of the observed benefit in the population (white bars). In contrast, when the model fitted on the control patients was used for stratification in risk quartiles (panel B), the observed benefit in the sample is too heterogeneous across risk quartiles compared with the observed benefit in the population. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

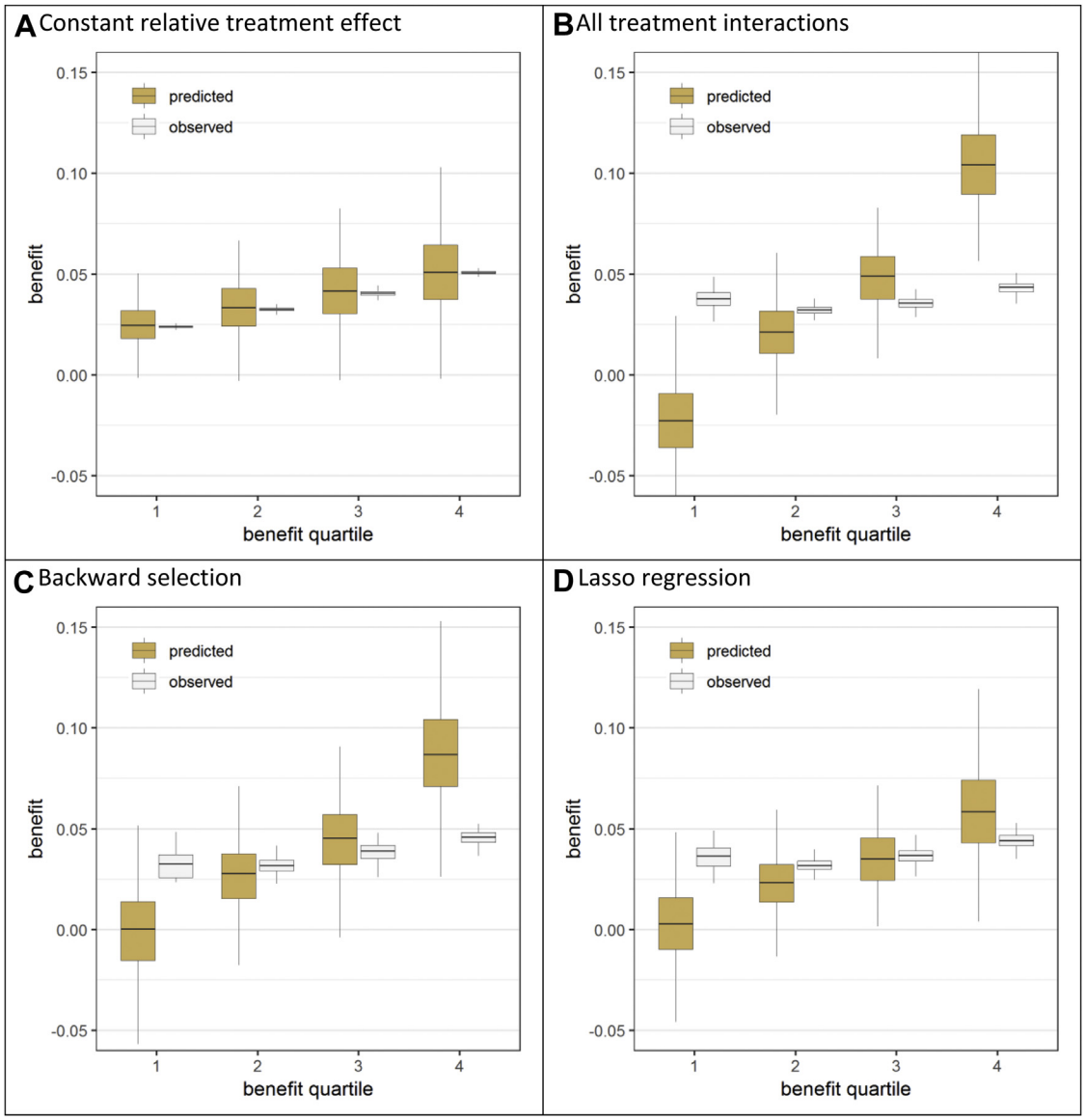

Fig. 2. Effect modeling approaches were seriously overfit and were prone to treatment mistargeting in the absence of true treatment interaction Benefit predictions were based on different models fitted in the samples: a model without treatment interactions (panel A), a model with all treatment interactions (panel B), a model with all treatment interactions using backward selection based on AIC (panel C), and a model with all treatment interactions fitted with Lasso regression (panel D). The agreement between predicted (brown bars) and observed (white bars) benefit in predicted benefit quartiles of the population was better for the risk modeling approach (A) compared with the effect modeling approaches (B–D). Moreover, the risk modeling approach (A) resulted in more heterogeneity of observed benefit, that is, is better able to distinguish between patients with low and patients with high benefit. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Fig. A.3). In the base case scenario (treatment OR of 0.8; sample size 3,600), the mean difference between predicted and observed benefit (i.e., miscalibration), averaged over the two extreme predicted benefit quartiles, was 6.1% for the model with all interactions (3.8% after backward selection) but was reduced by 60–2.5% when that model was fitted with Lasso regression (Fig. A.3D). For the population generated from the "hard null" assumptions (no treatment effect, no interactions), the models spuriously identified groups with treatment-related harm and benefit. The mean difference between predicted and observed benefit, averaged over the two extreme predicted benefit quartiles, was even larger than in the base case scenario (6.4% for the model with all interactions, 4.4% after backward selection, and 3.0% with Lasso regression; Fig. A.4).

In the presence of true treatment interactions, all models overfitted treatment benefit, but to a varying extent (Fig. 3; Fig. A.5). Overfitting increased with the number of modeled treatment interactions. In the base case scenario (treatment OR of 0.8; sample size 3,600), the mean difference between predicted and observed benefit, averaged over the two extreme predicted benefit quartiles, was only 0.9%

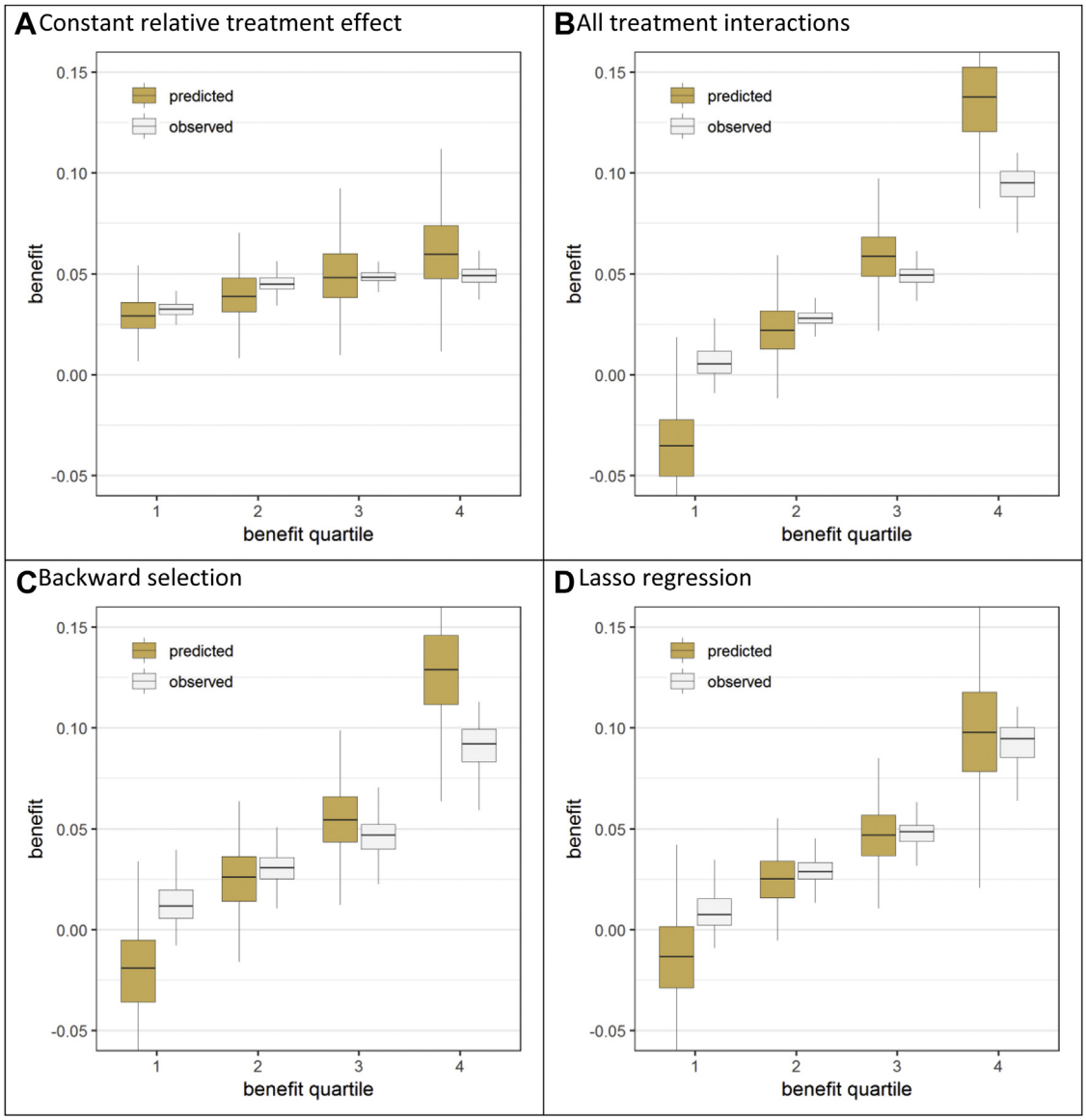

Fig. 3. Effect modeling approaches were better able to separate patients with differential benefit in the presence of true treatment interaction, but required penalized regression to prevent overfitting. Benefit predictions were based on different models fitted in the samples: a model without treatment interactions (panel A), a model with all treatment interactions (panel B), a model with all treatment interactions using backward selection based on AIC (panel C), and a model with all treatment interactions fitted with Lasso regression (panel D). The agreement between predicted benefit (brown bars) and observed benefit (white bars) in predicted benefit quartiles of the population is better for both the risk modeling approach (A) and the Lasso regression approach (D) compared with the unpenalized effect modeling approaches (B–C). However, the Lasso regression approach (D) resulted in more heterogeneity of observed benefit than the risk modeling approach (A), that is, is better able to distinguish between patients with low and patients with high benefit. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

for the model with a constant relative treatment effect and 4.4% for the model with all interactions. Backward selection only slightly reduced the overfitting (3.5% difference). In contrast, Lasso regression considerably reduced, but did not eliminate, the overfitting (1.5% difference), even for the larger sample size (Figs. A.6–A.9). In addition, even when just two "established" predictors were forced into the model without any statistically driven model selection, overfitting for benefit prediction persisted (1.0% difference). Overfitting for the prediction of benefit was relatively much stronger than overfitting for outcome prediction (Fig. A.10). Finally, overfitting for the prediction of benefit also persisted although Lasso regression with minimum mean cross-validated error essentially eliminated miscalibration for outcome prediction (Fig. A.11).

Models that were fitted with Lasso regression performed consistently better than models that were fitted with Ridge regression, especially in the absence of true treatment interaction (Figs. A.7 and A.9). Furthermore, models that were fitted using penalized regression with minimum mean

cross-validated error (“Standard Lasso”) consistently outperformed models that were fitted using penalized regression with maximum penalization, such that the error is within one standard error of the minimum (“Strong Lasso”). In the remainder of the results, we will therefore limit the presentation of penalized regression models to Lasso regression with minimum mean cross-validated error (“Standard Lasso”).

3.3.3. Discrimination for benefit

In the absence of true treatment interactions, the EQD of benefit in the population of the model with the constant relative treatment effect (2.7% in the base case scenario) was almost equal to the potential maximum as defined by the true model (Fig. 4). This means that we observe 2.7% more absolute treatment benefit in the 25% of patients with the highest compared with the 25% of patients with the lowest predicted benefit according to the model; this corresponds to more than a doubling of the effect size. Despite the fact that models with interaction terms misleadingly appeared to have substantially higher EQDs within the sample (indicating better discrimination for benefit), the population EQD decreased as the number of interactions in the model increased, indicating worse discrimination than risk modeling. Lasso regression moderately improved the population discrimination (0.8% EQD) compared with conventional maximum likelihood regression (0.6% EQD) but did not reach the potential maximum (2.7% EQD), even for the largest sample size (Fig. A.12).

In the presence of true treatment interactions, the population EQD of the model with a constant relative treatment effect was lowest (1.7% in the base case scenario), representing poor benefit discrimination. In the presence of true treatment interactions, Lasso regression showed very similar discriminative ability compared with conventional maximum likelihood regression (population EQD of 8.6% vs. 9.0% in the base case scenario) and somewhat better discriminative ability than backward selection (population EQD of 8.1%). The difference in population EQD between the model with a constant relative treatment effect and other models was limited when the sample size was low (1,800), and the treatment effect was large (OR 0.5; Fig. A.12).

3.3.4. Prediction error for benefit

In the absence of true treatment interactions, the rMSE of the individual estimated treatment benefits vs. the true individual treatment benefits increased with the number of modeled treatment interactions (Fig. 5). In the base case scenario (treatment OR of 0.8; sample size 3,600), the rMSE was 0.013 for the model with a constant relative treatment effect and 0.053 for the model with all

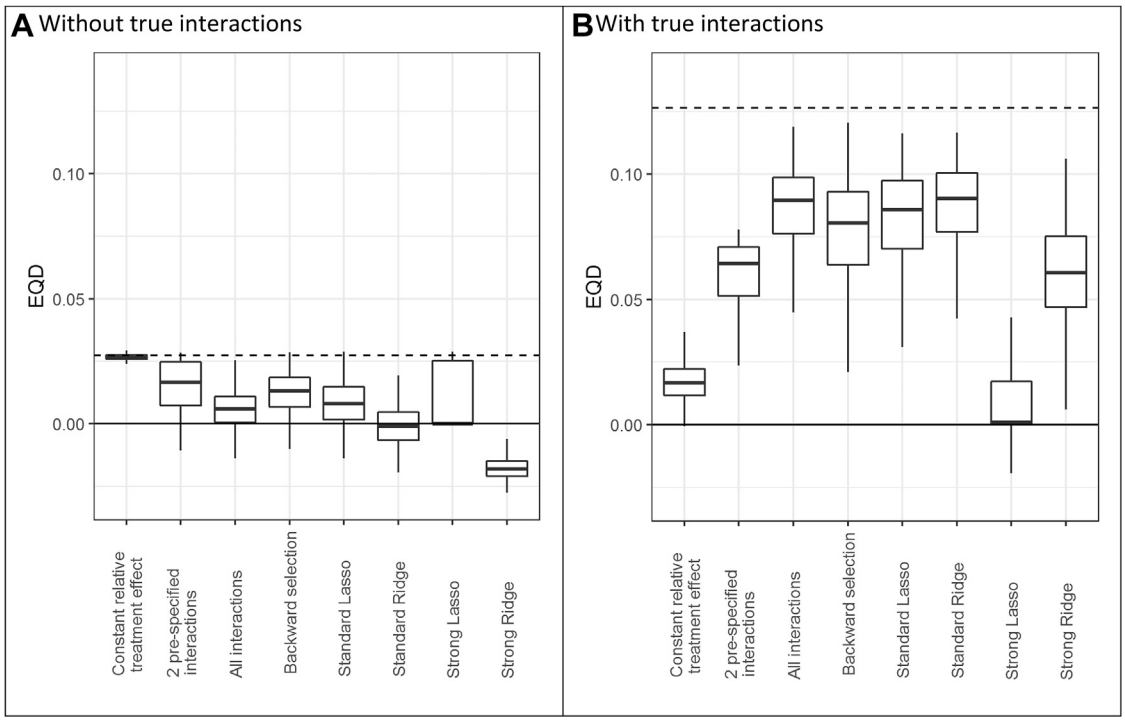

Fig. 4. The risk modeling approach (constant relative treatment effect) discriminated best in the absence of true interactions, but left considerable treatment benefit heterogeneity undetected in the presence of true interactions Extreme quartile difference (EQD) represents the difference between the observed benefit in the fourth quartile and the observed benefit in the first quartile of the population in base case simulation scenarios without true interactions (panel A) and with true interactions (panel B). The maximum achievable EQD of the true model is represented by the dashed horizontal lines. In the presence of true interactions (B), penalized regression approaches (Standard Lasso and Standard Ridge) discriminated similarly to unpenalized regression (All interactions).

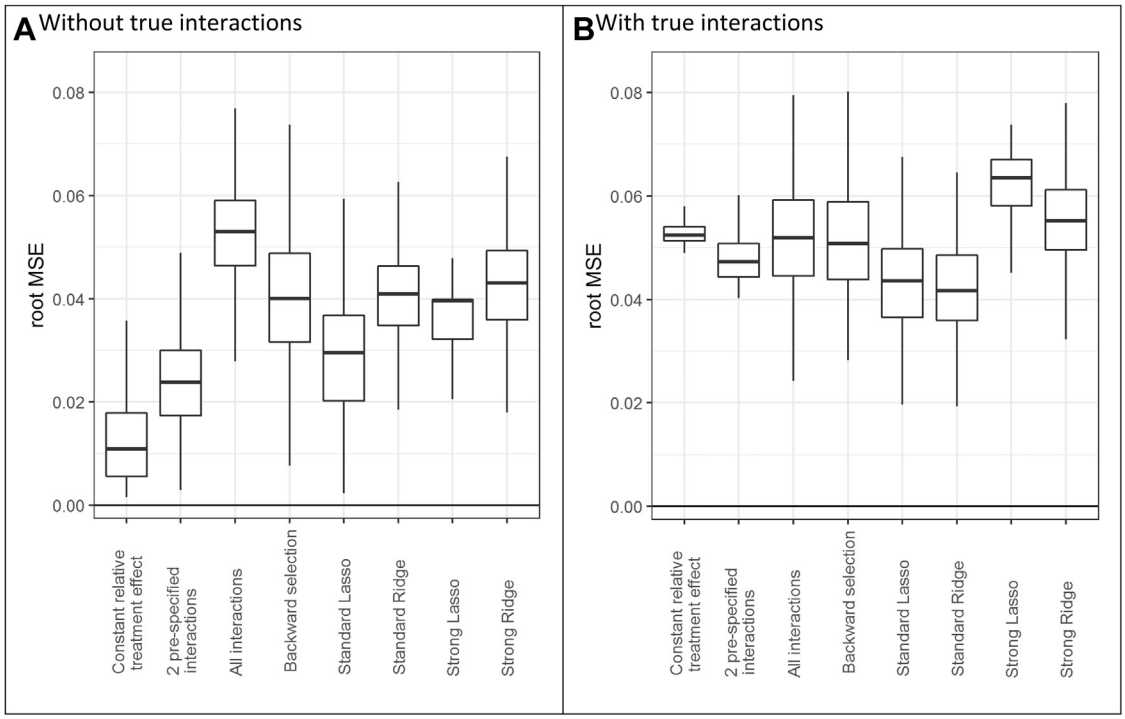

Fig. 5. The risk modeling approach (constant relative treatment effect) had minimal prediction error in the absence of true interactions but was outperformed by penalized regression approaches (Standard Lasso or Ridge) in the presence of true interactions The root mean squared error (rMSE) represents the root of the mean of the square differences between predicted benefit and true benefit in the population for base case simulation scenarios without true interaction (panel A), and with true interaction (panel B).

interactions. Lasso regression decreased the rMSE considerably to 0.029 but did not outperform the model with two prespecified treatment interactions (0.024), even in the largest sample size (Fig. A.13).

In the presence of true treatment interactions, (unpenalized) models with interactions only outperformed the model with a constant relative treatment effect with large sample size (n = 7,200; Fig. A.14). For example, in the base case scenario (treatment OR of 0.8; sample size 3,600), the rMSE was 0.053 for the model with a constant relative treatment effect and 0.052 for the model with all interactions (Fig. 5). Lasso regression decreased the rMSE considerably (0.043), but moderate sample size (n = 3,600) was needed for Lasso to outperform the model without treatment interactions or with two prespecified treatment interactions.

3.3.5. Sensitivity analyses

When continuous rather than binary baseline characteristics were simulated, the results were much the same (Figs. A.15−A.18). When the baseline characteristics were correlated rather than independent, the results were also very similar in the absence of true treatment interactions (Figs. A.19, A.21A, and A.22.A). In the presence of true treatment interactions however, when baseline characteristics were correlated, the advantage of Lasso regression over a model with a constant relative treatment effect attenuated (Figs. A.20−A.22.B). Less diversity in the possible combinations of strongly correlated baseline characteristics clearly reduced the additional discriminative ability that can be achieved by treatment interactions with individual baseline characteristics.

4. Discussion

This simulation exercise illustrates the serious overfitting that can arise when developing models to predict heterogeneous treatment effects in randomized samples. We have shown that overfitting of treatment effect heterogeneity can significantly overestimate (and underestimate) the treatment benefit for a substantial proportion of patients, potentially leading to suboptimal treatment decisions. The magnitude of the miscalibration of predicted benefit when treatment interaction terms are incorporated into a prediction model is relatively much larger than the miscalibration typically observed when stratifying a trial by predicted outcome risk and potentially much more clinically influential. Standard heuristics used to avoid overfitting models to predict outcome risk (e.g., 10 or 20 outcomes for each variable or interaction term considered [10,11]) are grossly inadequate to avoid this overfitting. Overfitting for predicted benefit persisted despite approaches that are usually effective for eliminating overfitting for outcome risk (such as penalized regression and a parsimonious non−data-driven approach including only two ‘‘established’’ effect modifiers), and even with large sample sizes.

In the absence of true treatment interactions, a prediction model without treatment interactions—that is, a risk modeling approach—has superior predictive performance, even for large sample size. Our simulations also confirm a prior study showing that, with a risk modeling approach, modeling on the control arm only induces serious overfitting that can be completely eliminated by modeling on the whole population, blinded to treatment [6]. Even in the absence of any treatment effect and with an EPV ratio over 80, modeling in the control arm induced a spurious risk by treatment effect interaction. To understand this phenomenon, one can imagine a model fitted only to the noise of the control arm. Using this model, the treatment arm would appear to have a uniform ''average'' risk, whereas the control arm would seem well stratified. In such a scenario, high-risk patients would erroneously be expected to benefit, and low-risk patients would erroneously be expected to be harmed by a totally ineffective treatment.

When true interactions are present, risk modeling still provides well-calibrated predictions, although it does not discriminate as well as models including interaction terms. Of the effect modeling approaches compared, Lasso regression proved to be the optimal approach to minimize overfitting of treatment benefit. Interestingly, some overfitting for the prediction of benefit persisted although Lasso regression leads to excellent calibration of predicted outcome risk in both trial arms, underscoring the importance of using metrics for benefit prediction rather than relying on conventional prediction metrics when validating models that predict benefit.

Intuitively, benefit prediction is more sensitive to miscalibration than risk prediction because relatively small discrepancies in observed vs. predicted outcome rates in the control and treatment arms can add up to large discrepancies in absolute treatment effects (Figs. A.10 and A.11). In addition, because the scale of absolute risk difference is generally much smaller than the scale of outcome risk, the error will be comparatively large (i.e., magnified). Furthermore, differences in treatment effects can be extremely consequential for clinical decision-making. This can be illustrated most dramatically by the effect models that were developed in samples from the population simulated with the ''hard null'' assumption (of no treatment effect). With 12 null treatment interactions, statistically significant HTE was found 87% (1−[1−0.15712]) of the time, an issue that cannot be corrected with even larger sample sizes. Inclusion of more candidate predictors with their potential treatment interactions—for example in a case study of high-dimensional data—would have increased the number of false positive treatment interactions and would presumably have increased the degree of overfitting when predicting benefit. For these reasons, effect modeling should be applied cautiously, particularly in the absence of a well-established overall treatment effect.

Previously published treatment effect models to support decision-making in reperfusion therapy for both coronary artery disease and acute stroke were developed with conventional maximum likelihood estimation [12−14]. This simulation shows that the treatment benefit predictions from these models are likely to be overfit, even when the treatment interactions are clinically and biologically plausible. Assessment of treatment benefit calibration in new settings is required to better understand the validity of these models. Recalibration using penalized regression techniques may improve the ability of these models to predict treatment benefit [15,16]; this was well-demonstrated in a re-analysis of the SPRINT trial comparing conventional regression to elastic net regularization (an amalgam of Ridge and Lasso) [17]. When tested on a separate but related clinical trial, predictions of benefit generated with conventional regression were virtually useless but those generated with elastic net regularization validated substantially better. Nevertheless, the best approach to penalization is a subject for future research.

In our simulations, models that were fitted with Lasso regression performed consistently better than models that were fitted with Ridge regression, although both were significantly overfit for benefit across most of the simulated scenarios. We hypothesize that Lasso more stringently penalizes null treatment interaction effects. More research is required to better understand the differences between using Lasso or Ridge regression for models that predict treatment benefit and also to compare these to other (regression- and nonregression-based) approaches used for data-driven benefit prediction [18,19]. Because penalization reduced the overfitting of benefit predictions, one might anticipate that stronger penalties would further improve benefit predictions. However, models that were fitted using penalized regression with maximum penalization, such that the error is within one standard error of the minimum, performed substantially worse than models that were fitted using penalized regression with minimum mean cross-validated error. The more stringent penalization of these methods seems to overly shrink the average treatment effect, resulting in poor calibration even in the middle quartiles of predicted benefit.

When true interactions are present in the data, discrimination improved with models incorporating interaction terms (i.e., an effect modeling approach always provided superior discrimination to a risk modeling approach). However, the effect modeling approaches were consistently overfit. The trade-off in terms of clinical decision-making between the improved discrimination and worse calibration resulting from the more aggressive effect modeling approaches will depend on the specific decisional context of each case. More research is needed to determine those circumstances where the trade-offs are likely to favor these more aggressive approaches. Because these effect modeling approaches can mislead investigators into falsely ''discovering'' subgroup treatment effects

even when no treatment effects exist, and even in large simulated databases, we recommend caution in applying these approaches, pending future research. These approaches might be applied in situations in which there is sufficient randomized data to permit fully external model validation and in situations with a priori evidence for important treatment interactions.

Benefit prediction approaches are useful in targeting treatment to those patients most likely to benefit, especially when treatment is associated with some treatment-related harm or burden. When we added a constant treatment-related harm—that is, we added outcomes with a 2.5% event rate in the treatment arm regardless of the risk factors—most of the findings were very similar (data not shown). The additive harm induced some interaction with risk on the relative scale, which the risk modeling approach unsurprisingly could not pick up. However, this led to a slight *under*fitting of the treatment effect heterogeneity by the risk modeling approach, in contrast with the general finding of *over*fitting the treatment effect heterogeneity by the effect modeling approaches.

As with all simulations, we were limited in covering the entire relevant space of risk and effect modeling in these simulated clinical trials. We attempted to emulate common scenarios. However, the benefits of the risk modeling approach (compared with a conventional one size fits all approach) may be limited because the risk heterogeneity within these simulated trials is relatively modest (c-index = 0.66). We have found extreme quartile risk ratios substantially larger in a recent empirical evaluation [20]. We also note that several aspects of the simulated trial population make it a relatively favorable setting for effect modeling, including the fact that 50% of tested interactions include ''true'' effects, some true effects were known a priori, and there were abundant outcomes. Nevertheless, our results consistently indicated problematic and resilient overfitting when an effect modeling approach was attempted. Risk modeling, on the other hand, when performed on the whole trial population, yielded well-calibrated treatment effects within risk strata, even in the presence of treatment interaction terms that were excluded from the model. We expressed calibration for benefit and discrimination for the benefit of each modeling approach in terms of quartiles of predicted benefit. A more detailed stratification—for example, into octiles or deciles of predicted benefit—would have emphasized the degree of overfitting through more extreme miscalibration in the lowest and highest strata of predicted benefit. Although calibration for benefit and discrimination for benefit can also be expressed on a continuous scale, for example, with a c-statistic and an E-statistic [21,22], the quartile-based presentation has the advantage of being illustrative for the actual use of these models in clinical practice. Future research could enable more individualized approaches to implementing predictive models in clinical practice.

In conclusion, predicting heterogeneous treatment effects has substantial challenges beyond the prediction of outcome risk. A risk modeling approach yields models consistently well-calibrated for benefit. Effect modeling may improve discrimination for benefit in the presence of true interactions but are prone to serious overfitting; they should only be considered when important treatment interactions are highly likely. Even under these circumstances, penalized approaches should be favored and external validation remains very important.

CRediT authorship contribution statement

David van Klaveren: Conceptualization, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Theodor A. Balan:** Conceptualization, Methodology, Visualization, Writing - review & editing. **Ewout W. Steyerberg:** Conceptualization, Methodology, Writing - review & editing. **David M. Kent:** Conceptualization, Methodology, Visualization, Writing - review & editing.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.05.029.

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. J Clin Epidemiol 2013;66:818−25.
[2] Kent DM, Steyerberg EW, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ 2018;363:k4245.
[3] Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007;298:1209−12.
[4] Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010;11:85.
[5] Abadie A, Chingos MM, West MR. Endogenous stratification in randomized experiments. National Bureau of Economic Research Working Paper Series; 2013: No. 19742.
[6] Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circ Cardiovasc Qual Outcomes 2014;7:163−9.
[7] Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. BMJ 2015;351:h5651.
[8] Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. JAMA Intern Med 2017;177:554−60.
[9] Wallach JD, Sullivan PG, Trepanowski JF, Steyerberg EW, Ioannidis JP. Sex based subgroup differences in randomized controlled trials: empirical evidence from cochrane meta-analyses. BMJ 2016;355:i5826.
[10] Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 2017;26:796−808.

[11] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373−9.

[12] Kent DM, Selker HP, Ruthazer R, Bluhmki E, Hacke W. The stroke-thrombolytic predictive instrument: a predictive instrument for intravenous thrombolysis in acute ischemic stroke. Stroke 2006;37:2957−62.

[13] 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. Lancet 2013;381:639−50.

[14] Kent DM, Ruthazer R, Decker C, Jones PG, Saver JL, Bluhmki E, et al. Development and validation of a simplified stroke-thrombolytic predictive instrument. Neurology 2015;85:942−9.

[15] van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. J Clin Epidemiol 2015;68:1366−74.

[16] Venema E, Mulder M, Roozenbeek B, Broderick JP, Yeatts SD, Khatri P, et al. Selection of patients for intra-arterial treatment for acute ischaemic stroke: development and validation of a clinical decision tool in two randomised trials. BMJ 2017;357:j1710.

[17] Basu S, Sussman JB, Rigdon J, Steimle L, Denton BT, Hayward RA. Benefit and harm of intensive blood pressure treatment: derivation and validation of risk models using data from the SPRINT and ACCORD trials. PLoS Med 2017;14:e1002410.

[18] Dmitrienko A, Millen B, Lipkovich I. Multiplicity considerations in subgroup analysis. Stat Med 2017;36:4446−54.

[19] Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 2011;30:2601−21.

[20] 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. Int J Epidemiol 2016;45:2075−88.

[21] Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.

[22] van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. J Clin Epidemiol 2018;94:59−68.