RESEARCH ARTICLE   **Open Access**

Predictive approaches to heterogeneous treatment effects: a scoping review

Alexandros Rekkas[1,2], Jessica K. Paulus[3], Gowri Raman[4], John B. Wong[5], Ewout W. Steyerberg[1,6], Peter R. Rijnbeek[2], David M. Kent[3*] and David van Klaveren[3,6]

Abstract

Background: Recent evidence suggests that there is often substantial variation in the benefits and harms across a trial population. We aimed to identify regression modeling approaches that assess heterogeneity of treatment effect within a randomized clinical trial.

Methods: We performed a literature review using a broad search strategy, complemented by suggestions of a technical expert panel.

Results: The approaches are classified into 3 categories: 1) Risk-based methods (11 papers) use only prognostic factors to define patient subgroups, relying on the mathematical dependency of the absolute risk difference on baseline risk; 2) Treatment effect modeling methods (9 papers) use both prognostic factors and treatment effect modifiers to explore characteristics that interact with the effects of therapy on a relative scale. These methods couple data-driven subgroup identification with approaches to prevent overfitting, such as penalization or use of separate data sets for subgroup identification and effect estimation. 3) Optimal treatment regime methods (12 papers) focus primarily on treatment effect modifiers to classify the trial population into those who benefit from treatment and those who do not. Finally, we also identified papers which describe model evaluation methods (4 papers).

Conclusions: Three classes of approaches were identified to assess heterogeneity of treatment effect. Methodological research, including both simulations and empirical evaluations, is required to compare the available methods in different settings and to derive well-informed guidance for their application in RCT analysis.

Introduction

Evidence based medicine (EBM) has heavily influenced the standards of current medical practice. Randomized clinical trials (RCTs) and meta-analyses of RCTs are regarded as the gold standards for determining the comparative efficacy or effectiveness of two (or more) treatments within the EBM framework. Within this framework, as described in Guyatt et al's classic User's Guide to the Medical Literature II [1], "if the patient meets all the [trial] inclusion criteria, and doesn't violate

any of the exclusion criteria—there is little question that the results [of the trial] are applicable". It has thus been argued that RCTs should attempt to include even broader populations to ensure generalizability of their results to more (and more diverse) individuals [2, 3].

However, generalizability of an RCT result and applicability to a specific patient move in opposite directions [4, 5]. When trial enrollees differ from one another in many observed determinants of the outcome of interest (both primary and safety), it can be unclear to whom the overall average benefit-harm trade-offs actually apply— even among those included in the trial [6, 7]. Precision medicine aims to target the appropriate treatment to the appropriate patients. As such, analysis of heterogeneity of treatment effect (HTE), i.e. non-random variation in

* Correspondence: dkent1@tuftsmedicalcenter.org
[3]Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies (ICRHPS), Tufts Medical Center, 800 Washington St, Box 63, Boston, MA 02111, USA
Full list of author information is available at the end of the article

the direction or magnitude of a treatment effect for subgroups within a population [8], is the cornerstone of precision medicine; its goal is to predict the optimal treatments at the individual level, accounting for an individual's risk for harm and benefit outcomes.

In this scoping review [9], we aim to identify and categorize the variety of regression-based approaches for predictive heterogeneity of treatment effects analysis. Predictive approaches to HTE analyses are those that provide individualized predictions of potential outcomes in a particular patient with one intervention versus an alternative or, alternatively, that can predict which of 2 or more treatments will be better for a particular patient, taking into account multiple relevant patient characteristics. We distinguish these analyses from the typical one-variable-at-a-time subgroup analyses that appear in forest plots of most major trial reports, and from other HTE analyses which explore or confirm hypotheses regarding whether a specific covariate or biomarker modifies the effects of therapy. To guide future work on individualizing treatment decisions, we aimed to summarize the methodological literature on regression modeling approaches to predictive HTE analysis.

Methods

The terminology in this scoping review hews closely to that in the PATH Statement and PATH Statement Explanation and Elaboration articles, and we refer readers to these papers for details. Generally, we use the term HTE to refer to a scale-dependent property. This is in distinction to other writers that have reserved the term HTE to refer specifically to heterogeneity on a relative scale [10]. Thus, when outcome risk varies across subgroups of patients, HTE must exist on some scale. If relative risk is constant, then there is HTE on the clinically important absolute scale. Nevertheless, since this review focuses on regression methods which are typically performed on the odds or hazard ratio scales, when we use the terms "effect modifier" and "effect modification" and "statistical interaction", we are generally referring to effect modification on a relative scale (e.g. hazard ratio or odds ratio), unless we otherwise specify—although we recognize that these too are scale dependent concepts [11–15]. Additionally, we note that we generally eschew the term "individual treatment effects", since person level effects cannot be observed or measured in parallel arm clinical trials (owing to the fundamental problem of causal inference, only one counterfactual outcome can be observed in a given patient). Nevertheless, the common goal of the different methods of predictive approaches to HTE we describe herein is to provide *"individualized"* treatment effect estimates from group-based data, since medical decisions are generally made at the individual person level [14]. These treatment effects are estimated conditional on many covariates, which are felt to be relevant for determining the benefits of therapy.

Due to the absence of medical subject headings (MeSH) for HTE, we used a relatively broad search strategy to maximize sensitivity. For the time period 1/1/2000 through 8/9/2018, we searched Medline and Cochrane Central using the text word search strategy from Table 1. We also retrieved seminal articles suggested by a technical expert panel (TEP). The TEP was comprised of 16 experts who represented various perspectives on predictive HTE analyses, including expertise in HTE, prediction modeling, clinical trials, and guideline development as well as a patient advocate. More details on the TEP are available in the PATH Statement [12, 13].

We sought papers that developed or evaluated methods for predictive HTE in the setting of parallel arm RCT designs or simulated RCT. Abstracts were screened to identify papers that developed or evaluated a regression-based method for predictive HTE on actual or simulated parallel arm RCT data. Papers describing a generic approach that could be applied using either regression or non-regression methods, or papers comparing regression to non-regression methods were also included. Similarly, papers comparing generic one-variable-at-a-time approaches to predictive HTE methods were also included. Finally, papers suggested by the TEP that fell outside the search window were considered for inclusion.

Table 1 Search strategy for the study

#	Results
1	((heterogen$ and effect$) or (effect and modif$)).tw.
2	regression.tw.
3	treatment$.tw.
4	(treatment adj1 effect$).tw.
5	(treatment adj1 difference$).tw.
6	exp risk/ or risk.tw.
7	3 or 4 or 5 or 6
8	*Models, Statistical/
9	*Randomized Controlled Trials as Topic/mt
10	Multicenter Studies as Topic/mt
11	*Randomized Controlled Trials as Topic/sn
12	Multicenter Studies as Topic/sn
13	*Clinical Trials as Topic/sn
14	*Precision Medicine/mt
15	or/8–14
16	1 and 2
17	2 and 7
18	15 and 17
19	15 and 16
20	18 or 19

We excluded papers solely related to cross-over, single-arm, and observational study designs. We also excluded papers that were primarily applications of existing methods, such as those that primarily aim to estimate a treatment effect of interest in a specific patient population, rather than papers with the primary aim of developing or evaluating methods of predictive HTE. We also excluded papers using only non-regression-based methods. Similarly, methods papers about ONLY non-predictive subgroup analysis, i.e. one-variable-at-a-time or conventional subgroups, were omitted. We excluded papers on trial enrichment or adaptive trial designs along with those that use predictive HTE approaches in the design. We also excluded papers primarily aiming at characterization or identification of heterogeneity in response rather than trying to predict responses for individual patients or subsets of patients; e.g. group based trajectory or growth mixture modeling. Papers on regression methods that make use of covariates post-baseline, or temporally downstream of the treatment decision were omitted. Review articles and primarily conceptual papers without accompanying methods development were also excluded.

Titles, abstracts and full texts were retrieved and double-screened by six independent reviewers against eligibility criteria. Disagreements were resolved by group consensus in consultation with a seventh senior expert reviewer (DMK) in meetings.

Results

We identified 2510 abstracts that were screened in duplicate. We retrieved 64 full-text articles and an additional 110 suggested by experts and identified from reference lists of eligible articles. These 174 full-text articles were again screened in duplicate with group consensus resolution of conflicts in meetings. A total of 36 articles met eligibility criteria (Fig. 1).

Categorization methods

We could classify all regression-based methods to predictive HTE into 3 broad categories based on whether and how they incorporated prognostic variables and relative treatment effect modifiers:

- **Risk-based methods** exploit the mathematical dependency of treatment benefit on a patient's baseline risk for the outcome under study [8, 9]. Even though relative treatment effect may vary across different levels of baseline risk, relative treatment effect modification by each covariate is not considered, i.e. no covariate by treatment interaction terms are considered (Table 2, eqs. 1–3).
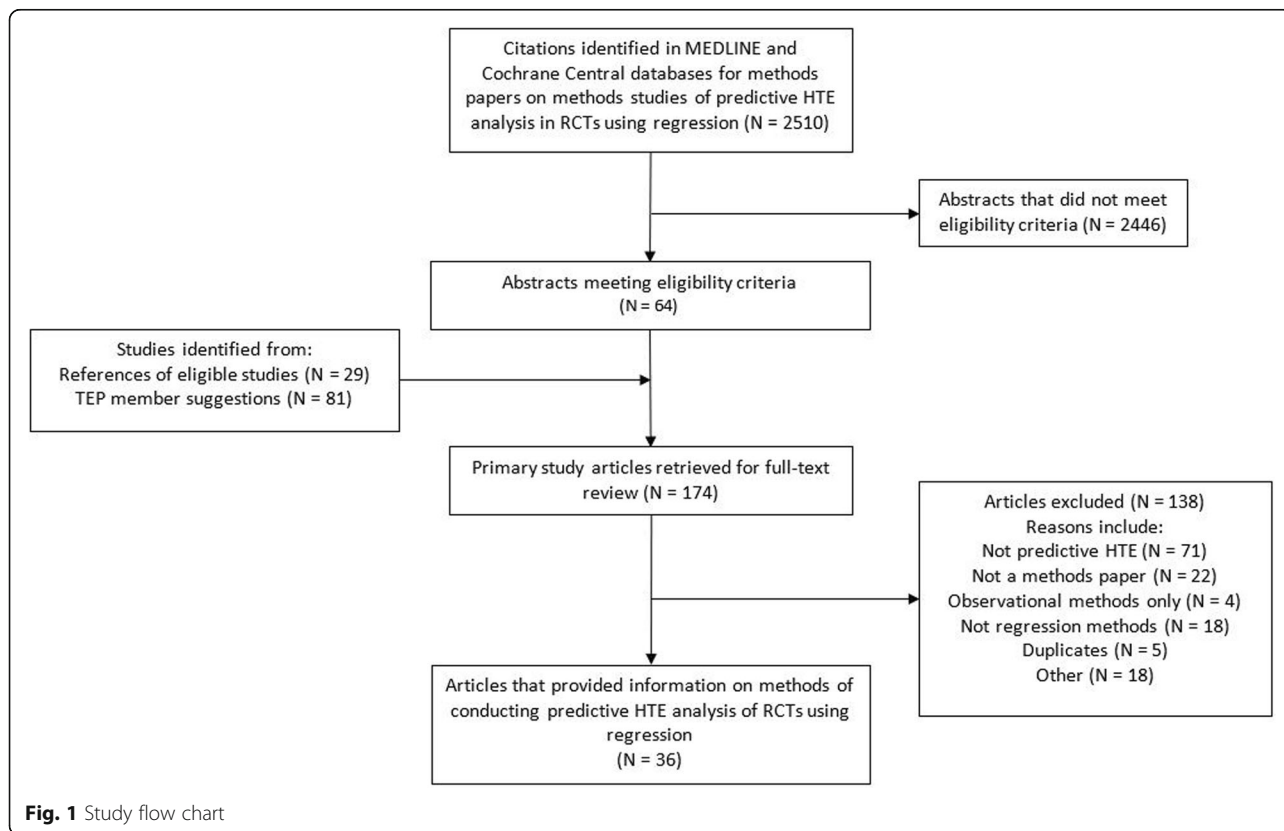

Fig. 1 Study flow chart

- **Treatment effect modeling methods** use both the main effects of risk factors and covariate-by-treatment interaction terms (on the relative scale) to estimate individualized benefits. They can be used either for making individualized absolute benefit predictions or for defining patient subgroups with similar expected treatment benefits (Table 2, eq. 4).
- **Optimal treatment regime methods** focus primarily on treatment effect modifiers (covariate by treatment interactions) for the definition of a treatment assignment rule dividing the trial population into those who benefit from treatment and those who do not (Table 2, eq. 5). Contrary to previous methods, baseline risk or the magnitude of absolute treatment benefit are not of primary concern.

Although risk-based methods emerged earlier (Fig. 2), methodology papers on treatment effect modeling (9 papers) and optimal treatment regimes (12 papers) are more frequently published since 2010 than papers on risk-based methods (8 papers). Even though extensive literature exists on model evaluation when it comes to prediction modeling, the same task can be quite challenging when modeling treatment effects [16]. That is due to the unavailability of counterfactual outcomes under the alternative treatment, providing a substantial challenge to the assessment of model fit. Methods included in the review concerning model evaluation in the setting of predictive HTE (4 papers) were assigned to a separate category as they are relevant to all identified approaches.

Risk-based methods

The most rigid and straightforward risk-based methods assume a constant relative treatment effect across different levels of baseline risk and ignore potential interactions with treatment. Dorresteijn et al. [17] studied individualized treatment with rosuvastatin for prevention of cardiovascular events. They combined existing prediction models (Framingham score, Reynolds risk score)

with the average rosuvastatin effect found in an RCT. To obtain individualized absolute treatment benefits, they multiplied baseline risk predictions with the average risk reduction found in trials. The value of the proposed approach is assessed in terms of improved decision making by comparing the net benefit with treat-none and treat-all strategies [18]. Julien and Hanley [19] estimated prognostic effects and treatment effect directly form trial data, by incorporating a constant relative treatment effect term in a Cox regression model. Patient-specific benefit predictions followed from the difference between event-free survival predictions for patients with and without treatment. A similar approach was used to obtain the predicted 30-day survival benefit of treatment with aggressive thrombolysis after acute myocardial infarction [20].

Risk stratification approaches analyze relative treatment effects and absolute treatment effects within strata of predicted risk, rather than assuming a constant relative effect. Both Hayward et al. [21] and Iwashyna et al. [22] demonstrated that these methods are useful in the presence of treatment-related harms to identify patients who do not benefit (or receive net harm) from a treatment that is beneficial on average. In a range of plausible scenarios evaluating HTE when considering binary endpoints, simulations showed that studies were generally underpowered to detect covariate-by-treatment interactions, but adequately powered to detect risk-by-treatment interactions, even when a moderately performing prediction model was used to stratify patients. Hence, risk stratification methods can detect patient subgroups that have net harm even when conventional methods conclude consistency of effects across all major subgroups.

Kent et al. [23] proposed a framework for HTE analysis in RCT data that recommended published trials routinely report the distribution of baseline risk in the overall study population and in the separate treatment arms using a risk prediction tool. Primarily binary or time-to-event outcomes were considered. Researchers should demonstrate how

Table 2 Equations corresponding to treatment effect heterogeneity assessment methods

Risk modeling

A multivariate regression model f that predicts the risk of an outcome y based on the predictors $x_1 \ldots, x_p$ is identified or developed:

$risk(x_1, \ldots, x_p) = E\{y|x_1, \ldots x_p\} = f(a + \beta_1 x_1 + \ldots \beta_p x_p)$ (1)

The expected outcome of a patient with measured predictors $x_1, \ldots, x_p$ receiving treatment T (where $T = 1$, when patient is treated and 0 otherwise) based on the linear predictor $lp(x_1, \ldots x_p) = a + \beta_1 x_1 + \ldots \beta_p x_p$ from a previously derived risk model can be described as:

$E\{y|x_1, \ldots, x_p, T\} = f(lp + \gamma_0 T + \gamma T|lp)$ (2)

When the assumption of constant relative treatment effect across the entire risk distribution is made (risk magnification), equation (2) takes the form:

$E\{y|x_1, \ldots, x_p, T\} = f(lp + \gamma_0 T)$ (3)
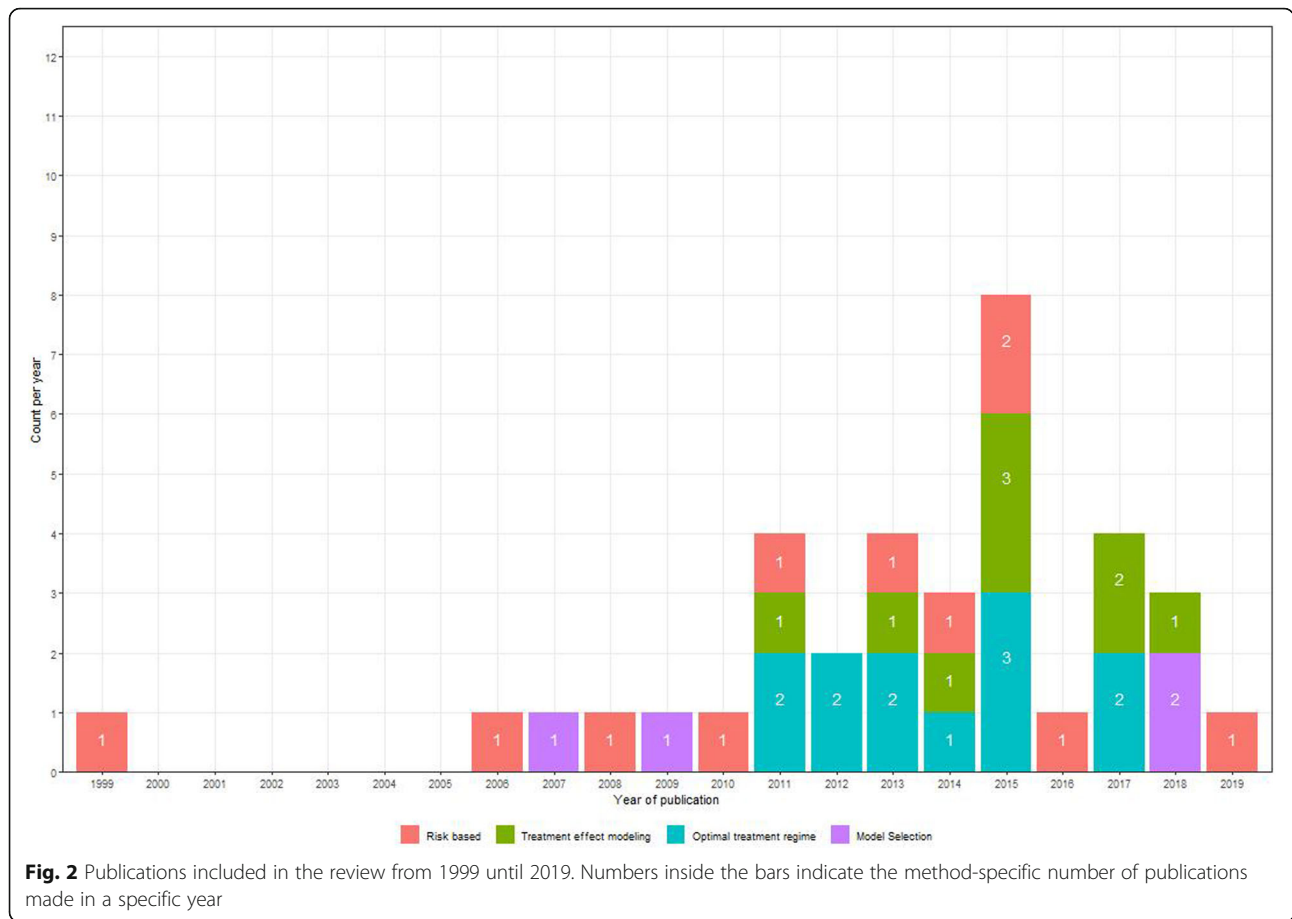
Treatment effect modeling

The expected outcome of a patient with measured predictors $x_1, \ldots, x_p$ receiving treatment T can be derived from a model containing predictor main effects and potential treatment interaction terms:

$E\{y|x_1, \ldots, x_p, T\} = f(a + \beta_1 x_1 + \cdots + \beta_p x_p + \gamma_0 T + \gamma_1 T x_1 + \cdots + \gamma_p T x_p)$ (4)

Optimal treatment regime

A treatment regime $T(x_1, \ldots, x_p)$ is a binary treatment assignment rule based on measured predictors. The optimal treatment regime maximizes the overall expected outcome across the entire target population:

$T_{optimal} = argmax_T E\{E\{y|x_1, \ldots x_p, T(x_1, \ldots, x_p)\}\}$ (5)

Fig. 2 Publications included in the review from 1999 until 2019. Numbers inside the bars indicate the method-specific number of publications made in a specific year

relative and absolute risk reduction vary by baseline risk and test for HTE with interaction tests. Externally validated prediction models should be used, when available.

In the absence of an adequate prediction model when performing a risk-based assessment of HTE, an internal risk model from the data at hand can be derived. Burke et al. [24] demonstrated that developing the risk model on the control arm of the trial may result in overfitting and, thus, exaggerate the presence of HTE. In extensive simulations, internally developed prediction models blinded to treatment assignment led to unbiased treatment effect estimates in strata of predicted risk. Using this approach to re-analyze 32 large RCT, Kent et al. [25] demonstrated that variation in the outcome risk within an RCT is very common, in the presence of adequately performing prediction models, which in turn leads to substantial HTE on the clinically important scale of absolute risk difference. Several trials from this analysis had clinically relevant results [26–28].

Similar to Burke et al. [24], Abadie et al. [29] presented evidence of large biases in risk stratified assessment of HTE in two randomized experiments rising

from the development of a prediction model solely from the control arm. They focused on financial outcomes that are primarily continuous. As a remedy, they considered both a leave-one-out approach, where individualized risk predictions are obtained from a model derived by excluding the particular individual, and a repeated split sample approach, where the original sample is repeatedly split into a sample for the development of the prediction model and a sample for treatment effect estimation within risk strata. These approaches were found to substantially reduce bias in a simulation study. Finally, Groenwold et al. [30] found in simulations that the inclusion of a constant relative treatment effect in the development of a prediction model better calibrates predictions to the untreated population. However, this approach may not be optimal for risk-based assessment of HTE, where accurate ranking of risk predictions is of primary importance for the calibration of treatment benefit predictions.

Follmann and Proschan [31] proposed a one-step likelihood ratio test procedure based on a proportional interactions model to decide whether treatment interacts with a linear combination of baseline covariates. Their

proportional interactions model assumes that the effects of prognostic factors in the treatment arm are equal to their effects in the control arm multiplied by a constant, the proportionality factor. Testing for an interaction along the linear predictor amounts to testing that the proportionality factor is equal to 1. If high risk patients benefit more from treatment (on the relative scale) and disease severity is determined by a variety of prognostic factors, the proposed test results in greater power to detect HTE on the relative scale compared to multiplicity-corrected subgroup analyses. Even though the proposed test requires a continuous response, it can be readily implemented in large clinical trials with binary or time-to-event endpoints.

Kovalchik et al. [32] expanded upon the previous approach by exploring misspecification of the proportional interactions model, when considering a fixed set of pre-specified candidate effect modifiers. A proportional interactions model is misspecified either when covariates with truly proportional effects are excluded or when covariates with non-proportional effects across treatment arms are included in the model. In this case the one-step likelihood ratio test of Follmann and Proschan [31] fails to achieve its statistical advantages. For model selection an all subsets approach combined with a modified Bonferroni correction method can be used. This approach accounts for correlation among nested subsets of considered proportional interactions models, thus allowing the assessment of all possible proportional interactions models while controlling for the familywise error rate.

Treatment effect modeling

Using data from the SYNTAX trial [33] Van Klaveren et al. [34] considered models of increasing complexity for the prediction of HTE using data from the SYNTAX trial. They compared different Cox regression models for the prediction of treatment benefit: 1) a model without any risk factors; 2) a model with risk factors and a constant relative treatment effect; 3) a model with treatment, a prognostic index and their interaction; and 4) a model including treatment interactions with all available prognostic factors, fitted both with conventional and with penalized ridge regression. Benefit predictions at the individual level were highly dependent on the modeling strategy, with treatment interactions improving treatment recommendations under certain circumstances.

Basu et al. [35] developed and validated risk models for predicting the absolute benefit (reduction of time to CVD events) and harm (serious adverse events) from intensive blood pressure therapy, using data from SPRINT. They compared traditional backward selection to an elastic net approach for selection and estimation of all treatment-covariate interactions. The two approaches selected different treatment-covariate interactions and—

while their performance in terms of CVD risk prediction was comparable when externally validated in the AC-CORD BP trial [36]—the traditional approach performed considerably worse than the penalized approach when predicting absolute treatment benefit. However, with regard to selection of treatment interactions, Ternes et al. [37] concluded from an extensive simulation study that no single methodology yielded uniformly superior performance. They compared 12 different approaches in a high-dimensional setting with survival outcomes. Their methods ranged from a straightforward univariate approach as a baseline, where Wald tests accounting for multiple testing were performed for each treatment-covariate interaction to different approaches for dealing with hierarchy of effects—whether they enforce the inclusion of the respective main effects if an interaction is selected—and also different magnitude of penalization of main and interaction effects.

Another approach to reducing overfitting of treatment effect models is separation of treatment effect estimation from subgroup identification. Cai et al. [38] fit "working" regression parametric models within treatment arms to derive absolute treatment benefit scores initially. In a second stage, the population is stratified into small groups with similar predicted benefits based on the first-stage scores. A non-parametric local likelihood approach is used to provide a smooth estimate of absolute treatment benefit across the range of the derived sores. The authors focused on continuous and binary endpoints, but their method can be extended to time-to-event outcomes. Claggett et al. [39] extended this two-stage methodology to RCTs with multiple outcomes, by assigning outcomes into meaningful ordinal categories. Overfitting can be avoided by randomly splitting the sample into two parts; the first part is used to select and fit ordinal regression models in both the treatment and the control arm. In the second part, the models that perform best in terms of a cross-validated estimate of concordance between predicted and unobservable true treatment difference— defined as the difference in probability of observing a worse outcome under control compared to treatment and the probability of observing a worse outcome under treatment compared to control—are used to define treatment benefit scores for patients. Treatment effects conditional on the treatment benefit score are then estimated through a non-parametric kernel estimation procedure.

Zhao et al. [40] proposed a two-stage methodology similar to the approach of Cai et al. [32], focusing on the identification of a subgroup that benefits from treatment. They repeatedly split the sample population based on the first-stage treatment benefit scores and estimate the treatment effect in subgroups above different thresholds. These estimates are plotted against the score thresholds to assess the adequacy of the selected scoring

rule. This method could also be used for the evaluation of different modeling strategies by selecting the one that identifies the largest subgroup with an effect estimate above a desired threshold.

Künzel et al. [41] proposed an "X-learner" for settings where one treatment arm is substantially larger than the alternative. They also start by fitting separate outcome models within treatment arms. However, rather than using these models to calculate treatment benefit scores, they imputed individualized absolute treatment effects, defined as the difference between the observed outcomes and the expected counterfactual (potential) outcomes based on model predictions. In a second stage, two separate regression models—one in each treatment arm—are fitted to the imputed treatment effects. Finally, they combined these two regression models for a particular covariate pattern by taking a weighted average of the expected treatment effects.

Most effect modeling methods start with outcome predictions conditional on treatment and then examine the difference in predictions with and without treatment. In contrast, Weisberg and Pontes [42] introduced a causal difference outcome variable ("cadit") which can be modeled directly. In case of a binary outcome, the binary cadit is 1 when a treated patient has a good outcome or when an untreated patient does not, and 0 otherwise. Thus, the dependent variable implicitly codes treatment assignment and outcome simultaneously. They first demonstrated that the absolute treatment benefit equals $2 \times P(\text{cadit} = 1) - 1$ and then they derived patient-specific treatment effect estimates by fitting a logistic regression model to the cadit. A similar approach was described for continuous outcomes with the continuous cadit defined as -2 and 2 times the centered outcome, i.e. the outcome minus the overall average outcome, for untreated and treated patients, respectively.

Finally, Berger et al. [43] proposed a Bayesian methodology for the detection of subgroup treatment effects in case of a continuous response and binary covariates. The approach identifies single covariates likely to modify treatment effect, along with the expected individualized treatment effect. The authors also extended their methodology to include two covariates simultaneously, allowing for the assessment of multivariate subgroups.

Optimal treatment regime methods

A treatment regime (TR) is a function mapping each patient's covariate pattern to a single treatment assignment. Any candidate TR can be evaluated based on its value, i.e. the expected outcome at the population level if the specific TR were to be followed. The TR achieving the highest value among all possible TRs is the optimal treatment regime (OTR). The majority of such methods follows a two-stage approach, where an outcome model—usually including treatment interactions—is used to derive expected treatment benefit in the first stage. In the second stage treatment assignment is optimized based on the expected outcome. Qian and Murphy [44] advocated a first-stage model including all covariate main effects and treatment interactions in combination with LASSO-penalization to reduce model complexity. Real-valued (continuous or binary) are considered without considering censoring.

When the outcome model is misspecified, however, the approach of Qian and Murphy may fail to identify the best possible treatment regime. As Zhang et al. [45] introduced an approach robust to such misspecifications that uses an augmented inverse probability weighted estimator of the value function. This is achieved by imposing a missing data framework, where the response under any candidate OTR is observed if the proposed treatment coincides with actual treatment and is considered missing otherwise. However, in commenting on this work, Taylor et al. [46] noted that the misspecification issues of the outcome models considered in the simulation study presented by Zhang et al. would have been easily spotted, if common approaches for the assessment of model fit had been examined. They argue that if adequately fitting outcome models had been thoroughly sought, the extra modeling required for the robust methods of Zhang et al. may not have been necessary.

Zhang et al. [47] proposed a novel framework for the derivation of OTRs for real-valued responses (continuous or binary), within which treatment assignment is viewed as a classification problem. The OTR is derived in two separate steps. In the first step, a contrast function is estimated, determining the difference between expected outcomes under different treatment assignments for each individual patient. The sign of the contrast function is then used to define class labels, i.e. -1 for negative contrast (harm) and + 1 for positive contrast (benefit). In the second step, any classification technique can be used to find the OTR by minimizing the expected misclassification error weighted by the absolute contrast. The authors demonstrated that many of the already existing OTR methods [44, 45] fit within their framework by defining a specific contrast function.

When the outcome of interest is continuous, the magnitude of absolute treatment benefit estimates derived from regression-based methods depends solely on treatment interactions. Therefore, Foster et al. [48] focus on non-parametric estimation of the function defining the structure of treatment-covariate interactions for a continuous outcome of interest. More specifically, they recursively update non-parametric estimates of the treatment-covariate interaction function from baseline risk estimates and vice-versa until convergence. The estimates of absolute treatment benefit are then used to

restrict treatment to a contiguous sub-region of the co-variate space.

Xu et al. [49] claimed that the identification of an OTR with high value depends on the adequate assessment of the sign of treatment-covariate interactions rather than on the estimation of the contrast function. They demonstrated that in many common cases (binary or time-to-event outcomes), even though the underlying structure of interactions can be quite complex, its sign can be approximated from a much simpler linear function of effect modifiers. Using the classification framework of Zhang et al. [47], they assign patients to class labels based on the resulting sign from these candidate linear combinations. The coefficients of that linear function are derived by minimizing the misclassification error weighted by the observed outcome—assuming higher values are preferable. In this way, the derived OTR is forced to contradict actual treatment assignment when the observed outcome is low. Tian et al. [50] proposed a different approach that solely focuses on treatment-covariate interactions by recoding the binary treatment indicator variable to $-1/2$ for control patients and $+1/2$ for treated patients and multiplying it with the covariates of a posited regression model to derive modified covariates so that the linear predictor of the model predicting the outcome from the modified covariates can be used as a score for stratifying patients with regard to treatment benefits. Starting from continuous responses they generalized their methodology to binary and time-to-event outcomes.

Kraemer [51] suggested a methodology that implicitly assesses treatment-covariate interactions using the correlation coefficient of the pairwise difference of the continuous outcome between treatment arms and their respective candidate predictive factor pairwise difference as a measure of effect modification. A stronger composite treatment effect modifier can then be constructed by fitting a regression model predicting pairwise outcome differences between treatments from the averages of the effect modifier values across treatment arms and then summing the individual effect modifiers weighted by the estimated regression coefficients. Treatment can then be assigned based on stratification on the composite treatment effect moderator. Two different approaches to model selection in Kraemer's effect modifier combination method were identified in clinical applications. Principal component analysis was used to select an uncorrelated subset from a large set of possibly correlated effect modifiers [52]. Alternatively, the cross-validated mean squared error of increasingly complex regression models was used to select the number of effect modifiers to construct the composite one [53].

Gunter et al. [54] proposed a method for the discovery of covariates that qualitatively interact with treatment.

Using LASSO regression to reduce the space of all possible combinations of covariates and their interaction with treatment to a limited number of covariate subsets, their approach selects the optimal subset of candidate covariates by assessing the increase in the expected response from assigning based on the considered treatment effect model, versus the expected response of treating everyone with the treatment found best from the overall RCT result. The considered criterion also penalizes models for their size, providing a tradeoff between model complexity and the increase in expected response. The method focuses solely on continuous outcomes, however, suggestions are made on its extension to binary type of outcomes.

Finally, Petkova et al. [55] proposed to combine baseline covariates into a single generated effect modifier (GEM) based on the linear model. The GEM is defined as the linear combination of candidate effect modifiers and the objective is to derive their individual weights. This is done by fitting linear regression models within treatment arms where the independent variable is a weighted sum of the baseline covariates, while keeping the weights constant across treatment arms. The intercepts and slopes of these models along with the individual covariate GEM contributions are derived by maximizing the interaction effect in the GEM model, or by providing the best fit to the data, or by maximizing the statistical significance of an F-test for the interaction effects—a combination of the previous two. The authors derived estimates that can be calculated analytically, which makes the method easy to implement.

A growing literature exists on estimating the effect of introducing the OTR to the entire population [56–59]. Luedtke and Van der Laan [56] provide an estimate of the optimal value—the value of the OTR—that is valid even when a subset of covariates exists for which treatment is neither beneficial nor harmful. It has been previously demonstrated that estimation of the optimal value is quite difficult in those situations [60]. Based on the proposed method, an upper bound of what can be hoped for when a treatment rule is introduced can be established. In addition, Luedtke and Van der Laan [59] provided an estimation method for the impact of treating the optimal subgroup, i.e. the subgroup that is assigned treatment based on the OTR. Their methodology returns an estimate of the population level effect of treating based on the OTR compared to treating no one.

Model evaluation

Schuler et al. [61] defined three broad classes of metrics relevant to model selection when it comes to treatment effect modeling. μ-risk metrics evaluate the ability of models to predict the outcome of interest conditional on treatment assignment. Treatment effect is either

explicitly modeled by treatment interactions or implicitly by developing separate models for each treatment arm. τ-risk metrics focus directly on absolute treatment benefit. However, since absolute treatment benefit is unobservable, it needs to be estimated first. Value-metrics originate from OTR methods and evaluate the outcome in patients that were assigned to treatment in concordance with model recommendations.

Vickers et al. [18] suggested a methodology for the evaluation of models predicting individualized treatment effects. The method relies on the expression of disease-related harms and treatment-related harms on the same scale. The minimum absolute benefit required for a patient to opt for treatment (treatment threshold) can be viewed as the ratio of treatment-related harms and harms from disease-related events, thus providing the required relationship. Net benefit is then calculated as the difference between the decrease in the proportion of disease-related events and the proportion of treated patients multiplied by the treatment threshold. The latter quantity can be viewed as harms from treatment translated to the scale of disease-related harms. Then, the net benefit of a considered prediction model at a specific treatment threshold can be derived from a patient-subset where treatment received is congruent with treatment assigned based on predicted absolute benefits and the treatment threshold. The model's clinical relevance is derived by comparing its net benefit to the one of a treat-all policy.

Van Klaveren et al. [62] defined a measure of discrimination for treatment effect modeling. A model's ability to discriminate between patients with higher or lower benefits is challenging, since treatment benefits are unobservable in the individual patient (since only one of two counterfactual potential outcomes can be observed). Under the assumption of uncorrelated counterfactual outcomes, conditional on model covariates, the authors matched patients from different treatment arms by their predicted treatment benefit. The difference of the observed outcomes between the matched patient pairs (0, 1: benefit; 0,0 or 1, 1: no effect; 1, 0: harm) acts as a proxy for the unobservable absolute treatment difference. The c-statistic for benefit can then be defined on the basis of this tertiary outcome as the proportion of all possible pairs of patient pairs in which the patient pair observed to have greater treatment benefit was predicted to do so.

Finally, Chen et al. [63] focused on the case when more than one outcomes—often non-continuous— are of interest and proposed a Bayesian model selection approach. Using a latent variable methodology, they link observed outcomes to unobservable quantities, allowing for their correlated nature. To perform model selection, they derive posterior probability estimates of false inclusion or false exclusion in the final model for the considered covariates. Following the definition of an outcome-space sub-region that is considered beneficial, individualized posterior probabilities of belonging to that beneficial sub-region can be derived as a by-product of the proposed methodology.

Discussion

We identified 36 methodological papers in recent literature that describe predictive regression approaches to HTE analysis in RCT data. These methodological papers aimed to develop models for predicting individual treatment benefit and could be categorized as follows: 1) risk modeling ($n = 11$), in which RCT patients were stratified or grouped solely on the basis of prognostic models; 2) effect modeling ($n = 9$), in which patients are grouped or stratified by models combining prognostic factors with factors that modify treatment effects on the relative scale (effect modifiers); 3) optimal treatment regimes ($n = 12$), which seek to classify patients into those who benefit and those who do not, primarily on the basis of effect modifiers. Papers on the evaluation of different predictive approaches to HTE ($n = 4$) were assigned to a separate category. Of note, we also found literature on the evaluation of biomarkers for treatment selection, which did not meet inclusion criteria [64–67].

Risk-based approaches use baseline risk determined by a multivariate equation to define the reference class of a patient as the basis for predicting HTE. Two distinct approaches were identified: 1) risk magnification [10, 68] assumes constant relative treatment effect across all patient subgroups, while 2) risk stratification analyzes treatment effects within strata of predicted risk. This approach is straightforward to implement, and may provide adequate assessment of HTE in the absence of strong prior evidence for potential effect modification. The approach might better be labeled 'benefit magnification', since benefit increases by higher baseline risk and a constant relative risk.

Treatment effect modeling methods focus on predicting the absolute benefit of treatment through the inclusion of treatment-covariate interactions alongside the main effects of risk factors. However, modeling such interactions can result in serious overfitting of treatment benefit, especially in the absence of well-established treatment effect modifiers. Penalization methods such as LASSO regression, ridge regression or a combination (elastic net penalization) can be used as a remedy when predicting treatment benefits in other populations. Staging approaches starting from—possibly overfitted— "working" models predicting absolute treatment benefits that can later be used to calibrate predictions in groups of similar treatment benefit provide another alternative. While these approaches should yield well calibrated personalized effect estimates when data are abundant, it is

yet unclear how broadly applicable these methods are in conventionally sized randomized RCTs. Similarly, the additional discrimination of benefit of these approaches compared to the less flexible risk modeling approaches remains uncertain. Simulations and empirical studies should be informative regarding these questions.

The similarity of OTRs to general classification problems—finding an optimal dichotomization of the covariates space—enables the implementation of several existing non-regression-based classification algorithms. For instance Zhao et al. [69] applied a support vector machine methodology for the derivation of an OTR for a binary outcome and was later extended to survival outcomes [70]. Because prognostic factors do not affect the sign of the treatment effect, several OTR methods rely primarily on treatment effect modifiers. However, when treatments are associated with adverse events or treatment burdens (such as costs) that are not captured in the primary outcome—as is often the case—estimates of the magnitude of treatment effect are required to ensure that only patients above a certain expected net benefit threshold (i.e. outweighing the harms and burdens of therapy) are treated. Similarly, these classification methods do not provide comparable opportunity for incorporation of patient values and preferences for shared decision making which prediction methods do.

While there is an abundance of proposed methodological approaches, examples of clinical application of HTE prediction models remain quite rare. This may reflect the fact that all these approaches confront the same fundamental challenges. These challenges include the unobservability of individual treatment response, the curse of dimensionality from the large number of covariates, the lack of prior knowledge about the causal molecular mechanisms underlying variation in treatment effects and the relationship of these mechanism to observable variables, and the very low power in which to explore interactions. Because of these challenges there might be very serious constraints on the usefulness of these methods as a class; while some methods may be shown to have theoretical advantages, the practical import of these theoretical advantages may not be ascertainable.

The methods we identified here generally approach the aforementioned challenges from opposite ends. Relatively rigid methods, such as risk magnification (in which relative effect homogeneity is assumed) and risk modeling (which examines changes in relative effect according to baselines risk only) deal with dimensionality, low power and low prior knowledge by restricting the flexibility of the models that can be built to emphasize the well understood influence of prognosis. Effect modeling approaches permit more flexible modeling and then subsequently try to correct for the overfitting that inevitably arises. Based on theoretical considerations and some simulations, it is

likely that the optimal approach depends on the underlying causal structure of the data, which is typically unknown. It is also likely that the method used to assess performance may affect which approach is considered optimal. For example, recent simulations have favored very simple approaches when calibration is prioritized, but more complex approaches when discrimination is prioritized—particularly in the presence of true effect modification [71]. Finally, it is uncertain whether any of these approaches will add value to the more conventional EBM approach of using an overall estimate of the main effect, or to the risk magnification approach of applying that relative estimate to a risk model.

We identify several limitations to our study. Because no MeSH identifying these methods exists, we anticipate that our search approach likely missed some studies. In addition, a recently growing literature of other non-regression based methods that assess predictive HTE in observational databases [72–74] would have been excluded. Finally, our review is descriptive and did not compare the approaches for their ability to predict individualized treatment effects or to identify patient subgroups with similar expected treatment benefits.

Based on the findings and the limitations of our review, several objectives for future research can be described. Optimal approaches to the reduction of overfitting through penalization need to be determined, along with optimal measures to evaluate models intended to predict treatment effect. General principles to judge the adequacy of sample sizes for predictive analytic approaches to HTE are required to complement the previous objectives. Also, methods that simultaneously predict multiple risk dimensions regarding both primary outcome risks and treatment-related harms need to be explored. The current regression-based collection of methods could be expanded by a review of non-regression approaches. Methods targeted at the observational setting need also to be considered. Additionally, a set of empirical and simulation studies should be performed to evaluate and compare the identified methods under settings representative of real world trials. The growing availability of publicly available randomized clinical trials should support this methodological research [75–77].

In conclusion, we identified a large number of methodological approaches for the assessment of heterogeneity of treatment effects in RCTs developed in the past 20 years which we managed to divide into 3 broad categories. Extensive simulations along with empirical evaluations are required to assess those methods' relative performance under different settings and to derive well-informed guidance for their implementation. This may allow these novel methods to inform clinical practice and provide decision makers with reliable individualized

information on the benefits and harms of treatments. While we documented an exuberance of new methods, we do note a marked dearth of comparative studies in the literature. Future research could shed light on advantages and drawbacks of methods in terms of predictive performance in different settings.

Availability of data and materials
Not applicable.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands. [2]Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands. [3]Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies (ICRHPS), Tufts Medical Center, 800 Washington St, Box 63, Boston, MA 02111, USA. [4]Center for Clinical Evidence Synthesis, ICRHPS, Tufts Medical Center, Boston, MA, USA. [5]Division of Clinical Decision Making, Tufts Medical Center, Boston, MA, USA. [6]Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands.

References
1. Guyatt GH, Sackett DL, Cook DJ, Guyatt G, Bass E, Brill-Edwards P, et al. Users' guides to the medical literature: II. How to use an article about therapy or prevention a. are the results of the study valid? JAMA. 1993; 270(21):2598–601.
2. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis. 1967;20(8):637–48.
3. Ford I, Norrie J. Pragmatic trials. N Engl J Med. 2016;375(5):454–63.
4. Caplan LR. Evidence based medicine: concerns of a clinical neurologist. J Neurol Neurosurg Psychiatry. 2001;71(5):569–74.
5. Kent DM, Kitsios G. Against pragmatism: on efficacy, effectiveness and the real world. Trials. 2009;10(1):48.
6. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet. 1995;345(8965):1616–9.
7. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA. 2007; 298(10):1209–12.
8. 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(8):818–25.
9. Daudt HM, van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. BMC Med Res Methodol. 2013;13(1):48.
10. Harrell F. Viewpoints on Heterogeneity of Treatment Effect and Precision Medicine [Internet]. Statistical Thinking. 2018 [cited 2020 Jun 14]. Available from: https://fharrell.com/post/hteview/.
11. Rothman K, Greenland S, Lash TL. Modern Epidemiology, 3rd Edition. 2007 31 [cited 2020 Jul 27]; Available from: https://www.rti.org/publication/modern-epidemiology-3rd-edition.
12. Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med. 2020;172(1):35–45.
13. Kent DM, van Klaveren D, Paulus JK, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement: explanation and elaboration. Ann Intern Med. 2020;172(1):W1–25.
14. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. Int J Epidemiol. 2016;45(6):2184–93.
15. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ. 2018;nil (nil):k4245.
16. Steyerberg E. Clinical prediction models: a practical approach to development, validation, and updating [internet]. New York: Springer-Verlag; 2009 [cited 2020 Jun 14]. (Statistics for Biology and Health). Available from: https://www.springer.com/gp/book/9780387772431.
17. Dorresteijn JAN, Visseren FLJ, Ridker PM, Wassink AMJ, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ. 2011;343:d5888.
18. Vickers AJ, Kattan MW, Daniel S. Method for evaluating prediction models that apply the results of randomized trials to individual patients. Trials. 2007;8:14.
19. Julien M, Hanley JA. Profile-specific survival estimates: making reports of clinical trials more patient-relevant. Clin Trials. 2008;5(2):107–15.
20. 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 GUSTO-I Investigator. Am Heart J. 1997;133(6):630–9.
21. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol. 2006;6:18.
22. Iwashyna TJ, Burke JF, Sussman JB, Prescott HC, Hayward RA, Angus DC. Implications of heterogeneity of treatment effect for reporting and analysis of randomized trials in critical care. Am J Respir Crit Care Med. 2015;192(9):1045–51.
23. 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.
24. 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(1):163–9.
25. 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;nil (nil):dyw118.
26. Kozminski MA, Wei JT, Nelson J, Kent DM. Baseline characteristics predict risk of progression and response to combined medical therapy for benign prostatic hyperplasia (BPH). BJU Int. 2015;115(2):308–16.
27. 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.
28. Upshaw JN, Konstam MA, van Klaveren D, Noubary F, Huggins GS, Kent DM. Multistate Model to Predict Heart Failure Hospitalizations and All-Cause Mortality in Outpatients With Heart Failure With Reduced Ejection Fraction: Model Derivation and External Validation. Circ Heart Fail. 2016;9(8).
29. Abadie A, Chingos MM, West MR. Endogenous stratification in randomized experiments. Rev Econ Stat. 2018;100(4):567–80.
30. Groenwold RHH, Moons KGM, Pajouheshnia R, Altman DG, Collins GS, Debray TPA, et al. Explicit inclusion of treatment in prognostic modeling

was recommended in observational and randomized settings. J Clin Epidemiol. 2016;78:90–100.

31. Follmann DA, Proschan MA. A multivariate test of interaction for use in clinical trials. Biometrics. 1999;55(4):1151–5.

32. Kovalchik SA, Varadhan R, Weiss CO. Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. Stat Med. 2013;32(28):4906–23.

33. Serruys PW, Morice M-C, Kappetein AP, Colombo A, Holmes DR, Mack MJ, et al. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med. 2009;360(10):961–72.

34. 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(11): 1366–74.

35. 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(10):e1002410.

36. Action to Control Cardiovascular Risk in Diabetes Study Group, Gerstein HC, Miller ME, Byington RP, Goff DC, Bigger JT, et al. Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med. 2008;358(24):2545–59.

37. Ternès N, Rotolo F, Heinze G, Michiels S. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. Biom J. 2017;59(4):685–701.

38. Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011;12(2):270–82.

39. Claggett B, Tian L, Castagno D, Wei L-J. Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. Biostatistics. 2015;16(1):60–72.

40. Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future comparative study. J Am Stat Assoc. 2013;108(502):527–39.

41. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci USA. 2019;116(10):4156–65.

42. Weisberg HI, Pontes VP. Post hoc subgroups in clinical trials: anathema or analytics? Clin Trials. 2015;12(4):357–64.

43. Berger JO, Wang X, Shen L. A Bayesian approach to subgroup identification. J Biopharm Stat. 2014;24(1):110–29.

44. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Stat. 2011;39(2):1180–210.

45. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012;68(4):1010–8.

46. Taylor JMG, Cheng W, Foster JC. Reader reaction to "a robust method for estimating optimal treatment regimes" by Zhang et al. (2012). Biometrics. 2015;71(1):267–73.

47. Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat. 2012;1(1):103–14.

48. Foster JC, Taylor JMG, Kaciroti N, Nan B. Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. Biostatistics. 2015;16(2):368–82.

49. Xu Y, Yu M, Zhao Y-Q, Li Q, Wang S, Shao J. Regularized outcome weighted subgroup identification for differential treatment effects. Biometrics. 2015; 71(3):645–53.

50. Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. J Am Stat Assoc. 2014;109(508):1517–32.

51. Kraemer HC. Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. Stat Med. 2013;32(11):1964–73.

52. Wallace ML, Frank E, Kraemer HC. A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. JAMA Psychiatry. 2013;70(11):1241–7.

53. Niles AN, Loerinc AG, Krull JL, Roy-Byrne P, Sullivan G, Sherbourne CD, et al. Advancing personalized medicine: application of a novel statistical method to identify treatment moderators in the coordinated anxiety learning and management study. Behav Ther. 2017;48(4):490–500.

54. Gunter L, Zhu J, Murphy S. Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. J Biopharm Stat. 2011;21(6):1063–78.

55. Petkova E, Tarpey T, Su Z, Ogden RT. Generated effect modifiers (GEM's) in randomized clinical trials. Biostatistics. 2017;18(1):105–18.

56. Luedtke AR, van der Laan MJ. Statistical Inference For The Mean Outcome Under A Possibly Non-Unique Optimal Treatment Strategy. Ann Stat. 2016; 44(2):713–42.

57. van der Laan MJ, Luedtke AR. Targeted learning of the mean outcome under an optimal dynamic treatment rule. J Causal Inference. 2015;3(1):61–95.

58. Chakraborty B, Laber EB, Zhao Y-Q. Inference about the expected performance of a data-driven dynamic treatment regime. Clin Trials. 2014; 11(4):408–17.

59. Luedtke AR, van der Laan MJ. Evaluating the impact of treating the optimal subgroup. Stat Methods Med Res. 2017;26(4):1630–40.

60. Robins J, Rotnitzky A. Discussion of "Dynamic treatment regimes: Technical challenges and applications". Electron J Statist. 2014;8(1):1273–89.

61. Schuler A, Baiocchi M, Tibshirani R, Shah N. A comparison of methods of model selection when estimating individual treatment effects. arXiv: 180405146 [cs, stat] [Internet]. 2018 13 [cited 2020 Jun 14]; Available from: http://arxiv.org/abs/1804.05146.

62. 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.

63. Chen W, Ghosh D, Raghunathan TE, Sargent DJ. Bayesian variable selection with joint modeling of categorical and survival outcomes: an application to individualizing chemotherapy treatment in advanced colorectal cancer. Biometrics. 2009;65(4):1030–40.

64. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. Ann Intern Med. 2011;154(4):253–9.

65. Janes H, Pepe MS, McShane LM, Sargent DJ, Heagerty PJ. The Fundamental Difficulty With Evaluating the Accuracy of Biomarkers for Guiding Treatment. J Natl Cancer Inst. 2015 Aug;107(8).

66. Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. Biometrics. 2012;68(3):687–96.

67. Polley M-YC, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. J Natl Cancer Inst. 2013;105(22):1677–83.

68. Harrell F. EHRs and RCTs: Outcome Prediction vs. Optimal Treatment Selection [Internet]. Statistical Thinking. 2017 [cited 2020 Jun 14]. Available from: https://fharrell.com/post/ehrs-rcts/.

69. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc. 2012;107(449): 1106–18.

70. Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR. Doubly robust learning for estimating individualized treatment with censored data. Biometrika. 2015;102(1):151–68.

71. van Klaveren D, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. J Clin Epidemiol. 2019;114(nil):72–83.

72. Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019; 47(2):1148–78.

73. Powers S, Qian J, Jung K, Schuler A, Shah NH, Hastie T, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. Stat Med. 2018;37(11):1767–87.

74. Louizos C, Shalit U, Mooij J, Sontag D, Zemel R, Welling M. Causal Effect Inference with Deep Latent-Variable Models. arXiv:170508821 [cs, stat] [Internet]. 2017 [cited 2020 Jun 14]; Available from: http://arxiv.org/abs/1705.08821.

75. Navar AM, Pencina MJ, Rymer JA, Louzao DM, Peterson ED. Use of open access platforms for clinical trial data. JAMA. 2016;315(12):1283–4.

76. Ross JS. Clinical research data sharing: what an open science world means for researchers involved in evidence synthesis. Syst Rev. 2016;5(1):159.

77. Ross JS, Waldstreicher J, Bamford S, Berlin JA, Childers K, Desai NR, et al. Overview and experience of the YODA Project with clinical trial data sharing after 5 years. Sci Data. 2018;5:180268.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.