

Evaluating individualized treatment effect predictions: A model-based perspective on discrimination and calibration assessment

J. Hoogland^{1,2} | O. Efthimiou^{3,4} | T. L. Nguyen⁵ | T. P. A. Debray^{1,6}

¹Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

²Epidemiology and Data Science, Amsterdam University Medical Center, Amsterdam, The Netherlands

³Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland

⁴Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland

⁵Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

⁶Smart Data Analysis and Statistics B.V., Utrecht, The Netherlands

Correspondence

J. Hoogland, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.
Email: j.hoogland@amsterdamumc.nl

Funding information

Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 180083; ZonMw, Grant/Award Numbers: 91215058, 91617050; European Union's Horizon 2020 research and innovation program, Grant/Award Number: 825746

In recent years, there has been a growing interest in the prediction of individualized treatment effects. While there is a rapidly growing literature on the development of such models, there is little literature on the evaluation of their performance. In this paper, we aim to facilitate the validation of prediction models for individualized treatment effects. The estimands of interest are defined based on the potential outcomes framework, which facilitates a comparison of existing and novel measures. In particular, we examine existing measures of discrimination for benefit (variations of the c-for-benefit), and propose model-based extensions to the treatment effect setting for discrimination and calibration metrics that have a strong basis in outcome risk prediction. The main focus is on randomized trial data with binary endpoints and on models that provide individualized treatment effect predictions and potential outcome predictions. We use simulated data to provide insight into the characteristics of the examined discrimination and calibration statistics under consideration, and further illustrate all methods in a trial of acute ischemic stroke treatment. The results show that the proposed model-based statistics had the best characteristics in terms of bias and accuracy. While resampling methods adjusted for the optimism of performance estimates in the development data, they had a high variance across replications that limited their accuracy. Therefore, individualized treatment effect models are best validated in independent data. To aid implementation, a software implementation of the proposed methods was made available in R.

KEYWORDS

treatment effect, individualized, prediction, discrimination, calibration

1 | INTRODUCTION

Prediction models for important health outcomes have long been a crucial aspect of personalized healthcare.¹⁻³ In line, methods for assessing their performance have been well established, and include overall accuracy, discrimination, and calibration assessment.¹⁻⁵

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd.

In recent years, there has been a growing interest in the prediction of key health outcomes under different treatment options.⁶⁻¹¹ Such individualized treatment effect (ITE) predictions are clearly of interest in many applied settings, including medical decision making. Due to the causal nature of ITE predictions, such models are typically developed in randomized data or explicitly account for confounding by other means. While there is a large and rapidly growing literature on the *development* of models for individualized treatment effect prediction, for example,¹⁰⁻¹⁷ literature on the corresponding *performance assessment* is scarce.^{11,18-20}

In this paper, we build on existing proposals for the assessment of clinical prediction models for ITE prediction,^{18,19,21} with a focus on measures of discrimination for benefit and calibration for benefit. Discrimination and calibration have been key quantities of interest for clinical prediction model evaluation in the case of binary outcomes,^{1,2,4,22} and previous proposals for discrimination for benefit and calibration for benefit have mainly focused on binary outcomes.^{18,21} Consequently, we focus on binary outcomes, which are also very common in randomized clinical trials.²³ Nevertheless, we will briefly digress into analogous procedures for continuous outcomes, since both quantities are also of interest there as well.

With respect to the types of ITE prediction models of interest, we focus on causal prediction models that contrast outcome predictions under different treatment options. This in contrast to (i) models that directly estimate ITEs without considering the outcomes per treatment condition as key estimands, and (ii) models that only predict the sign of treatment effect (ie, benefit or harm). The main reason is that medical decision making is best informed by *both* prognostic information *and* treatment effect information, and should balance the benefits and harms of initiating particular treatments, the underlying risk of disease without treatment, and patient preferences.²⁴ For example, an effective treatment with significant potential harm and cost may be of interest only to those at high risk without the treatment.²⁵

Previous work by our group has addressed a number of evaluation metrics for ITE models for both binary and continuous outcomes, including decision accuracy, discrimination for benefit and calibration for benefit.¹⁹ The main contribution of this paper is a more in depth exploration of discrimination and calibration for benefit that (i) utilizes the potential outcomes framework to increase clarity of exposition and to clearly define the estimands of interest,^{26,27} (ii) incorporates recently proposed modifications to the c-for-benefit,^{20,21} and (iii) proposes model-based estimators of the defined discrimination and calibration estimands that avoid the need for matching. Simulation results are provided for illustrative purposes. An applied example using data from the third International Stroke Trial (IST-3)²⁸ serves to further illustrate implementation in practice. To aid implementation, a software implementation of the proposed methods is made available in R.

2 | INDIVIDUALIZED TREATMENT EFFECT PREDICTION

Most outcome prediction research focuses on capturing statistical association in absence of interventions. Individualized treatment effect (ITE) prediction is a different type of prediction since it has a causal interpretation: the quantity to be predicted is the effect caused by the treatment (or intervention, in a larger sense) on the outcome. Therefore, before moving to the performance measures of interest, this section shortly outlines causal prediction. Subsequently, issues surrounding the use of binomial outcome data for ITE modeling are shortly discussed (further details are available as Online Supplementary Material A).

2.1 | Causal prediction

To emphasize the causal nature of the predictions, it is helpful to write the individualized treatment effect of interest in terms of the potential outcomes framework.^{26,27} For treatment taking values $a \in \mathcal{A}$, $Y^{A=a}$ denotes the potential outcome under treatment a . When comparing two treatments, the ITE for individual i, \dots, n can be defined as

$$\delta(\mathbf{x}_i) = \mathbb{E}(Y^{a=1} | \mathbf{X} = \mathbf{x}_i) - \mathbb{E}(Y^{a=0} | \mathbf{X}_i = \mathbf{x}_i) \quad (1)$$

where \mathbf{x}_i is a row vector of individual-level characteristics in matrix \mathbf{X} . The degree of granularity or individualization reflected by $\delta(\mathbf{x}_i)$ relates to the number of predictors included in \mathbf{X} , to the strength and shape of their association with the potential outcomes, and especially to the degree to which they have a differential effect across potential outcomes (ie, modify the effect of treatment). Ideally, the set of measured individual-level characteristics includes all relevant characteristics with respect to individualized treatment effect. In practice however, this set of all relevant characteristics is often

unknown and the best way forward is to aim for conditioning on the most important characteristics. Correspondingly, equation (1) reflects ITE as a conditional treatment effect for some set of characteristics.

Since in practice only one potential outcome is observed per individual,²⁹ assumptions are required to estimate $\delta(\mathbf{x}_i)$ based on the observed data. These assumptions are discussed in detail elsewhere.^{11,30} In short, the key assumptions are *exchangeability* (the potential outcomes do not depend on the assigned treatment), *consistency* (the observed outcome under treatment $a \in \mathcal{A}$ corresponds to the potential outcomes $Y^{A=a}$), and *positivity* (each individual has a nonzero probability of each treatment assignment). An additional assumption that eases inference is *no interference* (the potential outcomes for individual i do not depend on treatment assignment to other individuals). Based on these assumptions, the individualized treatment effect can be identified given the observed data:

$$\begin{aligned}\delta(\mathbf{x}_i) &= \mathbb{E}(Y^{a=1} | \mathbf{X} = \mathbf{x}_i) - \mathbb{E}(Y^{a=0} | \mathbf{X} = \mathbf{x}_i) \\ &= \mathbb{E}(Y^{a=1} | A = 1, \mathbf{X} = \mathbf{x}_i) - \mathbb{E}(Y^{a=0} | A = 0, \mathbf{X} = \mathbf{x}_i) \quad (\text{by exchangeability}) \\ &= \mathbb{E}(Y_i | A = 1, \mathbf{X} = \mathbf{x}_i) - \mathbb{E}(Y_i | A = 0, \mathbf{X} = \mathbf{x}_i) \quad (\text{by consistency}).\end{aligned}\tag{2}$$

Equation (2) shows that ITE predictions $\hat{\delta}(\mathbf{x}_i)$ can be estimated using a prediction model for outcome risk $\mathbb{E}(Y_i | A = a_i, \mathbf{X} = \mathbf{x}_i)$. Many modeling tools can be used for this endeavor and the details are beyond the scope of this paper and are given elsewhere (eg, ^{11,31}). When conditioning on \mathbf{x} is clear from the context, we at times abbreviate $\hat{\delta}(\mathbf{x}_i)$ as $\hat{\delta}_i$, or write the vector of predictions for all individuals $1, \dots, n$ as $\hat{\delta}$.

2.2 | Observed outcome data

For binary outcomes, we observe outcome $Y_i \in \{0, 1\}$ and covariate status \mathbf{x}_i for each individual i . In this context, the ITE estimate $\delta(\mathbf{x}_i)$ is a difference between two risk predictions ($P(Y_i = 1 | A = 1, \mathbf{X} = \mathbf{x}_i) - P(Y_i = 1 | A = 0, \mathbf{X} = \mathbf{x}_i)$). While other quantities are available to express treatment effect (eg, relative risk, odds ratio), risk differences are generally preferred and better understood by clinicians.³²⁻³⁴ The range of $\hat{\delta}(\mathbf{x}_i)$ includes all values in the $[-1, 1]$ interval, while the observed difference between any two outcomes can only be one of $\{-1, 0, 1\}$. Therefore, the observations come with large and irreducible binomial error and provide only limited information. Also, predictions for binary outcome data commonly involve nonlinear functions of the covariates, and hence the effects of treatment and the covariates are typically not additive on the risk difference scale of interest here. Consequently, the resulting ITE predictions conflate variability from different sources: between-subject variability in $P(Y^{a=0} = 1 | X = x)$ and genuine treatment effect heterogeneity on the scale used for modeling. This is the price to pay for the benefit in terms of interpretation of measures on the scale of $\delta(x)$.³⁵

In case of a continuous outcome, the problems at hand simplify considerably. First, continuous outcomes are far less noisy than binary outcomes. Second, many continuous outcome models have an identity link function, which puts the parameter effects directly on the outcome scale, and, most importantly, avoids the conflation of the effects of treatment effect related parameter and other model parameters.

Regardless of the type of outcome measure, the fact that only one potential outcome can be observed is a key challenge at the time of model development *and* evaluation. As opposed to the evaluation of regular prediction models of directly observable outcomes, a direct comparison between predictions $\hat{\delta}(\mathbf{x}_i)$ and observed outcomes is not feasible.

3 | DISCRIMINATION FOR INDIVIDUALIZED TREATMENT EFFECTS

Discriminative model performance reflects the degree to which model predictions are correctly rank-ordered and is a common performance measure in prediction models for binary and survival endpoints.^{2,3} In the binary endpoint setting, individualized treatment effects estimates $\hat{\delta}(\mathbf{x}_i)$ for i, \dots, n range from $[-1, 1]$, and differences between potential outcomes $\Delta_i = Y_i^{a=1} - Y_i^{a=0}$ take values in $\{-1, 0, 1\}$. The aim is to quantify the degree to which $\hat{\delta}(\mathbf{x})$ correctly rank-orders the probability to observe benefit in terms of Δ . The continuous case is discussed in Supplementary Material C.

3.1 | Discrimination estimand

We base our estimand on the well-known c-statistic, which describes the proportion of concordant pairs amongst pairs not tied on the outcome. Analogously, for a randomly sampled pair of cases with discordant outcomes, it describes the

probability of concordant predictions (ie, for current purposes: where the case with the lower predicted treatment effect in the pair also has the lower probability of benefit). In particular, we start from the c-statistic as formulated by Harrell,²² since it allows for ordinal outcomes. Note that such measures of concordance with respect to rank order have a long history, dating back to Kendall's proposal of τ as a measure of rank correlation. Supplementary Material B provides a short overview of measures of association leading up to the c-statistic by Harrell to build some more intuition. Adapting to the current setting, for a randomly selected pair of cases $k, l \in 1, \dots, n$ ($k \neq l$), the c-statistic in absence of tied predictions can be defined as

$$\begin{aligned} P(\text{concordance} \mid \text{differential benefit}) &= \frac{P(\text{concordance} \cap \text{differential benefit})}{P(\text{differential benefit})} \\ &= \frac{P(\Delta_k < \Delta_l \cap \hat{\delta}_k < \hat{\delta}_l) + P(\Delta_k > \Delta_l \cap \hat{\delta}_k > \hat{\delta}_l)}{P(\Delta_k < \Delta_l \cup \Delta_k > \Delta_l)}, \end{aligned} \quad (3)$$

where $P(\text{concordance} \mid \text{differential benefit})$ is conditional on $\hat{\delta}_k$ and $\hat{\delta}_l$ (treated as fixed), and the selection of the pairs (k, l) as well as the outcomes Δ are taken to be random. Thus, $P(\Delta_k < \Delta_l)$ is the probability to observe a more beneficial treatment effect for individual k as compared to individual l . When the proportion of concordant pairs increases, equation (3) goes to 1; conversely, it goes to 0 when the proportion of discordant pairs increases. Likewise, Equation (3) moves to 0 when the proportion of ties in $\hat{\delta}$ for pairs with a nonzero $P(\Delta_k \neq \Delta_l)$ increases. In line with the c-statistic, we opt for an estimand that moves toward 0.5 for pairs tied on $\hat{\delta}$ and with nonzero $P(\Delta_k \neq \Delta_l)$. Our main discrimination estimand of interest can be formulated as

$$\theta_d = \frac{P(\Delta_k < \Delta_l \cap \hat{\delta}_k < \hat{\delta}_l) + P(\Delta_k > \Delta_l \cap \hat{\delta}_k > \hat{\delta}_l) + \zeta}{P(\Delta_k < \Delta_l \cup \Delta_k > \Delta_l)}, \quad (4)$$

where $\zeta = \frac{1}{2}P(\Delta_k \neq \Delta_l \cap \hat{\delta}_k = \hat{\delta}_l)$. Consequently, 0.5 represents a neutral value for the case where the order of $\hat{\delta}_k, \hat{\delta}_l$ does not provide information on the probability of benefit (or harm). Since the comparisons are made between all k and l where $k \neq l$, each pair is evaluated twice and benefit in the comparison k to l means harm in the comparison l to k . This allows for an equivalent formulation of θ_d in terms of just benefit as

$$\theta_d = \frac{P(\Delta_k < \Delta_l \cap \hat{\delta}_k < \hat{\delta}_l) + \zeta^*}{P(\Delta_k < \Delta_l)}, \quad (5)$$

where $\zeta^* = \frac{1}{2}P(\Delta_k < \Delta_l \cap \hat{\delta}_k = \hat{\delta}_l)$.

3.2 | Discrimination estimators based on matching

In practice, differences between the potential outcomes of interest are not observable at the individual level. Consequently, the required observations of the Δ 's in θ_d are unavailable and have to be approximated. Thus, the truly individual Δ 's are out of reach, and in the following we will use approximations conditional on covariates \mathbf{x} . One of the possibilities is to use matching as in the proposed c-for-benefit that aims to quantify discriminative performance on the ITE level.¹⁸ Below we outline the c-for-benefit, its properties, and a modification. Thereafter, we introduce an alternative model-based approach to estimate θ_d .

3.2.1 | C-for-benefit definition

In the setting of a randomized two-arm study measuring a binary outcome of interest, the c-for-benefit aims to assess discrimination at the level of ITE predictions (referred to as 'predicted [treatment] benefit' in the original paper).^{*} The problem of unobserved individual treatment effects is approached from a matching perspective. One-to-one matching is used to match treated individuals to control individuals based on their predicted treatment effects. The subsequent data pairs hence consist of a treated individual and a control individual with similar predicted treatment effect. Observed treatment effect within the pair is defined as the difference in outcomes between these two individuals. Of note, observed

(within-pair) treatment effect can only be in $\{-1,0,1\}$. Subsequently, the c-for-benefit has been defined as “the proportion of all possible pairs of matched individual pairs with unequal observed benefit in which the individual pair receiving greater treatment benefit was predicted to do so.”¹⁸ The predicted treatment effect within each pair used in this definition is taken to be the (within-pair) average of predicted treatment effects. That is, for a control individual i out of $1, \dots, n_C$ (with n_C being the number of controls) and a treated individual j out of $1, \dots, n_T$ (with n_T being the number treated), predicted treatment effects are taken to be

$$\hat{\delta}_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \{(\hat{P}(Y_i = 1|A_i = 1, \mathbf{X} = \mathbf{x}_i) - \hat{P}(Y_i = 1|A_i = 0, \mathbf{X} = \mathbf{x}_i)) + (\hat{P}(Y_j = 1|A_j = 1, \mathbf{X} = \mathbf{x}_j) - \hat{P}(Y_j = 1|A_j = 0, \mathbf{X} = \mathbf{x}_j))\}/2. \quad (6)$$

The ‘observed’ treatment effect is subsequently taken to be $O_{ij} = Y_i - Y_j$. If the two (binary) outcomes in such a pair are discordant, then there supposedly is some evidence of a treatment effect (ie, benefit or harm); conversely, there is no such evidence when the outcomes are concordant (ie, the predicted treatment effect did not manifest as a difference in outcomes). The implicit assumption is that individual i and j are similar enough to serve as pseudo-observations of the unobserved potential outcomes. The c-for-benefit is an application of the c-statistic by Harrell²² as applied to predictions $\hat{\delta}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ and observations O_{ij} from the matched pairs. That is, indexing the matched pairs by $p, q \in 1, \dots, n_{pairs}$, with n_{pairs} the total number of pairs, the c-for-benefit can be written as

$$\text{cben-}\hat{\delta} = \frac{\sum_p \sum_{q \neq p} [I(O_p < O_q)I(\hat{\delta}_p < \hat{\delta}_q) + \frac{1}{2}I(O_p < O_q)I(\hat{\delta}_p = \hat{\delta}_q)]}{\sum_p \sum_{p \neq q} [I(O_p < O_q)]} \quad (7)$$

with $I(\cdot)$ denoting the indicator function; O_p and O_q the observed outcome differences within pair p and pair q respectively; and $\hat{\delta}_p$ and $\hat{\delta}_q$ the average predicted treatment effect within pair p and pair q respectively (as defined in equation (6)). Further details on the derivation of Harrell’s c are given in the Supplementary Material B. It is important to note that the given definition describes an estimator, while the estimand was not explicitly defined in the original paper.¹⁸ Throughout, our goal here is to estimate θ_d defined in Equation (4). Modification in the c-for-benefit algorithm discussed below hence pertain to different estimators, while θ_d remains the target estimand.

3.2.2 | C-for-benefit properties

Although the c-for-benefit has been applied on several occasions (eg,³⁶⁻³⁸), its properties have not been fully elucidated. Van Klaveren et al¹⁸ recommended further work on its theoretical basis and simulation studies, which we present here. In parallel with our work, Xia et al have considered related and complementary methodological aspects of the c-for-benefit, which we will also relate to here.³⁹

While the c-for-benefit was not developed with our estimand θ_d in mind (in fact, no estimand was specified¹⁸), it is helpful to further dissect what it is estimating. As described, the c-for-benefit compares concordance between differences in i) average predicted treatment effect within matched control-treated pairs $\hat{\delta}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ and ii) observed outcome differences within those same pairs O_{ij} . In general, however, $\delta_{ij}(\mathbf{x}_i, \mathbf{x}_j) \neq \mathbb{E}(O_{ij}|\mathbf{x}_i, \mathbf{x}_j)$ unless $\mathbf{x}_i = \mathbf{x}_j$. Specifically, for controls $i \in 1, \dots, n_C$ and treated individuals $j \in 1, \dots, n_T$ and writing $g_0(\mathbf{x})$ for $P(Y_i = 1|A = 0, \mathbf{X} = \mathbf{x})$ and $g_1(\mathbf{x})$ for $P(Y_i = 1|A = 1, \mathbf{X} = \mathbf{x})$,

$$\mathbb{E}(O_{ij}|\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}(Y_j|\mathbf{x}_j) - \mathbb{E}(Y_i|\mathbf{x}_i) \quad (8)$$

$$= g_1(\mathbf{x}_j) - g_0(\mathbf{x}_i) \quad (9)$$

$$= [g_0(\mathbf{x}_j) + \delta(\mathbf{x}_j)] - g_0(\mathbf{x}_i) \quad (9)$$

$$= g_1(\mathbf{x}_j) - [g_1(\mathbf{x}_i) - \delta(\mathbf{x}_i)]. \quad (10)$$

Perfect matching. In case of perfect matching (ie, $\mathbf{x}_i = \mathbf{x}_j$), it can be seen from Equations (6), (9), and (10) that $\mathbb{E}(O_{ij}|\mathbf{x}_i, \mathbf{x}_j) = \delta(\mathbf{x}_j) = \delta(\mathbf{x}_i)$ and $\hat{\delta}_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \hat{\delta}(\mathbf{x}_j) = \hat{\delta}(\mathbf{x}_i)$. Relating this to our estimand θ_d , perfect matching on \mathbf{x} provides

the required information on Δ (since $\mathbb{E}(O_{ij}|\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}(\Delta_i|\mathbf{x}) = \mathbb{E}(\Delta_j|\mathbf{x})$) and uses the correct treatment effect estimates $\delta(\mathbf{x}_i)$). Thus, the c-for-benefit based on predictions $\hat{\delta}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ and outcomes O_{ij} estimates θ_d under perfect matching.

Imperfect matching. Two matching procedures were proposed for the c-for-benefit: (i) based on $\hat{\delta}(\mathbf{x})$ (ie, minimize the distance between pairs $\hat{\delta}_i$ and $\hat{\delta}_j$), and an alternative (ii) based on the Mahalanobis distance between covariate vectors.¹⁸ In theory, matching on covariates \mathbf{x} leads to appropriate matches as described above. However, it is notoriously difficult in case of increasing dimension of \mathbf{x} and requires appropriate scaling or weighting (importance assignment) for all elements of \mathbf{x} . Matching on $\hat{\delta}(\mathbf{x})$ is one-dimensional and hence much easier, but does not necessarily lead to appropriate matches on \mathbf{x} since $\hat{\delta}(\mathbf{x})$ is generally not an injective function of \mathbf{x} (ie, multiple configurations of \mathbf{x} can give rise to the same value of $\hat{\delta}(\mathbf{x})$). In general, when matches are only approximate in terms of \mathbf{x} , $\mathbb{E}(O_{ij}|\mathbf{x}_i, \mathbf{x}_j)$ is not equal to either $\delta(\mathbf{x}_j)$ or $\delta(\mathbf{x}_i)$. Specifically, as most easily seen in Equation (9), O_{ij} will reflect treatment effect in the treated $\delta(\mathbf{x}_j)$ and differences in risk under the control condition between case i and j (ie, $g_0(\mathbf{x}_j)$ and $g_0(\mathbf{x}_i)$ may differ when $\mathbf{x}_i \neq \mathbf{x}_j$). Hence, $\mathbb{E}(O_{ij}|\mathbf{x}_i, \mathbf{x}_j) = \Delta(\mathbf{x}_j) + \xi_{ij}$, where $\xi_{ij} = g_0(\mathbf{x}_j) - g_0(\mathbf{x}_i)$ is not related to treatment but to variability in control outcome risk, and will typically have $\mathbb{E}(\xi_{ij}) \neq 0$. Also, when $\mathbf{x}_i \neq \mathbf{x}_j$ leads to $\hat{\delta}(\mathbf{x}_j) \neq \hat{\delta}(\mathbf{x}_i)$, $\hat{\delta}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ is no longer equal to either $\hat{\delta}(\mathbf{x}_j)$ or $\hat{\delta}(\mathbf{x}_i)$ (but to their average). In terms of our estimand θ_d , (i) the approximation of $P(\Delta_k < \Delta_l)$ and $P(\Delta_k > \Delta_l)$ is too variable due to ξ_{ij} , and (ii) the pairwise averaged $\hat{\delta}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ is less variable than the individual level treatment effect estimates that are to be evaluated. While the effect that this may have on rank-ordering is not straightforward, it might at least be expected that the presence of ξ_{ij} , that is unexplained by the treatment effects under evaluation, leads to a bias in $\hat{\theta}_d$ toward the neutral value 0.5.

Loss of data. Both matching procedures were proposed for 1 : 1 matching, which requires either equal groups size for both study arms or loss of data. A simple remedy that stays close to the original idea is to perform repeated analysis with random sub-samples of the larger arm.¹⁹ Alternatively, many-to-one matching (eg, full matching) or many-to-many matching⁴⁰⁻⁴² might be implemented, but none of these has been studied in the context of the c-for-benefit.

3.2.3 | C-for-benefit modifications

Matching on predicted control outcome risk. From Equation (9) it can be seen that adjusting the pairwise outcome difference O_{ij} based on known $g_0(\cdot)$ leaves just $\delta(\mathbf{x}_j)$ (in expectation). That is, $O_{ij}^* = (Y_j - g_0(\mathbf{x}_j)) - (Y_i - g_0(\mathbf{x}_i))$ has a useful expectation that equals the true individualized treatment effect for the treated individual j

$$\mathbb{E}(O_{ij}^*|g_0(\mathbf{x}_i), g_0(\mathbf{x}_j)) = \mathbb{E}(Y_j - g_0(\mathbf{x}_j)) - \underbrace{\mathbb{E}(Y_i - g_0(\mathbf{x}_i))}_0 = \delta(\mathbf{x}_j). \quad (11)$$

Analogously, from Equation (10), and similarly adjusting for known $g_1(\cdot)$, the expectation equals the true individualized treatment effect for the control individual i

$$\mathbb{E}(O_{ij}^*|g_1(\mathbf{x}_i), g_1(\mathbf{x}_j)) = \underbrace{\mathbb{E}(Y_j - g_1(\mathbf{x}_j))}_0 - \mathbb{E}(Y_i - g_1(\mathbf{x}_i)) = \delta(\mathbf{x}_i). \quad (12)$$

Adjusting for $g_0(\cdot)$ in equation (11) aims to achieve prognostic balance, which bears resemblance to prognostic score analysis.^{43,44} Conditioning on $g_1(\cdot)$ in equation (12) is just the mirror image for $g_1(\cdot)$. In practice, $g_0(\cdot)$ and/or $g_1(\cdot)$ will of course have to be estimated and the exact equalities will become approximations. However, estimates of either one provide a matching target that is (i) one-dimensional, and (ii) is a weighted function of \mathbf{x} aiming to retain just those elements that are required to reach $\mathbb{E}(O_{ij}^*|\mathbf{x}_i, \mathbf{x}_j) = \Delta(\mathbf{x}_j)$ (adjusting for $g_0(\cdot)$) or $\mathbb{E}(O_{ij}^*|\mathbf{x}_i, \mathbf{x}_j) = \Delta(\mathbf{x}_i)$ (adjusting for $g_1(\cdot)$). Hence, we implemented a 1:1 matching procedure similar to the c-for-benefit, but with two important differences. First, matching was performed based on $\hat{g}_0(\mathbf{x})$ as opposed to predicted treatment effect. Second, concordance was evaluated between the individual level $\hat{\delta}(\mathbf{x}_j)$, as opposed to the averaged $\hat{\delta}_{ij}$, and the corresponding adjusted O_{ij}^* 's, which fits our estimand θ_d under perfect matching on $g_0(\mathbf{x})$. Imperfect matching may arise from error in $\hat{g}_0(\cdot)$ and unavailable matches on the level of $\hat{g}_0(\cdot)$. We will further refer to this implementation as *cben-y*⁰. Note that a mirror image alternative could be performed when matching on $\hat{g}_1(\mathbf{x})$; the choice between the two might be guided by the expected quality in terms of prediction accuracy of $\hat{g}_0(\cdot)$ and $\hat{g}_1(\cdot)$, and the size of the group in which ITE predictions will be evaluated.

Predicted pairwise treatment effects. Recent work by van Klaveren et al²¹ and Maas et al²⁰ suggests a modification of the c-for-benefit procedure. This novel work emphasizes the benefit of 1:1 nearest neighbour matching of treated and control cases on Mahalanobis distance, since this avoids model-dependence of the matching procedure.

Furthermore, they recognize the difficulty of the original definition of predicted treatment effect for a treated-control pair (Equation (6)), and instead propose to use ‘predicted pairwise treatment effects’. The latter is defined as the predicted difference in outcome risk within the matched treated-control pair (ie, $\hat{g}_1(\mathbf{x}_j) - \hat{g}_0(\mathbf{x}_i)$). This aligns the within-pairs observed outcome differences O_{ij} and the predictions (ie, that now specifically target this outcome difference). However, as can be seen from Equations (9) and (10), this ‘predicted pairwise treatment effects’ reflects *both* the treatment effect of interest *and* the degree to which the model correctly predicts the within-pair difference in prognosis under the *same* treatment allocation (ie, matching error in g_0 or g_1). Thus, it attributes correctly predicted within-pair differences in outcome risk that are unrelated to treatment to the ‘predicted pairwise treatment effects’. In line, we expect this novel modification to overestimate θ_d . For the remainder of this paper, we will use $c_{ben-\hat{\delta}}$ to refer to the original c-for-benefit using 1:1 matching on predicted treatment effect, and we will use $c_{ben_{ppte}}$ to refer to this recently proposed modification.

3.3 | Model-based c-statistic for benefit

Extending earlier work on model-based concordance assessment in the context of risk prediction,⁴⁵ we propose model-based estimation of θ_d (Equation (4)): the concordance statistic between ITE predictions and the true difference in probabilities to observe benefit between pairs of individuals. As such, model-based estimates are used to approximate $P(\Delta_k < \Delta_l)$ in Equation (5). For randomly selected pairs $k, l \in 1, \dots, n$ ($k \neq l$), and taking $Y = 1$ to be harmful, there are five potential outcome configuration that signal more benefit for case k than for case l .

1. $Y_k^{a=1} = 0, Y_k^{a=0} = 1, Y_l^{a=1} = 0, Y_l^{a=0} = 0$ (benefit for k , no benefit for l),
2. $Y_k^{a=1} = 0, Y_k^{a=0} = 1, Y_l^{a=1} = 1, Y_l^{a=0} = 1$ (benefit for k , no benefit for l),
3. $Y_k^{a=1} = 0, Y_k^{a=0} = 1, Y_l^{a=1} = 1, Y_l^{a=0} = 0$ (benefit for k , harm for l),
4. $Y_k^{a=1} = 0, Y_k^{a=0} = 0, Y_l^{a=1} = 1, Y_l^{a=0} = 0$ (no benefit for k , harm for l),
5. $Y_k^{a=1} = 1, Y_k^{a=0} = 1, Y_l^{a=1} = 1, Y_l^{a=0} = 0$ (no benefit for k , harm for l).

The corresponding probability estimates for these patterns follow easily from the model(s) for both potential outcomes. For instance, for the first pattern: $[1 - \hat{P}(Y_k^{a=1} = 1)] \cdot \hat{P}(Y_k^{a=0} = 1) \cdot [1 - \hat{P}(Y_l^{a=1} = 1)] \cdot [1 - \hat{P}(Y_l^{a=0} = 1)]$. The sum of the five patterns is further referred to as $P_{benefit,k,l}$. Subsequently, the required elements for an estimate of θ_d can be derived. From equation (5), plugging in $P_{benefit,k,l}$ for $P(\Delta_k < \Delta_l)$, we estimate model-based concordance probability for benefit (mbcb) as

$$mbcb = \frac{\sum_k \sum_{l \neq k} \left[I(\hat{\delta}_k < \hat{\delta}_l) \hat{P}_{benefit,k,l} + \frac{1}{2} I(\hat{\delta}_k = \hat{\delta}_l) \hat{P}_{benefit,k,l} \right]}{\sum_k \sum_{l \neq k} \left[\hat{P}_{benefit,k,l} \right]} \quad (13)$$

Note that the mbcb is sensitive to error in the estimation of both $\hat{\delta}(\mathbf{x})$ and $P_{benefit,k,l}$. For the first this is clearly desirable. Regarding the latter, it is the price for the inherently unobservable individual-level treatment effects and replaces the matching error of the estimators described earlier. The model underlying $\hat{P}_{benefit,k,l}$ should be fairly robust (ie, low risk of misspecification and overfitting), since it essentially replaces the role of the observed outcomes. When evaluating training sample performance (‘apparent’ performance), both $\hat{\delta}(\mathbf{x})$ and $\hat{P}_{benefit,k,l}$ are estimated using the same data. For external validation of predictions $\hat{\delta}(\mathbf{x})$, that is, in new data, the model for $\hat{P}_{benefit,k,l}$ should be estimated in the new data. As a different use-case, the effect of changing covariate distributions (also known as case-mix) on the mbcb can be predicted under given models for $\hat{\delta}(\mathbf{x})$ and $P_{benefit,k,l}$. That is, assuming that both models are transportable to the new setting, equation (13) gives an estimate of discriminative ability in the new case-mix (ie, for a new matrix of covariate and treatment data). This is useful since c-statistics are case-mix sensitive.^{45,46} For example, it may be helpful to estimate the effect of a more homogeneous or more heterogeneous sample on the model’s discriminative ability.

4 | CALIBRATION OF INDIVIDUALIZED TREATMENT EFFECT PREDICTIONS

In addition to discrimination, calibration represents a distinct and important criterion for prediction models. We use calibration here in the sense of the work of van Calster,^{4,5} Steyerberg,² and Harrell,³ where perfect calibration describes

the situation where the expected outcome conditional on the prediction equals the prediction, across the range of predictions. Empirically, this involves regressing observed outcomes on model predictions and then evaluating the slope and intercept or a smooth fitted to the data. Note that these measures of calibration are sensitive to both bias and the degree of spread in the predictions.[‡] In case of binary outcome, calibration is typically assessed by means of logistic regression. Here we will stay with this convention, which does assume that the logistic link function is appropriate for the model under evaluation. The continuous outcome case is discussed in Supplementary Material C. Regardless of the type of outcome, the challenge is to cope with the unobserved nature of the outcome of interest in case of individualized treatment effect predictions. Several methods have previously been proposed. A common descriptive method to assess individualized treatment effect calibration is to form k groups based on predictions $\hat{\delta}$ and to compare the within-group observed and predicted treatment effect.^{11,18,19} While intuitive and straightforward, this approach requires arbitrary choices for split points and is often hampered by small group sizes. Also, matching based solutions have been proposed.²⁰ However, in line with the arguments for model-based calibration in outcome risk prediction,^{5,48} we here argue for model-based calibration of individualized treatment effect predictions. Below we discuss the estimands and estimation for such a model-based approach. Note that the estimands for the alternative methods mentioned (eg, split-group and matching approaches) are different, and that a direct comparison of the estimators would have no clear interpretation.

Model-based calibration of treatment effect predictions. The main aim is to find the calibration intercept and slope for estimated treatment effects on the linear predictor scale. The anticipated intercept β_0 and slope β_1 in case of a perfect prediction are 0 and 1 respectively, as for regular prognostic model calibration.^{2,3} Thus, the estimands of interest are regression parameters. Slopes under 1 reflect overfitting of the treatment effect predictions, and conversely, slopes over 1 reflect underfitting. Deviations of the estimated calibration intercept relate to average error in the ITE predictions (for a fixed slope).

To cope with the unobserved outcome of interest, we first assume that control outcome risk $g_0(\mathbf{x})$ is known. Then, for the n_T treated cases indexed by j in $1, \dots, n_T$,

$$\text{logit}(P(Y_j = 1)) = \beta_0 + \beta_1 \hat{\delta}_{lp}(\mathbf{x}_j) + g_{lp,0}(\mathbf{x}_j), \quad (14)$$

where $\hat{\delta}_{lp}(\mathbf{x}_j) = \text{logit}(\hat{g}_1(\mathbf{x}_j)) - \text{logit}(\hat{g}_0(\mathbf{x}_j))$ and offset $g_{lp,0}(\mathbf{x}_j) = \text{logit}(g_0(\mathbf{x}_j))$. Conversely, assuming known outcome risk in the treated $g_1(\mathbf{x})$, for the n_C control cases indexed by i from $1, \dots, n_C$,

$$\text{logit}(P(Y_i = 1)) = -\beta_0 - \beta_1 \hat{\delta}_{lp}(\mathbf{x}_i) + g_{lp,1}(\mathbf{x}_i). \quad (15)$$

Both can be combined for all observed outcomes Y (so both arms), omitting the subject indices for simplicity, as

$$\text{logit}(P(Y = 1)) = [\beta_0 + \beta_1 \hat{\delta}_{lp}(\mathbf{x})](2A_i - 1) + g_{lp,0}(\mathbf{x})A + g_{lp,1}(\mathbf{x})(1 - A) \quad (16)$$

with $A = 1$ for the treated and $A = 0$ for controls. In the remainder of the paper, we take the calibration intercept β_0 and calibration slope β_1 in equation (16) as our calibration estimands, conditioning on a sample \mathbf{x} and under known $P(Y = 1)$, $g_0(\mathbf{x})$, and $g_1(\mathbf{x})$. In practice, Y is a noisy manifestation of $P(Y = 1)$ and both $g_0(\mathbf{x})$ and $g_1(\mathbf{x})$ have to be estimated. To obtain the required estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we plug $\hat{g}_0(\mathbf{x})$ and $\hat{g}_1(\mathbf{x})$ into equation (16).

While not the focus of the current study, the assessment of calibration of predicted individualized treatment effects is also possible without the need for estimates $\hat{\mathbb{E}}(Y^{A=a} | \mathbf{x})$ (ie, $\hat{g}_0(\mathbf{x})$ and $\hat{g}_1(\mathbf{x})$). Supplementary Material E introduces an approach based on the transformed covariate method by Tian et al,¹² which has different estimands but a similar underlying idea.

5 | SIMULATION STUDY

A simulation study was performed with the aim to compare performance of the different discrimination and calibration measures for ITE predictions discussed across varying sample sizes. The simulation study was performed and reported in line with recommendations by Morris et al⁴⁹ and using R statistical software version 4.2.⁵⁰

5.1 | Simulation study procedures

Data generating mechanisms: Synthetic trial data were simulated for a trial comparing two treatments on a binary outcome. Covariates x_1 and x_2 were generated from independent standard normal distributions and treatment assignment was 1:1 and independent of \mathbf{X} . Data generating mechanism 1 (DGM-1) was based on a simple logistic model

$$\text{logit}(P(Y_i^{A=a} = 1)) = -1 - 0.75a_i + x_{i1} + 0.5a_ix_{i2}. \quad (17)$$

DGM-1 includes main effects of treatment and X_1 and an interaction between treatment and X_2 . For each of $n_{sim} = 500$ simulation runs, development data sets (referred to as data sets D for development) of size 500, 750, and 1000 were randomly drawn. Validation data sets from DGM-1 were of size 1000 for each simulation run (referred to as data sets V1 for validation in data from DGM-1). Marginal event probabilities were $P(Y^{a=0} = 1) \approx 0.31$ and $P(Y^{a=1} = 1) \approx 0.20$. Additionally, independent validation sets of $n=1000$ cases were sampled from a second data generating mechanism (DGM-2) with changes in the coefficients to reflect a different population (referred to as data sets V2)

$$\text{logit}(P(Y_i^{A=a} = 1)) = -0.5 - 0.5a_i + 0.75x_{i1} + 0.25x_{i2} + 0.25a_ix_{i1} + 0.25a_ix_{i2}, \quad (18)$$

Marginal event probabilities for the second DGM were $P(Y^{a=0} = 1) \approx 0.39$ and $P(Y^{a=1} = 1) \approx 0.31$. With differences in both average treatment effect and heterogeneity of treatment effect between DGM-1 and DGM-2, a model developed in a sample from DGM-1 should not perform well in individuals from DGM-2.

Estimands: For discrimination, our estimand was θ_d as defined in equation (5). For calibration, our estimands were β_0 and β_1 as defined in equation (16).

Methods: The ITE model fitted to the development data was a logistic regression model estimated by means of maximum likelihood of the form

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 a_i + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 a_i x_{i1} + \beta_5 a_i x_{i2}. \quad (19)$$

Discrimination performance was assessed by means of the original c-for-benefit ($\text{cben-}\hat{\delta}$), the c-for-benefit using 1:1 matching on predicted outcome risk under the control treatment ($\text{cben-}\hat{y}^0$), recently proposed c-for-benefit based on Mahalanobis distance matching and predicted pairwise treatment effect ($\text{cben}_{\text{ppte}}$), and the proposed model-based c-for-benefit (mbcb). Calibration performance was estimated according to equation (16). Note that estimation of $\text{cben-}\hat{y}^0$, the mbcb , and calibration assessment require estimates $\hat{g}_0(\mathbf{x})$ and $\hat{g}_1(\mathbf{x})$. For 'apparent' evaluation, these are predictions from the ITE model. In practice, they should be based on data not used to fit the ITE model. Therefore, in bootstrap evaluations and external data simulations, $\hat{g}_0(\cdot)$ and $\hat{g}_1(\cdot)$ were estimated according to model (19) in independent samples. Performance measures were evaluated in settings (1) apparent (based on the same sample as on which the ITE model was fitted³), (2) in interval validation using bootstrap 0.632+ adjustment (3) in interval validation using bootstrap optimism correction, (4) in external validation samples V1 generated from DGM-1, and (5) in external validation data samples V2 generated from DGM-2. A more detailed account of the procedures is available in Online Supplementary Material D. In addition to the main simulations, settings (1), (4), and (5) were also evaluated under misspecification of the ITE model with the omission of the (important) interaction between treatment and x_2 . For these simulations, $\hat{g}_0(\mathbf{x})$ and $\hat{g}_1(\mathbf{x})$ were either estimated according to model (19) as in the main simulations, or using separate regression forests per treatment arm to provide an alternative robust to misspecification.^{15,51}

Performance measures: Writing θ_s for the estimand value in simulation run s , and $\hat{\theta}_s$ for the corresponding estimate, performance measures were averaged across simulations $s \in 1, \dots, n_{sim}$ in terms of root mean squared prediction error $\sqrt{\frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} (\theta_s - \hat{\theta}_s)^2}$ and visualized in terms of mean ± 1 SD for both the errors ($\theta_s - \hat{\theta}_s$) and the absolute values ($\hat{\theta}_s$). To obtain more stable estimates for calibration outcomes in presence of extreme values, those summary statistics were computed after trimming away the ten percent most extreme values.

5.2 | Discrimination results

Figure 1 (deviations from the estimands) and Supplementary Figure F.1 (absolute value summaries) show the main simulation results with respect to the discrimination statistics. With respect to the apparent estimates, all statistics showed

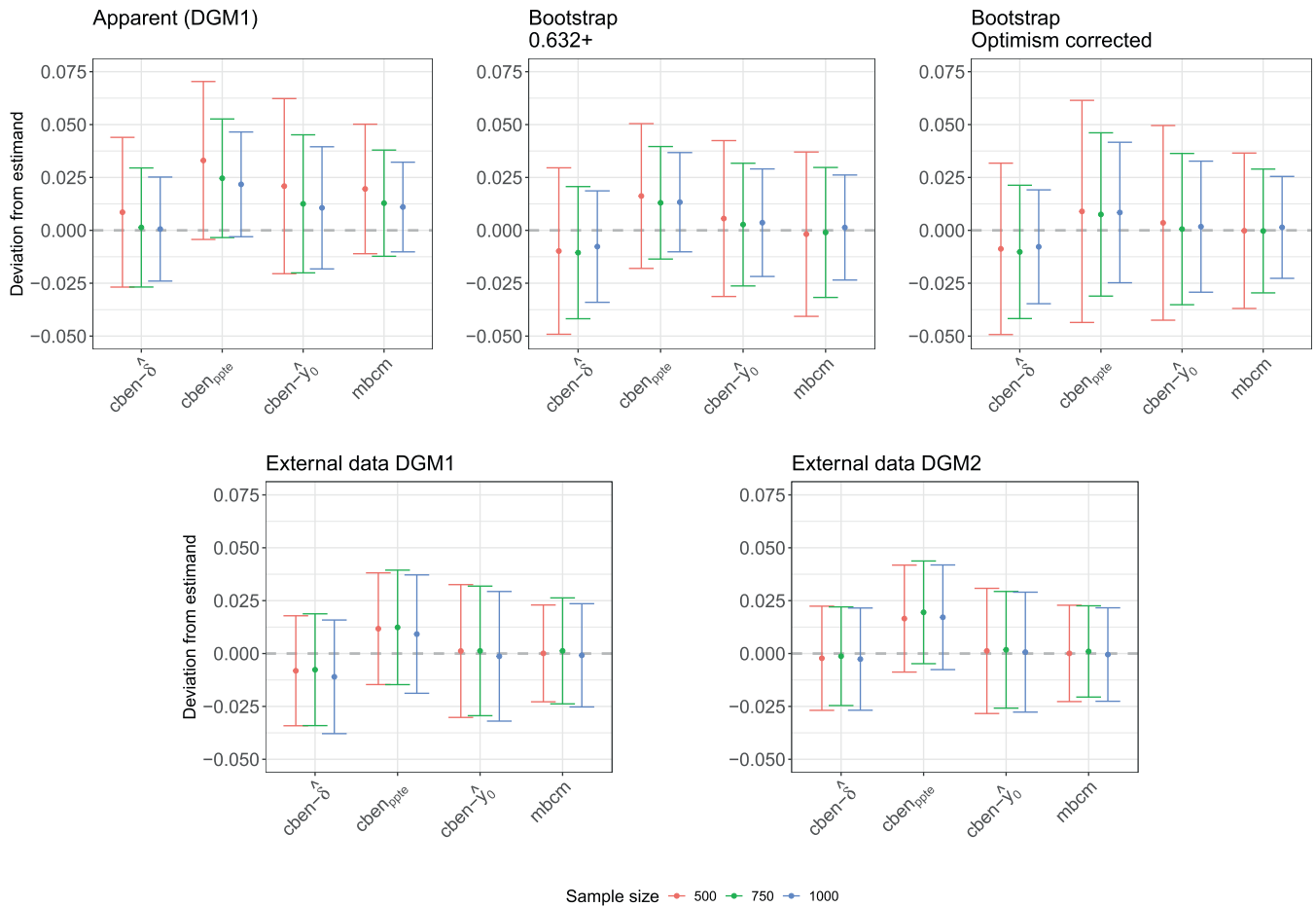


FIGURE 1 Simulation results for the discrimination statistics in terms of mean (± 1 SD) deviation from the estimand θ_d .

optimism that decreased with sample size. The original $\text{cben-}\hat{\delta}$ seemed virtually unbiased for sample sizes 750 and 1000. However, according to the results based on independent data evaluations discussed below, it was actually slightly biased downwards (in agreement with Section 3.2.2, and this canceled out the optimism here. The cben_{optnw} was most optimistic in agreement with Section 3.2.3.

As shown in the bootstrap panels in Figure 1, both types of bootstrap evaluations successfully adjusted for optimism in the apparent evaluations. On average, bias was almost eliminated from $\text{cben-}\hat{y}_0$ and the mbcb. However, in terms of accuracy, this decrease in bias was offset by increased variability as shown by the increase in rmse between apparent and bootstrap evaluations for the best performing methods (Table 1).

For assessment in independent validation samples from either DGM-1 or DGM-2, both $\text{cben-}\hat{y}_0$ and the mbcb were virtually unbiased, with the mbcb having the better rmse. As expected, the original $\text{cben-}\hat{\delta}$ was slightly pessimistic and the recently proposed cben_{optnw} was optimistic. Both had larger rmse than the mbcb.

Detailed results of the simulations under misspecification are provided in Supplementary Material F. In short, the mbcb and $\text{cben-}\hat{\delta}$ performed well, while $\text{cben-}\hat{y}_0$ was more sensitive to the choice of model used to approximate the potential outcomes in the external data.

5.3 | Calibration results

Figure 2 (deviations from the estimands) and Supplementary Figure F.2 (absolute value summaries) show the main simulation results with respect to the calibration evaluation. In line with regular calibration of outcome risk, apparent calibration assessment is not of interest, and apparent intercepts and slopes were uniformly 0 and 1 respectively. The estimand did clearly show a decrease in overfitting with increasing sample size (Figure F.2).

TABLE 1 Root mean squared error against θ_d as averaged over simulation runs for each measure and for each of the sample sizes (500, 750, and 1000).

Statistic	cben- $\hat{\delta}$			cben _{ppte}			cben- $\hat{\gamma}^0$			mbcb		
	500	750	1000	500	750	1000	500	750	1000	500	750	1000
Development data												
Apparent	0.036	0.028	0.025	0.050	0.037	0.033	0.046	0.035	0.031	0.036	0.028	0.024
0.632+	0.041	0.033	0.027	0.038	0.030	0.027	0.037	0.029	0.026	0.039	0.031	0.025
Opt. corrected	0.041	0.033	0.028	0.053	0.039	0.034	0.046	0.036	0.031	0.037	0.029	0.024
External												
DGM-1	0.027	0.027	0.029	0.029	0.030	0.029	0.031	0.031	0.031	0.023	0.025	0.024
DGM-2	0.025	0.023	0.024	0.030	0.031	0.030	0.030	0.028	0.028	0.023	0.022	0.022

Note: Bold for best performance per setting (multiple if tied).

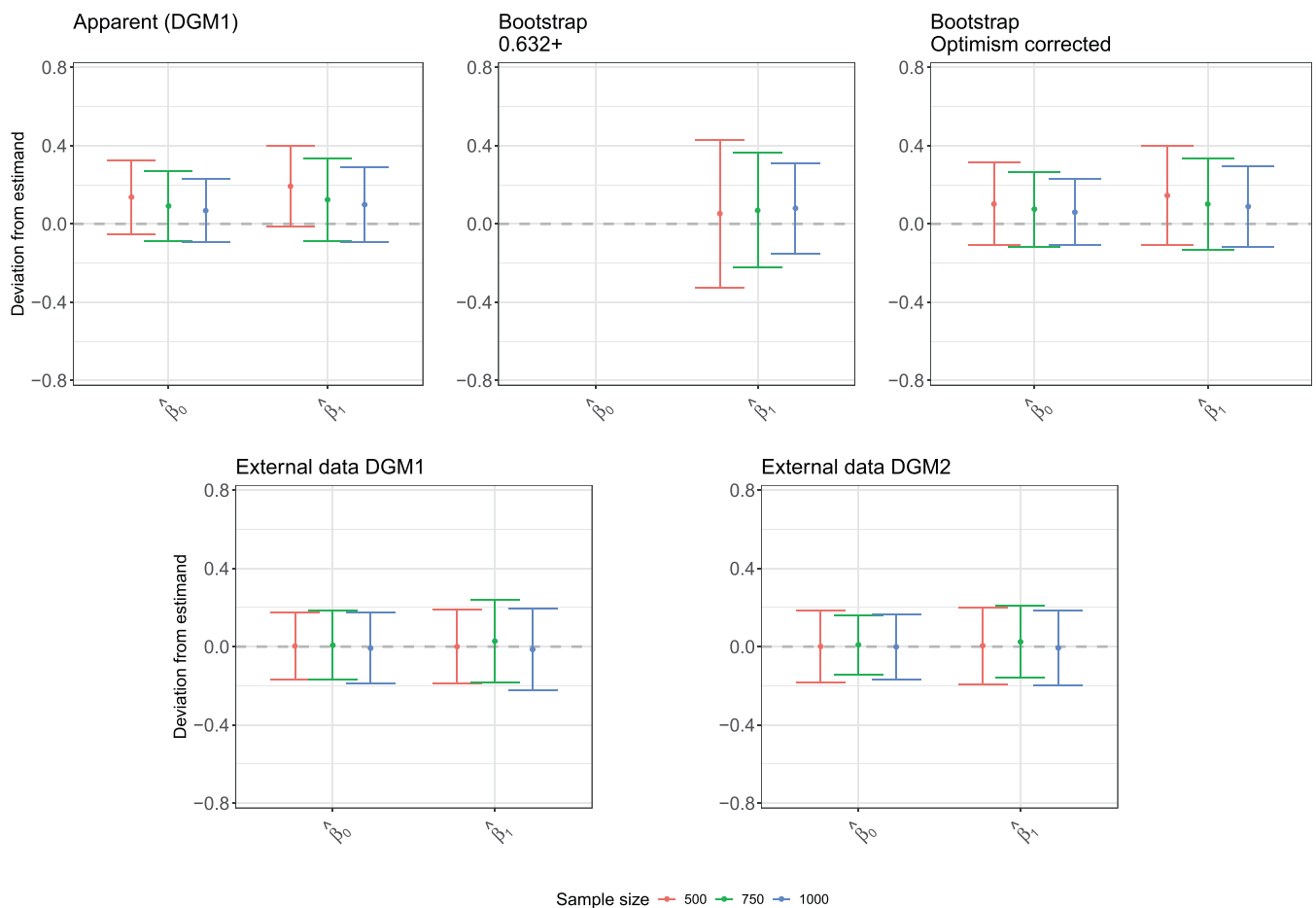


FIGURE 2 Simulation results for the ITE calibration intercept and slope estimates. Each figures shows the deviation from the estimand (10% trimmed mean \pm 1 SD).

TABLE 2 Root mean squared error against β_0 and β_1 , after ten percent trimming, and as averaged over simulation runs for each of the sample sizes (500, 750, and 1000).

Statistic	$\hat{\beta}_0$			$\hat{\beta}_1$		
	500	750	1000	500	750	1000
Development data						
Apparent	0.23	0.20	0.17	0.28	0.24	0.21
0.632+	-	-	-	0.38	0.30	0.24
Opt. corrected	0.23	0.21	0.18	0.29	0.26	0.22
External						
DGM-1	0.17	0.18	0.18	0.19	0.21	0.21
DGM-2	0.18	0.15	0.16	0.20	0.18	0.19

Both bootstrap procedures removed some optimism from the apparent estimates and showed the decreasing risk of overfitting with increasing sample size, but were still optimistic. In fact, in terms of rmse (Table 2), the bootstrap estimates were worse than the noninformative apparent evaluation. This implies that there is not enough information in a single sample to obtain reliable out-of-sample ITE calibration estimates based on this method. This is likely related to the need to estimate $g_0(\cdot)$ and $g_1(\cdot)$ in the small number of independent out-of-sample cases.

When validating a model in a new independent sample, calibration assessment was unbiased as desired (bottom panels Figure 2) in both DGM-1 and DGM-2, correctly identifying problems when applying the model to DGM-2.

Detailed results of the simulations under misspecification are provided in Supplementary Material F. In short, and in line with the main simulation results, the apparent calibration estimates were not informative, and the estimates of the calibration intercept and slope in external data agreed with the estimands on average, albeit with considerable variability.

6 | APPLIED EXAMPLE: THE THIRD INTERNATIONAL STROKE TRIAL

Patients with an ischemic stroke have sudden onset of neurological symptoms due to a blood clot that narrows or blocks an artery that supplies the brain. A key component in the emergency medical treatment of these patients includes clot-busting drug alteplase (intravenous thrombolysis recombinant tissue-type plasminogen activator).⁵²

The third International Stroke Trial (IST-3) was a randomized trial and investigated the benefits and harms of intravenous thrombolysis with alteplase in acute ischemic stroke.²⁸ This large trial included 3035 patients receiving either alteplase or placebo in a 1:1 ratio. The primary outcome was proportion of patients that was alive and independent at 6-month follow-up, which we used as outcome of interest here. Primary analyses of the treatment effect were performed with logistic regression adjusted for linear effects of age, National Institutes of Health stroke scale (NIHSS) score, time from onset of stroke symptoms to randomization, and presence (vs absence) of ischemic change on the prerandomization brain scan according to expert assessment. This analysis showed weak evidence of an effect (OR 1.13, 95% CI 0.95-1.35), but subgroup analyses suggested possibly heterogeneous treatment effect by age, NIHSS score, and predicted probability of a poor outcome.

For illustrative purposes, we here compare a main effects logistic regression model similar to the original adjusted analysis (model 1) with a model where all covariate-treatment interactions were included (model 2), and with random forest estimates.¹⁵ The outcome was coded as 0 for those independent and alive after 6 months and 1 otherwise. The included variables were treatment, age, NIHSS, time (from onset of stroke symptoms to randomization), and imaging status (presence vs absence of ischemic change on the prerandomization brain scan). In the regression models, continuous variables age, NIHSS, and time, were modeled using smoothing splines. Models 2 included covariate-treatment interactions for these variables. Model 1 and 2 were fitted using the `mgcv` package in R with default smoothing parameter selection based on generalized cross-validation.⁵³ ITE predictions were based on the difference between potential outcome predictions under alteplase and the control condition. For the random forest predictions, control outcome risk was modeled as a regression forest in the control group with the same covariates as used in model 1 and 2 (models 3a). ITE predictions were based on a causal forest with the same covariates (models 3b). Where required for evaluation purposes,

potential outcomes under the treated condition were inferred based on control predictions plus ITE predictions. Both random forests were fitted using R package **grf**⁵¹ and default settings (ie, 2000 trees, honest splitting, minimum node size 5, try all variables for splits given $p < 20$). All in all, this applied example illustrates different ways to assess the quality of individualized treatment effect predictions. The evaluated models were emphatically chosen for this purpose and were not developed in collaboration with clinical experts in the field. Hence, they are not meant to be applied in practice.

The exact parameter estimates for both models are not of key interest, but the apparent performance with respect to outcome risk prediction under allocated treatment was good: c-statistics were 0.826 (model 1), 0.831 (model 2), and 0.818 (model 3a/b), with corresponding Brier scores of 0.160, 0.158, and 0.163, and Nagelkerke R^2 of 0.389, 0.402, and 0.370. Important to note, only the random forest implementations directly provide out-of-bag estimates for the training data, so their apparent evaluation metrics should be less optimistic. However, optimism in outcome predictions should be small given the very large n to p ratio. Differences between methods were small and Spearman's correlation between outcome risk predictions for the different methods were all > 0.97 .

The predicted ITEs, however, were very different. Model 1 predicted ITEs with median -0.020 (95% between -0.29 and 0.00), model 2 predicted ITEs with median -0.026 (95% between -0.161 and 0.120), and model 3b predicted ITEs with a median of -0.023 (95% between -0.147 and 0.113). That is, the predicted treatment effect was very similar across individuals when predicted by model 1 (assuming a constant treatment effect on the log odds scale), but not when predicted by model 2 (assuming a heterogeneous treatment effect on the log odds scale) or causal random forest model 3b. Table 3 shows the apparent and bootstrap corrected results for discrimination and calibration assessment at the ITE level for the applied example as averaged over 1000 bootstrap samples.

With respect to ITE discrimination, both apparent and bootstrap-corrected discrimination estimates favored model 2 and 3a/b over model 1, with model 1 estimates around the no discriminative ability value of 0.5. The apparent mbcb for both model 2 and 3a/b were around 0.566 and correspond to the expected c for benefit for the covariate distribution in this sample assuming the model is correct. Bootstrap corrected results generally preferred model 2, with the exception of optimism corrected mbcb (preferring the random forest-based predictions). As a general remark, boot0.632+ corrections were quite stable, while optimism correction gave larger corrections. Based on the simulation results, their accuracy is similar in large sample. All in all, model 2 and model 3a/b are clearly better than model 1 in terms of discriminative ability.

In terms of calibration, bootstrap corrected slope estimates suggested that both model 2 and model 3a/b were overfitted with respect to ITEs. The amount of shrinkage suggested varies considerably between the 0.632+ and optimism corrected estimates, with the 0.632+ estimate suggesting more shrinkage. Based on the simulation study, the 0.632+ was more accurate. Note that the calibration slope for model 1 is not estimable (since the ITEs have no variability on the logit scale) and the intercept estimate for model 1 clearly showed that the degree of predicted benefit was underestimated.

The results indicate that model 1 did not provide useful differentiation in terms of ITEs. While the discriminative ability of model 2 and model 3a/b seemed modest, clear benchmarks are lacking. With respect to treatment decisions, Table 4

TABLE 3 Applied example discrimination and calibration statistics for predicted individualized treatment effect.

Model	$\text{cben}-\hat{\delta}$	$\text{cben}_{\text{ppte}}$	$\text{cben}-\hat{y}^0$	mbcb	$\hat{\beta}_0$	$\hat{\beta}_1$
Apparent						
M1	0.488	0.536	0.489	0.510	-0.117	
M2	0.562	0.584	0.570	0.567	0.012	1.096
M3a/b	0.538	0.544	0.557	0.566	-0.023	0.788
Bootstrap 0.632+						
M1	0.489	0.563	0.499	0.505		
M2	0.535	0.584	0.559	0.536		0.484
M3a/b	0.536	0.566	0.557	0.531		0.382
Optimism corrected						
M1	0.485	0.486	0.475	0.507	-0.118	
M2	0.534	0.538	0.544	0.518	-0.047	0.900
M3a/b	0.486	0.459	0.502	0.537	-0.065	0.486

TABLE 4 Applied example: out-of-sample benefit estimates of treating according to the model (treat is benefit <0, do not treat otherwise. Median and 95% bootstrap interval are shown.

Comparison	M1	M2	M3b
Model vs opposite	-0.009 (-0.050, 0.053)	-0.031 (-0.081, 0.020)	-0.038 (-0.082, 0.010)
Model vs treat all	0.002 (-0.046, 0.050)	-0.010 (-0.056, 0.042)	-0.013 (-0.061, 0.038)
Model vs treat none	-0.013 (-0.055, 0.044)	-0.025 (-0.071, 0.027)	-0.027 (-0.073, 0.020)

shows out-of-sample treatment benefit estimates for treatment according to the model versus (i) not treated according to the model,¹⁹ (ii) treatment all, and (iii) treat none are provided. All are nonsignificant. Nonetheless, while model 2 and model 3b are very different methods, they provided similar ITE estimates that agree with respect to sign in 80% of cases. Also, patient characteristics of those predicted to have benefit agree with clinical knowledge.⁵² For model 2 comparing the 1969 patients predicted to have benefit according to model 2 ($\hat{\delta}_{model2} < 0$) with the remaining 1066 patients ($\hat{\delta}_{model2} \geq 0$), the first were older [median(IQR) age 83 (78-87) vs 73 (63-82)], had worse symptoms [median(IQR) nihss 15(10-20) vs 6(4-9)], had less delay to randomization and thus treated earlier [median(IQR) time in hours 3.5 (2.5-4.9) vs 4.2 (3.6-4.8)], were more likely to have visual infarction on imaging (43% vs 36%). These figures were very similar for model 3b. Concluding, there was insufficient signal to reliably guide treatment decisions, but the proposed measures clearly differentiated between the models with and without potential (1 vs 2 & 3a/b), and aligned with estimates of decision accuracy.¹⁹

7 | SOFTWARE

R package **iteval** (<https://github.com/jeroenhoogland/iteval>) provides an implementation of the $c_{ben-\hat{\delta}}$, $c_{ben-\hat{y}^0}$, $mbcb$, and calibration measures as defined in this paper in the freely available R software environment for statistical computing.⁵⁰ The $c_{ben_{ppite}}$ has been implemented in R package **HTEPredictionMetrics** available on GitHub (<https://github.com/CHMMAas/HTEPredictionMetrics>).²⁰

8 | DISCUSSION

Measures of calibration and discrimination have a long history in the context of prediction models for observed outcome data, especially of the binary type. However, the evaluation of prediction models for individualized treatment effects (ITE) is more challenging due to the causal nature of the predictions and the resulting unobservable nature of individualized treatment effects. In this article, we used the potential outcomes framework²⁷ to gain insight into existing performance measures,^{18,20,21} clearly defined the estimands of interest, and developed model-based measures of discrimination and calibration for ITE prediction models. The model-based proposals avoid the need for matching and at the same time avoid the bias associated with existing measures, and are applicable for all prediction methods that can provide predictions for both potential outcomes. Measures of discrimination provide insight into the degree to which the model correctly ranked the predicted treatment benefit, while measures of calibration provide information on bias and over/underfitting. While the primary focus was on dichotomous outcomes, we also provided residual-based approaches for continuous outcome models. As such, our work provides generally applicable tools for the endeavor of evaluating ITE prediction models in randomized data.

In terms of discriminative ability, the proposed model-based c-for-benefit ($mbcb$) provides both a normal performance measure and an expected (case-mix adjusted) reference level for new data (in line with the model-based c-statistic.⁴⁵ The latter is relevant because concordance probabilities are known to be sensitive to case-mix.⁴⁶ Also, bootstrap procedures were proposed that adjust for optimism when no external data are available. In the simulation study, the $mbcb$ estimates were best in terms of both bias and root mean square error across settings. The original $c_{ben-\hat{\delta}}$ ¹⁸ has a more difficult interpretation and was downward biased in the simulation study, but was very stable throughout. In contrast, the recent $c_{ben_{ppite}}$ ^{20,21} was consistently optimistic. Our adaptation of the matching-based c-for-benefits (the $c_{ben-\hat{y}^0}$) removed the bias, but at too high a cost in terms of variability. We hypothesize that the stability of the $mbcb$ is due to the lack of need for a matching algorithm.

In terms of calibration, the potential outcomes framework provided a model-based method for evaluating ITE prediction calibration. Again, the model-based nature avoided the need for matching algorithms, and the results of the simulation study showed unbiased estimation in independent validation data. However, the main proposal depends on the accuracy of the underlying potential outcome predictions under both treatment conditions. Thus, misspecification of the potential outcome models may invalidate the proposed calibration measure. While this may seem like a significant cost, we emphasize that the medical decision making context for which these models were developed requires accurate potential outcome predictions in conjunction with ITE predictions. Therefore, we consider accurate outcome prediction models to be a prerequisite for ITE modeling, and their evaluation should be a principle part of performance evaluation in practice.^{2,3} Nonetheless, for use cases where only the ITEs are relevant, we proposed an alternative calibration method based on the work Tian et al¹² that does not depend on a potential outcome model.

A key finding for both the ITE discrimination and calibration measures was that bootstrapping procedures were able to remove optimism (ie, reduce bias), but that the increase in variance of the estimator generally led to increased root mean squared error compared to apparent evaluation. This implies that external data are needed to accurately evaluate ITE predictions. The underlying reason is the need for accurate potential outcome predictions based only on the out-of-sample cases, for which the 36.8% of out-of-sample cases in a bootstrap procedure were apparently insufficient. Nevertheless, bootstrap procedures are still to be preferable to apparent assessment when nothing else is available, as they provide a fairer estimate on average. Also, there was little overfitting in our examples due to large n to p ratios, but optimism in apparent estimates can be much more severe and can vary widely between methods.

Related work. The key literature underlying the developments in this article mainly arose from the fields of (medical) statistics and epidemiology, but there are recent and connected developments in econometrics and machine learning, some of which venture beyond randomized data and the large n to p ratio in this paper. While ITE estimation was not the focus of our work, it is worth discussing how it relates to recent machine learning methods. In particular, recent years have seen a rapid growth in regression and machine learning methods that aim to estimate individualized treatment effects directly, without estimating the potential outcomes themselves. These methods benefit in settings where the functional form of the treatment effect is less complex than the response surfaces of the potential outcomes, and thus easier to learn or model. Popular methods include transformed covariate regression¹² and in particular tree-based methods.^{14,15,54-56} In addition, there has been much interest in meta-learners that decompose the estimation of individualized treatment effects into separate prediction problems that can be approached using any prediction method (eg, regression, machine learning).^{13,16,17,57,58} Typically, these methods require prior estimation of the potential outcomes and propensity scores, which are then combined into transformed outcomes that can be regressed on the covariates to predict individual treatment effects. Prominent examples include the X learner,¹⁶ the DR learner,⁵⁸ and the R learner.¹⁷ Extensive simulation studies investigating the properties of these causal machine learning methods are available elsewhere.⁵⁹⁻⁶¹

What all of these methods have in common is that they either do not estimate potential outcomes, or treat them as nuisance parameters that are used to obtain more accurate personalized treatment effect estimates. Also, they either inherently provide shrinkage and selection techniques, or can build on methods that do, and thus have an advantage in settings with a higher tension between model complexity and sample size (eg, high-dimensional settings). This has been shown to be beneficial in terms of the accuracy of predicted ITEs, particularly for observational data and strong treatment selection.^{58,61} However, if the potential outcome predictions are of interest, separately predicting the potential outcomes and individualized treatment effects leads to the unwanted situation that $\hat{\mathbb{E}}(Y^{A=1}|\mathbf{x}_i) - \hat{\mathbb{E}}(Y^{A=0}|\mathbf{x}_i) \neq \hat{\delta}(\mathbf{x}_i)$. The pragmatic solution taken in our applied example was to separately estimate $\hat{\mathbb{E}}(Y^{A=0}|\mathbf{x})$ and $\hat{\delta}(\mathbf{x})$ and infer $\hat{\mathbb{E}}(Y^{A=1}|\mathbf{x})$ from them. While this leads to consistency between predicted potential outcomes and ITEs, it does not use the most accurate (directly estimated) model for $\hat{\mathbb{E}}(Y^{A=1}|\mathbf{x})$.

With respect to the proposed measures of discrimination and calibration for individualized treatment effects, the discrimination estimand formulated in section 3 is not within reach for methods estimating individualized treatment effects $\delta(\mathbf{x})$ without informing on $\hat{\mathbb{E}}(Y^{A=a}|\mathbf{x})$. This is because the probability of observing a benefit for an individual, or observing a difference in benefit between two individuals, depends on the risk in the absence of treatment. In general, medical decision making for individuals is difficult when only the treatment effect estimate is available, and not the outcome prediction under a reference condition. Obvious applications that depend only on $\delta(\mathbf{x})$ are beyond the scope of our work, but include ranking groups according to their predicted benefit for judicious allocation of limited resources. Research specifically aimed at evaluating a model's prioritization qualities includes recent work on prioritization rules via rank-weighted average treatment effects.⁶² With respect to calibration, we have proposed an alternative method that does not depend on the potential outcome model(s), but it requires further study and comparison with a recent arXiv paper by Xu and Yadlowsky⁶³ that provides a fairly general calibration metric for methods that only provide ITEs. Their proposal examines the ℓ_p norm of the expected calibration error for predicted treatment effect heterogeneity using nonparametric methods,

which have different estimands compared to our proposal.⁶³ Also, Chernozhukov has proposed an estimation framework that targets key features of individualized treatment effect, with a particular interest in approximation based on a linear function of proxy ITE predictions from auxiliary data.¹³ Although intended for estimation and inference purposes and, unlike the current work, based on transformed outcome models, their estimation objective bears resemblance to our calibration objective without assuming a specific parametric form.

Limitations. Limitations of the current work include the relatively narrow scope of the simulation study, which was conducted primarily for illustrative purposes. We also limited our use case to settings with randomized data and models that provide individualized treatment effect estimates as the difference between two potential outcome predictions. In addition, we focused on parametric measures of discrimination and calibration in line with the existing literature and the common context of very limited sample size, but recently proposed nonparametric measures are promising.⁶³ Important questions remain regarding the relationship between discrimination and calibration at the outcome and ITE levels, and the relationship between discrimination and calibration statistics and the clinical utility of the models. With respect to uncertainty estimates, bootstrap procedures provide a viable option.

Future work. In terms of future research, it would be interesting to evaluate whether some level of grouping is beneficial for evaluating model performance. Paradoxically, the goal of precision underlying the development of ITE models may hinder the ability to evaluate them, since individual-level treatment effects are inherently unobservable. In large sample situations, it would also be interesting to allow for a more flexible estimation of the calibration slope beyond the current linear implementation, which would allow for the construction of E-statistics and an integrated calibration index.^{20,64}

Conclusion. In summary, we have provided a principled review of existing measures of discrimination and calibration for models predicting individualized treatment effects, and proposed model-based methods that avoid the need for matching and reduce bias. Further research is needed to improve understanding of the precise properties of these measures under different conditions of sample size, degree of treatment effect heterogeneity, and explained variation, and to explore the relationship with novel estimators for related estimands in the machine learning literature.

ACKNOWLEDGEMENTS

This project received funding from the European Union's Horizon 2020 research and innovation program under ReCoDID grant agreement No 825746. Jeroen Hoogland and Thomas P. A. Debray acknowledge financial support from the Netherlands Organisation for Health Research and Development (grant 91215058). Thomas P. A. Debray also acknowledges financial support from the Netherlands Organisation for Health Research and Development (grant 91617050). Orestis Efthimiou was supported by the Swiss National Science Foundation (Ambizione grant number 180083). We like to thank the researchers involved in the original stroke trial for use of their data.^{28,65}

DATA AVAILABILITY STATEMENT

Data for the International Stroke Trial-3 applied example are publicly available.⁶⁵ R package **iteval** is available on GitHub (<https://github.com/jeroenhoogland/iteval>) and provides functions to derive the $\text{cben-}\hat{\delta}$, $\text{cben-}\hat{\gamma}^0$, mbcb , and calibration measures as defined in this paper. Github repository **iteval-sims** (<https://github.com/jeroenhoogland/iteval-sims>) provides the required files and instructions for replication of the simulation study.

ENDNOTES

*The original paper did not focus on the required conditions for *causal* interpretation of the predicted individualized treatment effects; here we assume that these assumptions, as described in Section 2.1, are met.


†Where the distance between \mathbf{x}_i and \mathbf{x}_j is defined as $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$ with \mathbf{S} the covariance matrix of the covariates \mathbf{x} .

‡Stevens and Poppe provide an interesting overview of different uses of the term calibration in different disciplines.⁴⁷

ORCID

J. Hoogland  <https://orcid.org/0000-0002-2397-6052>

O. Efthimiou  <https://orcid.org/0000-0002-0955-7572>

T. L. Nguyen  <https://orcid.org/0000-0002-6376-7212>

T. P. A. Debray  <https://orcid.org/0000-0002-1790-2719>

REFERENCES

1. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361-387.

2. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health, Cham: Springer International Publishing; 2019.
3. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. 2nd ed. Cham Heidelberg New York: Springer; 2015 OCLC: 922304565.
4. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176.
5. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230.
6. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*. 2018;363:k4245.
7. Kent DM, Paulus JK, van Klaveren D, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. *Ann Intern Med*. 2020;172:35.
8. Senn S. Statistical pitfalls of personalized medicine. *Nature*. 2018;563(7733):619-621.
9. Rekkas A, Paulus JK, Raman G, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Med Res Methodol*. 2020;20:264.
10. Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagnos Prognos Res*. 2021;5:3.
11. Hoogland J, Int'Hout J, Belias M, et al. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Stat Med*. 2021;40:9154.
12. 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;109:1517-1532.
13. Chernozhukov V, Demirer M, Duflo E, Fernández-Val I. Generic machine learning inference on heterogeneous treatment effects in randomized experiments. 2018 arXiv:1712.04802 [econ, math, stat].
14. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2018;113:1228-1242.
15. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Stat*. 2019;47:1148-1178.
16. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci*. 2019;116:4156-4165.
17. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*. 2021;108:299-319.
18. 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.
19. Efthimiou O, Hoogland J, Debray TP, et al. Measuring the performance of prediction models to personalize treatment choice. *Stat Med*. 2023;42:1188-1206.
20. Maas CCHM, Kent DM, Hughes MC, Dekker R, Lingsma HF, Van Klaveren D. Performance metrics for models designed to predict treatment effect. *BMC Med Res Methodol*. 2023;23(1):165.
21. Van Klaveren D, Maas CCHM, Kent DM. Measuring the performance of prediction models to personalize treatment choice: defining observed and predicted pairwise treatment effects. *Stat Med*. 2023;42:4514-4515.
22. Harrell FE, Califf RM. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):4.
23. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaut P. Reporting of sample size calculation in randomised controlled trials: review. *BMJ*. 2009;338:b1732.
24. Sackett DL, Rosenberg WM, Gray J, Haynes B, Richardson W. Evidence based medicine: what it is and what it isn't: it's about integrating individual clinical expertise and the best external evidence. *BMJ*. 1996;312(7023):71-72.
25. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA*. 2017;318:1377.
26. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701.
27. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc*. 2005;100:322-331.
28. The IST-3 collaborative group. The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. *Lancet*. 2012;379:2352-2363.
29. Holland P. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945-960.
30. Hernán MA, Robins JM. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.
31. Lamont A, Lyons MD, Jaki T, et al. Identification of predicted individual treatment effects in randomized clinical trials. *Stat Methods Med Res*. 2018;27(1):142-157.
32. Bobbio M, Demichelis B, Giustetto G. Completeness of reporting trial results: effect on physicians' willingness to prescribe. *Lancet*. 1994;343:1209-1211.
33. Naylor CD, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann Intern Med*. 1992;117:916-921.
34. Sorensen L, Gyrd-Hansen D, Kristiansen IS, Nexøe J, Nielsen JB. Laypersons' understanding of relative risk reductions: randomised cross-sectional study. *BMC Med Inform Decis Mak*. 2008;8:31.
35. Murray EJ, Caniglia EC, Swanson SA, Hernández-Díaz S, Hernán MA. Patients and investigators prefer measures of absolute risk in subgroups for pragmatic randomized trials. *J Clin Epidemiol*. 2018;103:10-21.

36. Bress AP, Greene T, Derington CG, et al. Patient selection for intensive blood pressure management based on benefit and adverse events. *J Am Coll Cardiol*. 2021;77:1977-1990.
37. Olsen MK, Stechuchak KM, Oddone EZ, Damschroder LJ, Maciejewski ML. Which patients benefit most from completing health risk assessments: comparing methods to identify heterogeneity of treatment effects. *Health Serv Outcomes Res Methodol*. 2021;21:527-546.
38. Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical value of predicting individual treatment effects for intensive blood pressure therapy: a machine learning experiment to estimate treatment effects from randomized trial data. *Circ Cardiovasc Qual Outcomes*. 2019;12:e005010.
39. Xia Y, Gustafson P, Sadatsafavi M. Methodological concerns about “concordance-statistic for benefit” as a measure of discrimination in predicting treatment benefit. *Diagnos Prognos Res*. 2023;7:10.
40. Rosenbaum PR. A characterization of optimal designs for observational studies. *J Royal Stat Soc B Methodol*. 1991;53:597-610.
41. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99:609-618.
42. Colanino J, Damian M, Hurtado F, et al. Efficient many-to-many point matching in one dimension. *Graphs Combinato*. 2007;23:169-178.
43. Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008;95:481-488.
44. Nguyen T, Debray TP. The use of prognostic scores for causal inference with general treatment regimes. *Stat Med*. 2019;38:2013-2029.
45. van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings: a new concordance measure for external validation of risk models. *Stat Med*. 2016;35:4136-4152.
46. Nieboer D, van der Ploeg T, Steyerberg EW. Assessing discriminative performance at external validation of clinical prediction models. *PLoS One*. 2016;11:e0148820.
47. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the “calibration slope” really measure? *J Clin Epidemiol*. 2020;118:93-99.
48. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. *Stat Methods Med Res*. 2016;25:1692-1706.
49. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods: Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):1-29.
50. R Core Team. R: a language and environment for statistical computing. 2022.
51. Tibshirani J, Athey S, Sverdrup E, Wager S. grf: Generalized Random Forests. 2023.
52. Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American heart association/American stroke association. *Stroke*. 2019;e344-e418:50.
53. Wood SN. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science. 2nd ed. Boca Raton: CRC Press/Taylor & Francis Group; 2017.
54. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20:217-240.
55. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci*. 2016;113:7353-7360.
56. Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal*. 2020;15:965-1056.
57. Imai K, Li ML. Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments. 2022 arXiv:2203.14511 [stat].
58. Kennedy EH. Towards optimal doubly robust estimation of heterogeneous causal effects. 2023 arXiv:2004.14497 [math, stat].
59. Knaus MC, Lechner M, Strittmatter A. machine learning estimation of heterogeneous causal effects: empirical Monte Carlo evidence. *Econ J*. 2021;24:134-161. arXiv:1810.13237 [econ].
60. Jacob D. Cross-fitting and averaging for machine learning estimation of heterogeneous treatment effects. 2020 arXiv:2007.02852 [stat].
61. Okasa G. Meta-learners for estimation of causal effects: finite sample cross-fit performance. 2022 arXiv:2201.12692 [econ, stat].
62. Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S. Evaluating treatment prioritization rules via rank-weighted average treatment effects. 2021 arXiv:2111.07966 [stat].
63. Xu Y, Yadlowsky S. Calibration error for heterogeneous treatment effects. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022*, Valencia, Spain. PMLR: Vol 151; 2022.
64. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med*. 2019;38:4051-4065.
65. Sandercock P, Wardlaw J, Lindley R, Cohen G, Whiteley, W. *The third International Stroke Trial (IST-3), 2000-2015 [dataset]*. University of Edinburgh & Edinburgh Clinical Trials Unit; 2016. doi:10.7488/ds/1350

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hoogland J, Efthimiou O, Nguyen TL, Debray TPA. Evaluating individualized treatment effect predictions: A model-based perspective on discrimination and calibration assessment. *Statistics in Medicine*. 2024;1-18. doi: 10.1002/sim.10186