https://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Articles in Advance, pp. 1–22
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

Calibration of Heterogeneous Treatment Effects in Randomized Experiments

Yan Leng,[a,*] **Drew Dimmery**[b]

[a] McCombs School of Business, The University of Texas at Austin, Austin, Texas 78705; [b] Research Network Data Science, University of Vienna, 1090 Vienna, Austria

*Corresponding author

Contact: yan.leng@mccombs.utexas.edu, https://orcid.org/0000-0002-7084-2700 (YL); drew.dimmery@univie.ac.at, https://orcid.org/0000-0001-9602-6325 (DD)

Received: June 28, 2021
Revised: August 31, 2022; June 26, 2023; October 9, 2023
Accepted: October 13, 2023
Published Online in Articles in Advance: January 12, 2024

https://doi.org/10.1287/isre.2021.0343

Abstract. Machine learning is commonly used to estimate the heterogeneous treatment effects (HTEs) in randomized experiments. Using large-scale randomized experiments on the Facebook and Criteo platforms, we observe substantial discrepancies between machine learning-based treatment effect estimates and difference-in-means estimates directly from the randomized experiment. This paper provides a two-step framework for practitioners and researchers to diagnose and rectify this discrepancy. We first introduce a diagnostic tool to assess whether bias exists in the model-based estimates from machine learning. If bias exists, we then offer a model-agnostic method to calibrate any HTE estimates to known, unbiased, subgroup difference-in-means estimates, ensuring that the sign and magnitude of the subgroup estimates approximate the model-free benchmarks. This calibration method requires no additional data and can be scaled for large data sets. To highlight potential sources of bias, we theoretically show that this bias can result from regularization and further use synthetic simulation to show biases result from misspecification and high-dimensional features. We demonstrate the efficacy of our calibration method using extensive synthetic simulations and two real-world randomized experiments. We further demonstrate the practical value of this calibration in three typical policy-making settings: a prescriptive, budget-constrained optimization framework; a setting seeking to maximize multiple performance indicators; and a multitreatment uplift modeling setting.

History: Ravi Bapna, Senior Editor; Gordon Burtch, Associate Editor.
Funding: This work was supported by the National Science Foundation, Division of Information and Intelligent Systems [Grant IIS 2153468].
Supplemental Material: The online appendices are available at https://doi.org/10.1287/isre.2021.0343.

Keywords: causal inference • heterogeneous treatment effects • randomized experiments • calibration • machine learning

1. Introduction

Randomized experiments have become essential tools in various domains, including academia, economics, and the medical field, as well as the technology industry where A/B testing is widely used by companies like Facebook, Uber, and Spotify for decision making and product optimization (Markov et al. 2021, Wu et al. 2022). A distinguishing characteristic of social phenomena is the inherent variability and heterogeneity among individuals (Xie 2007). Consequently, individual responses to policies, treatments, and stimuli also exhibit differences (Xie et al. 2012). Estimating heterogeneity in treatment effects (HTEs) has emerged as a pervasive research and practical problem across diverse fields in the social sciences and healthcare (Imai and Strauss 2011, Wager and Athey 2018). In recent years, the field of causal inference has embraced flexible estimation tools derived from machine learning to tackle the challenges associated with estimating HTEs (Hill 2011; Imai and Ratkovic 2013; Athey and Imbens 2015, 2016; Künzel et al. 2019; Chernozhukov et al. 2023; Kennedy 2023).

Our paper introduces a diagnostic tool for assessing whether these HTE estimates are "well calibrated." Additionally, we develop a model-agnostic[1] calibration approach, which is a process that reduces dependence between true HTEs and errors in HTE estimates using model-free subgroup estimates from a randomized experiment. Accurate HTE estimation is crucial for achieving four key goals: (1) investigating heterogeneity in the treatment as part of the basic scientific understanding of the experiment's mechanism (Athey and Imbens 2015); (2) assessing whether the experiment can be generalized to a different population (Athey and Imbens 2015); (3) selecting the best intervention for targeted individuals (Prosperi et al. 2020); and (4) designing policies that maximize utilities for

the policy maker (McFowland et al. 2021). These goals are prevalent in various fields.

Example 1 (IS, Marketing, and Public Policy). Decision makers face the challenge of constrained optimization when allocating individuals to different treatment options to maximize utility, considering the ex ante unknown and heterogeneous benefits and costs of these options (McFowland et al. 2021). For instance, a marketer may incentivize past consumers to generate referrals, where the costs and benefits are not known beforehand (Jung et al. 2020). A similar application can be found in stimulating blood donations in public policy (Sun et al. 2019). In this context, decision makers aim to achieve goals (1) and (4).

Example 2 (Healthcare). Precision medicine, a tailored healthcare initiative promoted by the White House[2] (Prosperi et al. 2020), aims to choose a personalized procedure among multiple options that maximize the probability of a favorable outcome for patients. For example, doctors may need to decide between administering amoxicillin or cephalosporin for an upper respiratory tract infection, considering factors like minimizing harm and maximizing efficacy. In this application, achieving goals (1), (2), and (3) are critical for the doctor.

Given the critical role of accurate HTE estimation in diverse fields, a variety of methodologies have been explored. Machine learning methods, in particular, have garnered substantial interest. Yet, despite the promise and interest, the application of machine learning to HTEs is not without challenges. The fundamental problem associated with counterfactual prediction—that only one potential outcome is observed for each individual—makes HTE estimation inherently difficult (Holland 1986). This inherent constraint makes the translation of HTE estimation into a traditional supervised learning problem impossible. Absent the individual treatment effects (ITEs), the task must be approximated.

Recent work in causal inference has used flexible estimation tools from machine learning to better estimate conditional causal effects (Hill 2011, Wager and Athey 2018, Künzel et al. 2019, Nie and Wager 2021, Kennedy 2023). Common approaches tend to focus on plugging in machine-learned response (i.e., outcome) surfaces directly. However, these surfaces are not the actual quantities of interest; the causal effects are. The mismatch between analyzing causal effects (the task of interest) and the task for which machine learning methods are particularly appropriate—response surface modeling—may lead to poor bias/variance tradeoffs in practice. Meanwhile, regularization of the response surface models may not be ideal for causal effect estimation (Hahn et al. 2018), even though traditional machine learning methods navigate that tradeoff well in the supervised setting. In particular, regularization, which is typically one of the great benefits of machine learning, can have negative implications for estimating HTEs and may lead to bias (Hahn et al. 2018).

This paper uncovers the miscalibration issue in HTE models by using two real-world randomized control trials (RCTs) and synthetic scenarios. Miscalibration occurs when the model-based conditional average treatment effects (CATEs) do not align with the *model-free* difference-in-means (DM). We introduce a diagnostic approach to detect miscalibration using a quantile-quantile (Q-Q) plot. This diagnostic plot is based on *model-free* difference-in-means (DM) conditional average treatment effects (CATEs) and *model-based* CATEs aggregated from a machine learning model. These model-free DM estimates are nonparametrically identified (owing to the RCT) using a DM estimator, which is often seen as the gold standard in RCTs. We fit a regression on the estimated HTEs to "calibrate" them, using additive and multiplicative scaling to align them with model-free CATEs. The calibration process can be viewed as regressing the model-free CATE on the estimates implied by the HTE model. The intercept and the slope of this regression define the additive and multiplicative scaling necessary to align the model-based CATE with the model-free CATE. The flexibility and model-agnostic nature of this approach makes it well suited for digital RCTs at scale. This paper offers three key contributions:

1. We highlight the issue of miscalibrated causal effects using RCTs by Facebook and Criteo AI Laboratory. We demonstrate that the model-based CATEs from many machine learning models may provide a poor estimator of and differ substantially from the model-free CATEs. This observation suggests a potential bias in many machine learning methods. We explore three mechanisms—regularization and misspecification in the response model, along with the application of causal forest in high-dimensional contexts—to elucidate why standard HTE models may yield biased estimates. Moreover, Proposition 2 reveals that a high degree of regularization directly translates into miscalibration when a T-learner, accompanied by ridge regression as the base learner, is used. This bias can be effectively mitigated through calibration.

2. We propose a two-step procedure to investigate this calibration issue. The first step is a diagnostic test, a Q-Q plot of the model-free CATE and model-based CATE, to investigate whether the standard HTE models align well with the model-free subgroup estimates. We recommend practitioners and researchers to implement this diagnostic test after they run an HTE model to determine whether the estimates are calibrated.

3. Our framework's second step is a model-agnostic calibration method for calibrating *any* HTE models. We prove that this calibration method provides the best

linear predictor (BLP) of the model-free subgroup effects from randomized experiments (Proposition 1). With both synthetic simulations and real-world randomized experiments, we show that calibration can improve the ITEs for a broad class of meta-algorithms and tree-based HTE methods. Furthermore, we conduct comprehensive policy simulations, solidifying the practical value of calibrated HTE estimation.

2. Related Work

We review related literature in this section. We first review the literature streams on two main methods for HTE estimation: meta-algorithms and sample splitting. We conclude this section with a discussion of the importance of HTE estimations for data-driven decision making.

2.1. Meta-Algorithms for HTE Estimation

Meta-algorithms, or meta-learners, offer a flexible framework that combines supervised learning models to estimate HTEs. These methods allow for base learners, the machine learning models used for response predictions or treatment assignments, to adopt any form. The S-learner is a simple HTE model that applies a single machine learning model to predict responses y based on treatment w and covariates $\mathbf{x}$ (Hill 2011, Green and Kern 2012): $\mu(\mathbf{x}, w) = \mathbb{E}[y|\mathbf{X} = \mathbf{x}, W = w]$. The CATE estimate is computed as $\hat{\tau}(\mathbf{x}) = \hat{\mu}(\mathbf{x}, 1) - \hat{\mu}(\mathbf{x}, 0)$.

Within this framework, Imai and Ratkovic (2013) provide a model that applies regularization separately to baseline covariates and covariate-treatment interactions. Grimmer et al. (2017) propose to estimate the response surface through an ensemble of models selected using SuperLearner (Van der Laan et al. 2007). It should be noted, however, that this ensemble optimizes performance on the response surface and not necessarily on the treatment effect.

The most widely applicable method for predicting HTEs is the T-learner due to its simplicity (Athey and Imbens 2015, Künzel et al. 2019, Jacob 2020). It uses two models for the responses of the treatment $\mu_1(\mathbf{x})$ and control group $\mu_0(\mathbf{x})$ given covariates $\mathbf{x}$:

$$\mu_1(\mathbf{x}) = \mathbb{E}[y|\mathbf{X} = \mathbf{x}, W = 1] \text{ and}$$
$$\mu_0(\mathbf{x}) = \mathbb{E}[y|\mathbf{X} = \mathbf{x}, W = 0]. \quad (1)$$

The CATE is then computed as $\hat{\tau}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x})$.

Another widely adopted method is the causal forest, which is closely related to the two approaches described previously, which fits a single random forest (e.g., S-learner). A crucial difference is that the causal forest enforces splits on treatment just before the terminal nodes and integrates this into its splitting criteria elsewhere in the tree (Wager and Athey 2018). It also uses sample splitting ("honesty") to provide guarantees on unbiasedness of some subgroup effects.

More recent meta-algorithms extend the modeling capabilities beyond just response functions to include treatment probabilities as well. For example, the X-learner, designed to manage imbalances between control and treatment groups, uses control group information to enhance treatment group estimates (Künzel et al. 2019). This approach is achieved by modeling the difference between observed outcomes and imputed counterfactuals and applies this in a vice versa fashion. Similarly, the R-learner seeks to mitigate selection bias that could arise from observed covariates using orthogonalization techniques (Nie and Wager 2021). This algorithm initially estimates the two nuisance functions, the conditional outcome mean and the propensity score, and then targets a loss function that separates the causal effects of interest from these nuisance components. A more recent development is the doubly-robust CATE estimator known as the DR-learner. This algorithm extends the T-learner and incorporates a version of inverse probability weighting on the residual of the response function models for both the control and treatment groups (Kennedy 2023). Details of these learners can be found in Online Appendix A.

A common trait of these meta-learners is that the supervised learning models they are built upon navigate a different bias/variance tradeoff than the one that would be optimal for the causal estimation task. Schuler et al. (2018) make this distinction clear in their comparison of a number of estimators of risk in the causal setting. Typically, individual models of the outcome (and of the propensity score in observational settings) are estimated by independently minimizing their loss functions (referred to as μ-risk by Schuler et al. (2018)). However, minimizing the loss function does not imply minimization of causal error—the gap between estimated and true effects. Indeed, these two do not typically align. As an illustration, Kennedy (2023) demonstrates that a transformed outcome meta-regression can, in theory, asymptotically match an optimal causal error, but this property does not necessarily hold for finite samples. In a simplified case of local polynomial regression in Theorem 3 of Kennedy (2023), *undersmoothing* the estimation of the propensity score is necessary to optimally reduce causal error: Bias must be reduced faster than in the case of a standard supervised learning problem. Schuler et al. (2018) provide a number of heuristics for model selection, but these heuristics entail additional assumptions and do not equate to causal error (which cannot be directly measured). Even methods with desirable asymptotic properties may not retain these properties in finite samples. The discrepancy between causal error and estimators can result in inappropriate bias/variance tradeoffs and consequently finite sample bias. We develop methods that help practitioners and applied researchers to determine whether the estimators they use on the data

they have in front of them has bias—not whether their approach *would* hold bias if they continue to collect more data.

We make several contributions to this literature. We first raise the awareness of the calibration issue using two real-world RCTs. We introduce a Q-Q diagnostic plot to assess whether bias exists in HTE estimates from a certain machine learning model on a given data set. Furthermore, we contribute a model-agnostic approach that is broadly applicable to *any* of the HTE methods in this literature. Our method requires no additional data beyond what is necessary for estimating HTEs. It can be scaled to arbitrarily large datasets. It also can be easily plugged into the experimentation infrastructure of tech firms and added to the analysis pipeline of applied researchers.

2.2. Sample Splitting for HTE Estimation

The second stream of literature uses sample splitting to estimate and infer the HTEs, using linear, semiparametric regression or tree-based methods to characterize HTEs. Athey and Imbens (2016) use a Horvitz-Thompson transformation on the outcome variable to estimate heterogeneity, most importantly showing that "honesty"—estimating splits and predicting on different subsets of the data—provides valid inference in this setting. Crucially, the approach of honesty provides unbiased estimates of leaf-specific average effects (in finite samples) when the covariates are low-dimensional relative to the sample size and are uniformly distributed within leaves. This does not, however, follow for *all* possible subgroup effects. Subject to regularity conditions, honesty implies *asymptotically* unbiased estimates of all HTEs as in section 3.2 of Wager and Athey (2018) so long as the size of individual leaves becomes increasingly small.[3] Other research also provide noteworthy methodologies for HTE estimation. Chernozhukov et al. (2018) estimate the sorted effect, which is a collection of estimated partial effects, ranked in increasing order and indexed by percentiles, that represent the heterogeneous effects. Zhao et al. (2017a) combine semiparametric regression and postselection inference for high-dimension regression. This method uses semiparametric regression to remove confounding bias and to increase the power of discovering effect variation. It then uses postselection inferential tools to examine whether a certain covariate interacts with the treatment, thus ascertaining the existence of effect modification. Chernozhukov et al. (2023) provide a generic method for estimating and performing inference; they use arbitrary HTE models that involve sample splitting, an approach that aligns closely with ours. Unlike existing works, we focus specifically on ensuring that aggregate subgroup effects align with benchmarks supplied by an RCT. In contrast to Chernozhukov et al. (2023), we begin from a perspective of *diagnostics* for a given model of HTEs through a Q-Q plot that we recommend. This is motivated by an empirical application in which extant methods provided poor characterizations of experimental heterogeneity. Rather than rely on the individual-level data as in Chernozhukov et al. (2023), we provide an approach to resolve these diagnosed problems using aggregated data, which can then be implemented and deployed simply through the use of the linear rescaling procedure defined in Section 4.

Dwivedi et al. (2020) consider a diagnostic approach similar to what we present here, but our work is distinctive from theirs in two primary ways. First, Dwivedi et al. (2020) focus on identifying subgroups that have larger than "average" treatment effects, whereas we focus on a different problem: improving the calibration of HTEs. This differentiation allows our method to be applied in policy-making settings, enabling the balancing of within-treatment and between-treatment heterogeneity (discussed in greater detail in Section 6). Such functionality is not offered by the subgroup discovery approach in Dwivedi et al. (2020). Furthermore, our contribution extends beyond diagnostic insights. We provide practical solutions to the identified issues and provide guidance on analytical choices that practitioners face, such as choosing bin sizes in this setting.

2.3. Causal Decision Making and HTE Estimation for Data-Driven Decision Making

Causal decision making (CDM), which entails deciding on the application of a specific intervention to a given individual, is a common reason for undertaking causal effect estimations (CEE). However, as astutely observed by Fernández-Loría and Provost (2022b), CDM and CEE may not be synonymous in certain contexts. More intriguingly, they put forth a novel proposition that certain CDM problems may not necessitate precise causal statistical modeling. In such situations, the decision-making process could potentially be substantially streamlined for policy makers (Fernández-Loría and Provost 2022a, Fernández-Loría et al. 2023).

Yet, in a variety of scenarios, obtaining treatment effect estimates remains essential. For instance, McFowland et al. (2021) devise a prescriptive framework for investigating a generalized budget-constrained optimization issue, with benefits and costs being unknown ex ante. Moreover, organizations and digital platforms generally endeavor to optimize multiple performance indicators, necessitating the simultaneous optimization of various metrics (Diemert et al. 2018). This presence of tradeoffs exemplifies the majority of practical decision-making problems encountered in the industry (Deng and Shi 2016, Letham et al. 2019). When more than one treatment effect needs to be estimated from one or more experiments, decision makers must be cognizant of, and reconcile, various forms of heterogeneity within and between these experiments. Such applications are prevalent in practical budget-constrained

optimizations, where achieving calibrated HTEs is a prerequisite. In reality, decision makers need to account for the context of decision-making when selecting data-driven decision-making methodologies (Fernández-Loría and Provost 2022c). Calibrated treatment effects and CDM nicely complement one another because of their distinct and specialized use cases. Moreover, decision makers can opt for their preferred methodology based on their preferences for the decision-making process.[4]

3. Problem Setup

In this section, we describe the problem setup. Data are collected from an RCT, where N units are assigned to a binary treatment, w, using a fair coinflip. Let N_1 be the number of units in the treatment group, with $w_i = 1$ and N_0, analogously. We rely on the stable unit treatment value assumption (SUTVA) throughout (Imbens and Rubin 2015) and often suppress indexing by units. In addition, we collect data on outcomes,

$$\mathrm{y} = \mathrm{w}\ \mathrm{y}(1) + (1 - \mathrm{w})\ \mathrm{y}(0), \tag{2}$$

where y(1) and y(0) correspond to the potential outcomes for the treatment group and the control group, respectively. The outcome is a realization of the *potential* outcome associated with the assigned treatment status. We collect pretreatment covariates, $\mathbf{X}$, which exist in the domain $\mathcal{X}$. We denote $\mathcal{S}$ as a subset of this domain, $\mathcal{S} \subseteq \mathcal{X}$. Let $N(\mathcal{S})$ indicate the number of units having a covariate value in the set $\mathcal{S}$ (using analogous notation for the treatment and control groups). The goal is to understand the conditional expectation function of the treatment effect. In particular, we propose that one desirable property of an HTE model (in addition to high accuracy in predicting effects) is calibration.

What is calibration? Think of a scatter plot between the estimated treatment effects and the error in each unit's treatment effect estimate. If these are uncorrelated, then we can call our effect estimates calibrated. We never directly observe the error for each unit's estimate. Intuitively, calibration means that positive errors tend to be balanced by negative ones at every level of estimated effect. The size of those errors might be large, indicating low accuracy, but if they are *balanced*, then calibration is achieved. This relationship closely accords with the definition of calibration in classification and regression problems (Kuleshov et al. 2018), with the added challenge resulting from the fundamental problem of causal inference: Labels (i.e., ITEs) are never observed (Holland 1986). Many options are available for estimating the conditional expectation function of a treatment effect (Athey and Imbens 2016, Künzel et al. 2019, Nie and Wager 2021, Kennedy 2023); we remain agnostic about which of these procedures should be used. Our method requires only a black box model to estimate the conditional expectation function of a treatment effect for every value of $\boldsymbol{x}$: $\hat{\tau}(\boldsymbol{x})$.[5]

We consider two ways by which subgroup average conditional effects can be estimated; each is defined on a subset $\mathcal{S}$ of the domain of $\mathcal{X}$, where $\mathcal{S} \subseteq \mathcal{X}$. The model-free CATE estimation just takes the difference in means within the subgroup:

$$\hat{\tau}^{\mathrm{DM}}(\mathcal{S}) = \frac{1}{N_1(\mathcal{S})} \sum_{i:\mathbf{X}_i \in \mathcal{S}} y_i w_i - \frac{1}{N_0(\mathcal{S})} \sum_{i:\mathbf{X}_i \in \mathcal{S}} y_i (1 - w_i). \tag{3}$$

This estimator is annotated with "DM" to indicate that it comes from the difference in means. The typical standard error of this quantity using the Neyman estimator is defined analogously as $\hat{s}^{DM}$.

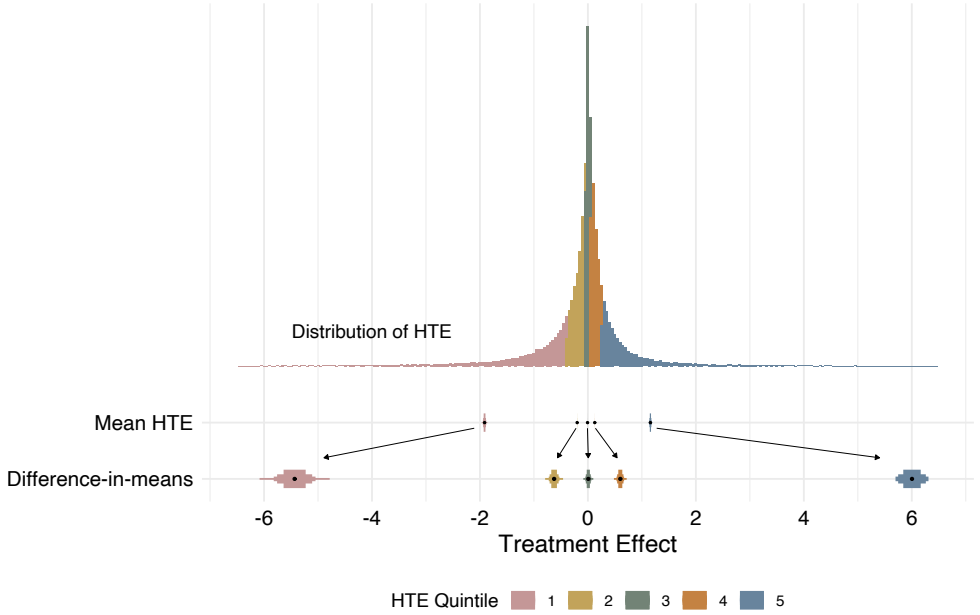
The model-based CATE estimation ($\hat{\tau}^{\mathrm{UNCAL}}(\mathcal{S})$) from a given HTE model is

$$\hat{\tau}^{\mathrm{UNCAL}}(\mathcal{S}) = \frac{1}{N(\mathcal{S})} \sum_{i:\mathbf{X}_i \in \mathcal{S}} \hat{\tau}^{\mathrm{UNCAL}}(\mathbf{X}_i). \tag{4}$$

The first method (Equation (3)) is simply the difference in observed outcomes within the subgroup. It makes no modeling assumptions when, as we have assumed, treatment is completely randomized (Aronow et al. 2021). This model-free estimator may not always be interesting (e.g., if the pretreatment covariates are not informative about heterogeneity). However, it is at least always unbiased for CATE within the subgroup $\mathcal{S}$. This estimator may be replaced by an augmented inverse propensity weighted estimator (AIPWE) for covariate adjustment, as long as it retains a lack of bias in finite samples (Zhang et al. 2008). The second method (Equation (4)) consists of marginalization over the estimated ITEs from some black box estimator of HTEs.

These two estimators differ because of the response surface models typically used in the latter. Suppose the response surface model is linear (in a T-learner[6]). In this case, the latter model is essentially just the covariate-adjusted estimator analyzed in Lin (2013). We analyze the misspecification of a linear response surface using quasi-Poisson generalized linear model in our simulation study in Section 5.1.3. In this linear case, the two models may provide different results in finite samples, but they both offer a consistent way to estimate subgroup effects (and the latter is more efficient). However, for more complicated models of $\hat{\tau}(\cdot)$, the differences can be stark because typical machine learning models may not approximate the conditional treatment effect well. For example, consider a tree-based S-learner HTE model, which tends to over-regularize estimated treatment effects to zero (Künzel et al. 2019). This model might provide effect estimates with high rank-correlation to the truth—but with severely misleading estimates of the central tendency and magnitude. We explore the effects of regularization on calibration in Sections 4.2 and 5.1.1. We now formally define the concept of linear calibration.

Figure 1. (Color online) HTE Models Can Have Poorly Calibrated Aggregate Effects



Notes. At the top is a histogram of the estimated HTEs, colored to indicate quintiles. The middle is the average HTE within each quintile. At the bottom are the model-free estimates within each quintile of the HTE.

Definition 1 (Linear Calibration). For a given HTE model, $\hat{\tau}$, and an equally sized partitioning of $\mathcal{X}$ called $\mathcal{P}$,

$$0 = \frac{1}{|\mathcal{P}|}\sum_{\mathcal{S}\in\mathcal{P}} \hat{\tau}(\mathcal{S})(\tau(\mathcal{S}) - \hat{\tau}(\mathcal{S})), \tag{5}$$

where $\tau(\cdot)$ and $\hat{\tau}(\cdot)$ are the ground-truth and the estimated conditional average treatment effects.

Calibration in Definition 1 implies that errors in subgroup effect estimates are zero, on average, and that they are uncorrelated with the estimated subgroup effects. From another perspective, we might say that linear calibration would obtain if we could run a linear regression of the true (unobservable) ITEs on some set of covariates. The residuals of this regression would be independent from the predictors, based on the standard properties of ordinary least squares (OLS) methods and the linearity of expectations. Of course, directly running this simple regression is not possible, as it is impossible to directly observe the true ITEs (Holland 1986). Although some HTE methods do exhibit this property of linear calibration (e.g., applying unregularized linear models in a T-learner when the true response functions are also linear), many do not. This especially holds true for methods that employ regularized nonlinear response functions based on machine learning. These sophisticated techniques often prioritize model simplicity and predictive accuracy on the response surfaces, which may lead to a potential bias-variance tradeoff, thereby compromising calibration. In the subsequent sections, we delve deeper into this issue and propose a solution to mitigate it. To demonstrate the practical implications of the aforementioned theoretical principles, Example 3 illustrates the calibration challenge in a real-world context, using a randomized experiment conducted on Facebook.

Example 3 (Randomized Experiment on Facebook). Figure 1 illustrates the problem of calibration on a real RCT run on Facebook. In this experiment, a T-learner model was estimated, with a random forest as the base learner (Breiman 2001). In this case, the uncalibrated HTEs (Equation (4)) are noticeably smaller in magnitude than the model-free estimators (Equation (3)). The aggregated effects estimated through the HTE model appear to preserve rank order, but they substantially understate the true magnitude of effects. Moreover, the two estimators are not consistent with one another—the differences between them are far larger than would be expected from different unbiased estimates of subgroup effects, implying that the subgroup effects from the HTE model are strongly biased. If Facebook took the biased estimates at face-value, its resulting decisions could be quite poor.

4. Method

Making effective decisions based on treatment estimates, especially when the decision makers have to prioritize multiple objectives (McFowland et al. 2021), requires knowing that these estimates are *calibrated*. Calibration is a property that errors between the true subgroup effect and the average HTE in that subgroup

will be zero on average over subgroups. Slightly more formally, if errors in subgroup effect estimates are zero on average and independent of the estimated subgroup effects, then they are calibrated. Calibration is distinct from risk minimization in machine learning: It implies that positive errors in HTE estimates are balanced by negative ones at every level of the predicted HTE. In supervised learning, this property is obtained by linear regression, but not necessarily by all learning methods (Greenfeld and Shalit 2020). We next formally introduce our calibration method.

4.1. HTE Calibration Method

Although we cannot directly observe the relationship between the ground-truth ITEs and estimated HTEs, we can observe the relationship between aggregated and noisy DM estimates and the estimated HTEs. We propose that researchers begin by examining this relationship (as we do for the Facebook experiment in Online Appendix D4). This empirical relationship forms the foundation of our method. The first step of this procedure is to partition the feature space $\mathcal{X}$ based on quantiles of the estimated HTEs. We estimate $\hat{\tau}^{\text{DM}}$ (along with the associated standard error) in each group. By comparing these completely agnostic estimates to the model-based estimates for those subgroups (the average of HTE estimates within each group), practitioners can uncover potential issues with their HTE estimates. When problems are uncovered, the practitioner can try to improve the the uncalibrated HTE estimates $\hat{\tau}^{\text{UNCAL}}$ through a linear transformation:

$$\hat{\tau}^{\text{CAL}}_{\alpha,\beta}(\mathcal{S}) = \alpha + \beta\hat{\tau}^{\text{UNCAL}}(\mathcal{S}). \tag{6}$$

We define $\hat{\tau}^{\text{CAL}}_{\alpha,\beta}(\mathbf{X})$ comparably. We estimate (α,β) by maximizing the likelihood of $\hat{\tau}^{\text{CAL}}$ under $\hat{\tau}^{\text{DM}}$. In the subsequent discussions, we simplify the notation of $\hat{\tau}^{\text{CAL}}_{\alpha,\beta}(\cdot)$ to $\hat{\tau}^{\text{CAL}}(\cdot)$ for clarity, provided the context remains unambiguous.

This estimation procedure allows us to find the additive and multiplicative factors of the estimated HTEs, providing the best approximation of the model-free estimates of the aggregated subgroup effects. In contrast to standard Platt scaling (Platt 1999), we focus on calibrating to aggregated effects due to the absence of true labels for the ITEs. This procedure does not modify the rank order of effects, which is a desirable property. In practice, our goal typically is not just to understand order statistics on a single dimension, but to trade off competing objectives on multiple dimensions. Thus, it is crucial to accurately measure the cardinal values of effects, not just their rank order as in the work of Chernozhukov et al. (2018). We show that when the HTE model (along with partitioning of the data based on the estimated HTEs) and the calibration procedure occur on separate subsets of the initial data, our procedure provides the best linear predictor (BLP) of the unbiased subgroup effects, as alternatively considered by Chernozhukov et al. (2023).

The pseudo-code for our calibration method is shown in Algorithm 1. Specifically, we segment the data into K subgroups, with an equal number of units in each group. Because we need to estimate two parameters and do not know the labels (i.e., ITEs) for any individual unit, we must aggregate the data into subgroups to determine calibration. We construct these subgroups by ordering units according to their ITEs (from the uncalibrated model) and then assigning the first $\frac{1}{K}$ of the units to the first subgroup, the subsequent $\frac{1}{K}$ of the units to the second, and so on. We need to choose an appropriate K (using parameter tuning) to estimate α and β and to avoid overfitting. Typically, a small K corresponds to poor performance in the training set, whereas a large K may lead to overfitting as a result of the bias-variance tradeoff.

We compute the model-free mean and standard errors of the subgroup treatment effect of $\mathcal{S}$, which are $\hat{\tau}^{\text{DM}}$ and $\hat{s}^{\text{DM}}$, respectively. The model-free estimates are directly derived from DM estimator, which is known to be unbiased and asymptotically normal under weak conditions (Aronow et al. 2021). Because these estimates are normally distributed according to the central limit theorem, we maximize the normal log-likelihood function to find the linear parameter α and multiplicative parameter β that maximize the log-likelihood function over the parameter space,

$$\ell(\alpha,\beta) = \sum_{i:\mathbf{X}_i\in\mathcal{S},\mathcal{S}\in\mathcal{P}} \log f(\hat{\tau}^{\text{CAL}}_{\alpha,\beta}(\mathbf{X}_i)), \tag{7}$$

where f is the probability density function of the normal distribution: $\mathcal{N}(\hat{\tau}^{\text{DM}}(\mathcal{S}),\hat{s}^{\text{DM}}(\mathcal{S}))$. Regarding our use of a likelihood-based approach to estimation, if we only need effects to attain linear calibration, we could simply estimate calibration through OLS and discard information about the variability of effects within each bin. Different subgroups may have very different standard errors.[7] The reason is that the conditional variance may vary substantially in different parts of the space, leading to wider standard errors in those regions. In essence, our likelihood-based estimation approach implies and recognizes that when the CATE is highly variable, small errors are less important than when the CATE is highly uncertain. This may be seen as the difference between estimation with OLS and weighted least squares (WLS).

Algorithm 1 (HTE Calibration)

1: **Input**: $\mathrm{y}, \mathrm{w}, K, \hat{\tau}^{\text{UNCAL}}$,
2: Partition individuals into K bins, $\mathcal{P}$, based on the estimated ITEs $\hat{\tau}^{\text{UNCAL}}$.
3: **for** $\mathcal{S} \in \mathcal{P}$ **do**
4: Compute $\hat{\tau}^{\text{DM}}(\mathcal{S})$ and its standard error, $\hat{s}^{\text{DM}}(\mathcal{S})$.
5: Compute subgroup estimates, $\hat{\tau}^{\text{UNCAL}}(\mathcal{S})$, by any method.

6: **end**
7: Optimize: $(\hat{\alpha}, \hat{\beta}) = \arg\max_{\alpha,\beta} \ell(\alpha, \beta)$
8: Compute the calibrated ITE: $\hat{\tau}^{\text{CAL}}(\mathbf{X}_i) = \hat{\alpha} + \hat{\beta}\hat{\tau}^{\text{UNCAL}}(\mathbf{X}_i), \quad \forall i \in \{1, \ldots, N\}$.
9: **return** $\alpha, \beta, \hat{\tau}^{\text{CAL}}$

Turning to our primary theoretical result, we first make the following two assumptions.

Assumption 1 (Honesty). *Calibration is performed on a set of data that is independent from the HTE model.*

Assumption 2 (Normality). *Each of the aggregated subgroup conditional average treatment effects ($\hat{\tau}^{\text{DM}}(\mathcal{S})$) is normally distributed.*

The first assumption can be ensured through sample splitting; the second holds asymptotically because of the central limit theorem.

Proposition 1. *Algorithm 1 provides the best linear predictor (BLP) of the subgroup effects.*

Proof [Proof Sketch]. Following Theorem 2.1 of Chernozhukov et al. (2023), honesty (Assumption 1) implies that $\hat{\tau}^{\text{UNCAL}}(\mathcal{S})$ can be taken to be an exogenous regressor. Normality of $\hat{\tau}^{\text{DM}}(\mathcal{S})$ implies all necessary regularity conditions for standard OLS results to apply. The implication of these two assumptions is that the calibrated effects are the best linear approximation (in the sense of likelihood) to the subgroup average treatment effect. □

The full proof is provided in Online Appendix F1. With stronger assumptions, a similar result holds that the calibrated model is the BLP of *individual-level* effects, the proof of this is provided in Online Appendix F2.

Our method can be viewed as a simplified and aggregated form of the second-stage regression proposed by Chernozhukov et al. (2023) through the use of grouped regression (Prais and Aitchison 1954). This proposition shows that estimates passed through Algorithm 1 have linear calibration. More precisely, they undergo the maximal likelihood linear transformation of the estimated subgroup effects. The HTE estimates benefit most from calibration when the precalibrated estimates are linearly correlated with (but not equal to) the true subgroup effects. However, if the HTE model is very poor (e.g., Figures 4(c) and 7(c)), our calibration approach may be of little help because no linear transformation of the estimated subgroup effects can replicate the actual subgroup effects. For example, a result like Figure 4(c) might be seen if the covariates used in the underlying model poorly predict the treatment effect or if there is excessive noise in the model-free estimates. Results like Figure 7(c) might be seen as if the HTE model overregularizes the covariates, leading to a lack of heterogeneity in the initial HTEs. Our calibration method operates directly on the scale of the treatment effects themselves, as in the second-stage models of X-learner (Künzel et al. 2019), but it provides a guarantee of the best linear prediction in the vein of Chernozhukov et al. (2023).

Our approach can be extended to the Bayesian framework by incorporating a prior over the estimated parameters. This prior reflects a prior belief that the HTE model of interest is well calibrated. The inclusion of this prior involves an addition to the log-likelihood defined in Equation (7):

$$\sum_{i=1}^{N} \log g(\hat{\tau}^{\text{CAL}}_{\alpha,\beta}(\mathbf{X}_i)). \tag{8}$$

where g is the probability density function of the normal distribution: $\mathcal{N}(\hat{\tau}^{\text{UNCAL}}(\mathbf{X}_i), \lambda)$ and λ is the standard deviation of the prior. It is clear that this prior is equivalent to a prior on the parameters (α, β) as being close to (0,1). It is essentially a ridge prior, due to the duality of ridge regression and the Normal-Normal conjugate Bayesian model. In our simulations, we adopt this regularized model, tuning λ (which determines the precision of our prior) using cross-validation. We describe this cross-validation procedure in Algorithm 2. Our cross-validation procedure partitions the data into V folds, wherein $V-1$ are used as training ($v=0$), and 1 is used as the holdout for validation ($v=1$). This process then is repeated so that each fold is held out once. The held-out log likelihood is averaged over all folds, and this result is used for model selection and for tuning both the number of bins and the amount of regularization (if used) in the calibration procedure. The computational cost of cross-validation in this setting is relatively low, as the calibration procedure does not require refitting of individual-level models.

Algorithm 2 (Out-of-Sample Validation)
1: **Input**: $\mathrm{y}, \mathrm{w}, K, \hat{\tau}^{\text{UNCAL}}, \mathrm{v}$,
2: Subset $\mathrm{y}, \mathrm{w}, \hat{\tau}^{\text{UNCAL}}$ to the training set where $\mathrm{v} = 0$.
3: $(\hat{\alpha}, \hat{\beta}, \hat{\tau}^{\text{CAL},v=0}) = \texttt{HTE Calibration}(\mathrm{y}^{v=0}, \mathrm{w}^{v=0}, K, \hat{\tau}^{\text{UNCAL},v=0})$
4: Compute the calibrated ITE in the holdout: $\hat{\tau}^{\text{CAL},v=1}(x_i) = \hat{\alpha} + \hat{\beta}\hat{\tau}^{\text{UNCAL},v=1}(\mathbf{X}_i), \quad \forall i \in \{j \in \{1, \ldots, N\} : v_j = 1\}$
5: **return** The log likelihood calculated in the holdout: $\ell^{v=1}(\hat{\alpha}, \hat{\beta})$

Our method uses a likelihood-based approach, with the uncalibrated HTE model nested within this model (corresponding to $\alpha = 0, \beta = 1$). Consequently, we apply a simple likelihood-ratio-based specification test to evaluate if the initial HTE model necessitates calibration. This test essentially checks for the presence of bias in the subgroup effects. This likelihood-ratio test statistic is given by $-2(\ell^* - \ell_0)$, where ℓ^* is the log likelihood under the optimized α and β, whereas ℓ_0 is the log likelihood under the linear restriction that $\alpha = 0$ and $\beta = 1$. This statistic is an asymptotically distributed χ^2 random variable, with two degrees of freedom (Casella and Berger 2002). However, given our parametric setup, this

specification test detects only *linear* miscalibration of the original HTE model.

This calibration method can also accommodate "stacking," a popular machine-learning technique for increasing generalization performance by combining multiple models (Wolpert 1992); the stacking can be done directly on the estimated treatment effects to improve the generalization performance. We provide more information about how it works in Online Appendix B.

4.2. Bias Under a T-learner with Ridge Regression

Regularization techniques are commonly used to mitigate overfitting in these models. However, these techniques can inadvertently introduce bias when used for HTE estimation. We will specifically explore a bias mechanism attributable to regularization and establish the bias within the T-learner framework. The occurrence of regularization-induced bias within HTE models has been recognized previously (Hahn et al. 2018). This issue arises because models are estimated on individual response surfaces, and the optimal regularization for each response surface may not correspond to what would be best for estimating the difference (i.e., the ITEs) between them. For instance, this could occur if the treatment effect function is smoother than the individual response functions. We derive our theoretical result using ridge regression as the base-learner.

Proposition 2. *Suppose $\boldsymbol{X}$ is an orthonormal basis, and the BLP of Y for the control and treatment potential outcome surface is β_0 and β_1, respectively. The BLP of the CATE function is $\beta_1 - \beta_0 \equiv \beta_\tau$. Suppose the CATE is estimated using a T-learner with ridge regression for base learners with regularization parameter λ, then*

- *Bias of ATE:* $\mathbb{E}[\tau(\boldsymbol{X}) - \hat{\tau}(\boldsymbol{X})] = -\frac{\lambda}{1+\lambda}\beta_\tau\ \mathbb{E}[\boldsymbol{X}]$
- *Bias of CATE*: $\mathbb{E}[\tau(\boldsymbol{X}) - \hat{\tau}(\boldsymbol{X})|\boldsymbol{X}] = -\frac{\lambda}{1+\lambda}\beta_\tau \boldsymbol{X}$
- *After calibration, the ATE is unbiased*, $\mathbb{E}[\tau(\boldsymbol{X}) - \hat{\tau}(\boldsymbol{X})] = 0$
- *After calibration, the CATE is unbiased*, $\mathbb{E}[\tau(\boldsymbol{X}) - \hat{\tau}(\boldsymbol{X})|\boldsymbol{X}] = 0$

The proof is deferred to Online Appendix F3. In particular, this result shows that we can express precisely what the calibration coefficient on a ridge regression T-learner will be in expectation: $1 + \lambda$. That is, a large amount of regularization directly translates into miscalibration in this simple setting, but can be removed with our calibration procedure.

5. Simulated and Real-World Experiments

In this section, we investigate the impact of HTE methods on bias and assess the effectiveness of calibration in mitigating this bias. We begin by running a series of simulations to demonstrate three types of bias and show how calibration can help reduce this bias. We then test the calibration method on both simulated and real-world RCTs. In all experiments in this section, we divide our data into training, validation, and test sets. The evaluation process is as follows:

1. We infer the HTEs using a machine learning model trained on the training data.
2. We use out-of-sample validation (Algorithm 2) to obtain the number of bins K and the amount of regularization λ on the validation set. We perform the proposed HTE calibration approach to learn α and β (Algorithm 1) using the tune K and λ on the validation set.
3. We evaluate the performance of the estimated ITEs on the test set.

We measure the performance using the mean absolute error (MAE) by computing the difference between the truth ITEs (known in simulations) and the estimated HTEs for each unit. Specifically, this quantity is defined as

$$\mathrm{MAE}(\hat{\tau}) = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{N}\sum_{i}^{N}|\hat{\tau}(\mathbf{X}_i) - \tau(\mathbf{X}_i)|\right), \tag{9}$$

where $\hat{\tau}(\mathbf{X}_i)$ is the estimated HTE for individual i, characterized by feature $\mathbf{X}_i$, and $\tau(\mathbf{X}_i)$ is the ground-truth ITE for individual i. The mean is taken over L simulations and we average the absolute errors in the HTEs for N individuals in the specific simulation.

5.1. Potential Mechanisms of Bias

5.1.1. Regularization-Induced Bias. We first assess how our method rectifies the bias introduced by regularization, as indicated in Proposition 2. We use a linear data generation process (DGP), adapted from Künzel et al. (2019), to illustrate this bias, with treatments assigned randomly, as in randomized experiments. The DGP is defined as follows:

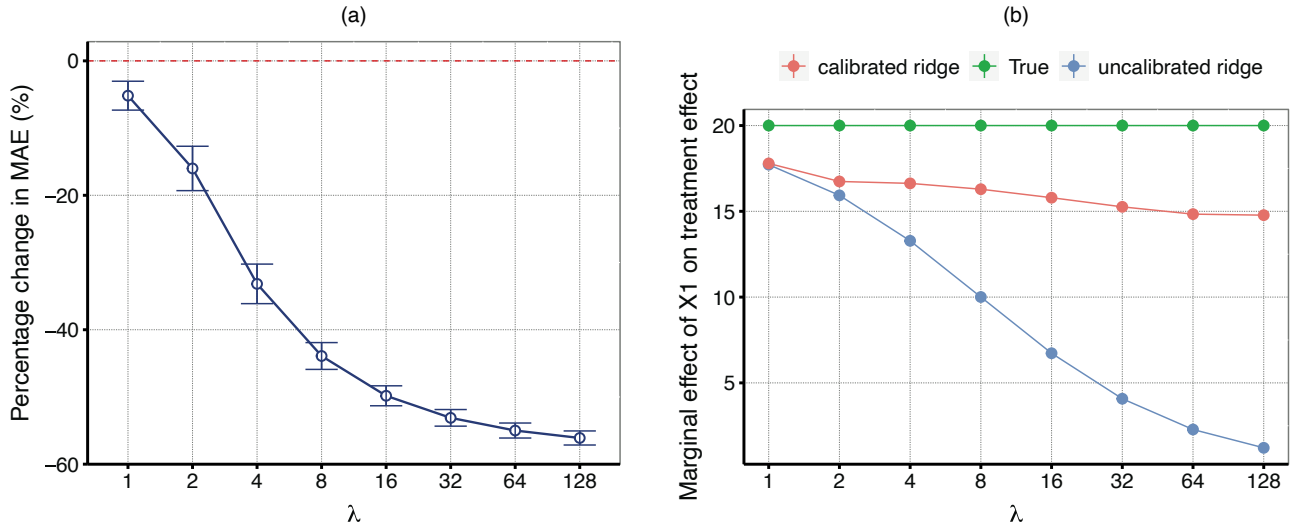
$$\begin{aligned} \mathrm{y}(0) &= \mathbf{X}\boldsymbol{\beta}_0 + \epsilon_0, \\ \mathrm{y}(1) &= \mathbf{X}\boldsymbol{\beta}_0 + 1 + 20\mathrm{X}_1 + \epsilon_1, \end{aligned} \tag{10}$$

where $\mathbf{X} \sim \mathrm{Uniform}([0,1]^{N\times D})$, $N = 3{,}000$, $D = 50$; treatment is randomly assigned $\mathrm{w} \sim \mathrm{Bernoulli}(0.5)$; X_1 is the first dimension of $\mathbf{X}$; ϵ_0 and ϵ_1 follow a normal distribution with parameters $\mathcal{N}(0, 0.1)$. Each coefficient in $\boldsymbol{\beta}_0$ follows uniform distribution, and $\boldsymbol{\beta}_0 \sim \mathrm{Uniform}([-5,5]^D)$.

Ridge regression is used as the base learner to demonstrate the regularization-induced bias, and serves as the input to a T-learner model.[8] To best estimate the heterogeneity in HTE, the covariate X_1 should be *undersmoothed* in the underlying response surface models relative to the covariate if the task was the standard supervised learning task on $\mathrm{y}(0)/\mathrm{y}(1)$. In short, HTE models falsely conflate performance on the response surfaces with performance on the *effect estimate*. Only the latter typically is of interest in causal inference.

As illustrated in Figure 2(a), as regularization (λ) increases, the benefits of our method increase, resulting in a greater reduction of the MAE. We explain this decline by examining the marginal effect of X_1 in

Figure 2. (Color online) Benefits of Calibration Grow with Increased Regularization



Notes. The x axis corresponds to the regularization parameter λ in ridge regression. In (a), the y axis shows the percentage change in the error of the HTE, which is computed as $\frac{\text{MAE}(\hat{\tau}^{\text{CAL}})-\text{MAE}(\hat{\tau}^{\text{UNCAL}})}{\text{MAE}(\hat{\tau}^{\text{UNCAL}})}\%$, where $\hat{\tau}^{\text{CAL}}$ and $\hat{\tau}^{\text{UNCAL}}$ are used as inputs to Equation (9). In (b), the y axis shows the marginal effect of X_1 on the treatment effect.

Figure 2(b), which shows that calibration ensures that the marginal effect on X_1 is not regularized away. Specifically, the bias in the coefficients of X_1 increases as regularization (λ) increases. In fact, we can characterize the bias in expectation in the slope of the heterogeneous effect on X_1 as $20\left(1-\frac{1}{1+\lambda}\right)$ (see Section 5.1.1 and Online Appendix F3 for details). This bias in the marginal effect is nonzero whenever regularization is applied to the response surfaces. Calibration on ridge regression effectively de-biases the marginal effects of the treatment effect by aligning the pink curve closer to the true effect. With very low regularization, ridge regression works well for estimating the marginal effect of the treatment effect with respect to X_1. However, as regularization increases, the slope is increasingly underestimated. Despite this, calibration ensures that the marginal effect is closer to the true effect than it would be under uncalibrated HTE methods, at all levels of regularization. Furthermore, the regularization of covariates orthogonal to the treatment effect will be left unchanged by calibration. As such, calibration accomplishes a similar aim as the one in Imai and Ratkovic (2013) by allowing different amounts of regularization to be applied to the underlying response surface and to the determinants of treatment efficacy.

5.1.2. Causal Forest in High-Dimensional Settings. The causal forest is a popular machine learning method to estimate HTE in causal inference (Wager and Athey 2018). It adapts the splitting criteria of random forests to focus on CATE estimation, partitioning bootstrap samples based on covariate values to explain HTEs. Wager and Athey (2018) proved unbiased inference for CATE when covariates both low-dimensional and uniformly distributed. (For more detailed discussions, please refer to Section 2.2). To delve deeper into the influence of calibration on the performance of causal forests in high-dimensional scenarios, we build on DGPs originally introduced by Wager and Athey (2018). We have adapted these DGPs to high-dimensional contexts to more thoroughly examine the efficacy of calibration in such settings. The first DGP, following equation (28) of Wager and Athey (2018), is as follows:

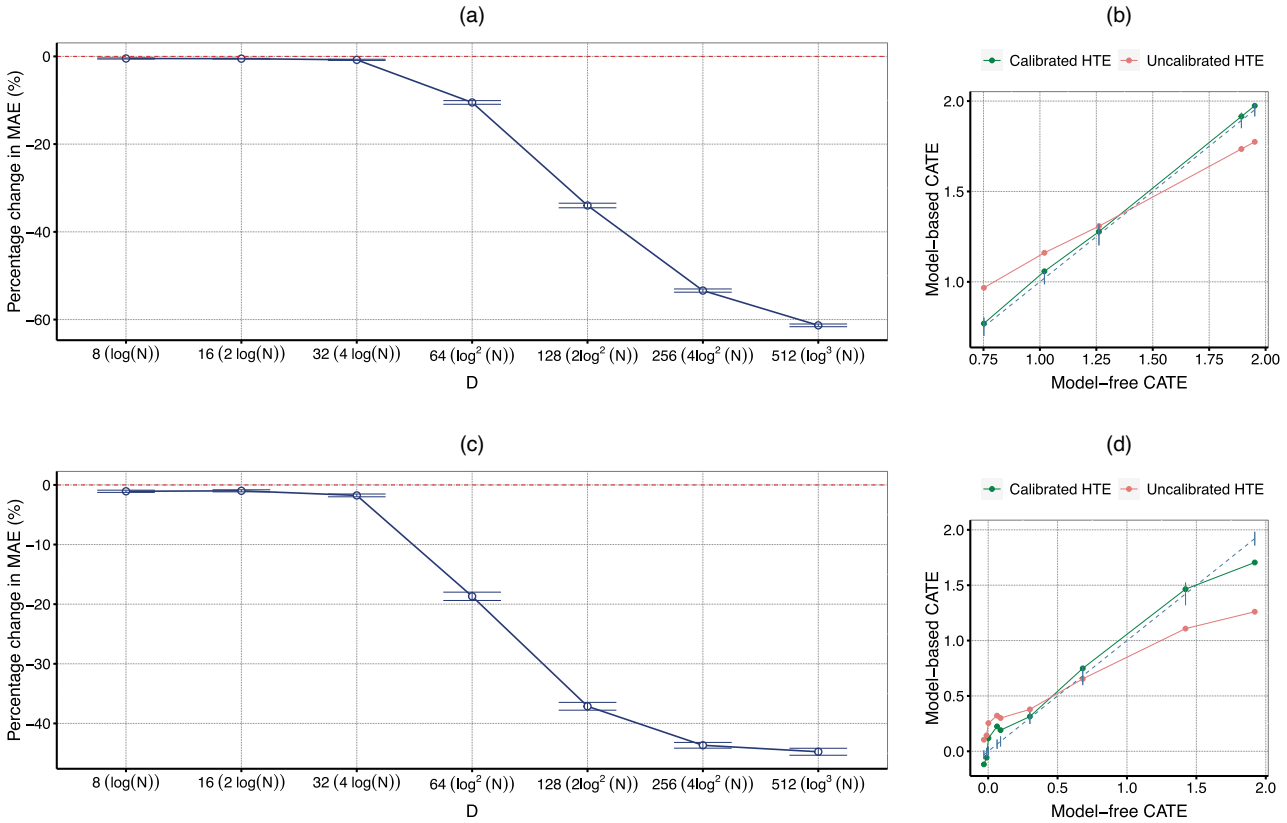
$$\begin{aligned} &y(0) \sim \mathcal{N}(0, 0.1), \\ &\tau(\mathbf{X}) = \frac{1}{2}\varsigma(X_1)\varsigma(X_2), \text{ where } \varsigma(x) = \frac{1}{1+e^{-20}(x-1/3)}, \\ &y(1) = y(0) + \tau(\mathbf{X}), \end{aligned} \tag{11}$$

where $\mathbf{X} \sim \text{Uniform}([0,1]^{N\times D})$, $N = 3{,}000$, and we vary the dimension D as multipliers of $\log(N)$; X_1 and X_2 are the first and second dimensions of $\mathbf{X}$; the treatments are randomly assigned with $w \sim \text{Bernoulli}(0.5)$.

In this second case, the ITE function, following equation (29) of Wager and Athey (2018), has a sharper spike when X_1 and X_2 approximate one. This DGP can demonstrate one known weakness of random forest–based methods. This method can fill in the valleys and flatten the peaks of the true ITE functions, especially near the edge of the feature space. The distribution of the true ITE is presented in Figure C1 in Online Appendix C. The DGP is similar to Equation (11), whereas the ITE function is

$$\tau(\mathbf{X}) = \frac{1}{2}\varsigma(X_1)\varsigma(X_2), \text{ where } \varsigma(x) = \frac{1}{1+e^{-12}(x-1/2)}. \tag{12}$$

Figure 3. (Color online) Benefits of Calibration on Causal Forest Increase as the Covariate Dimension (D) Increases



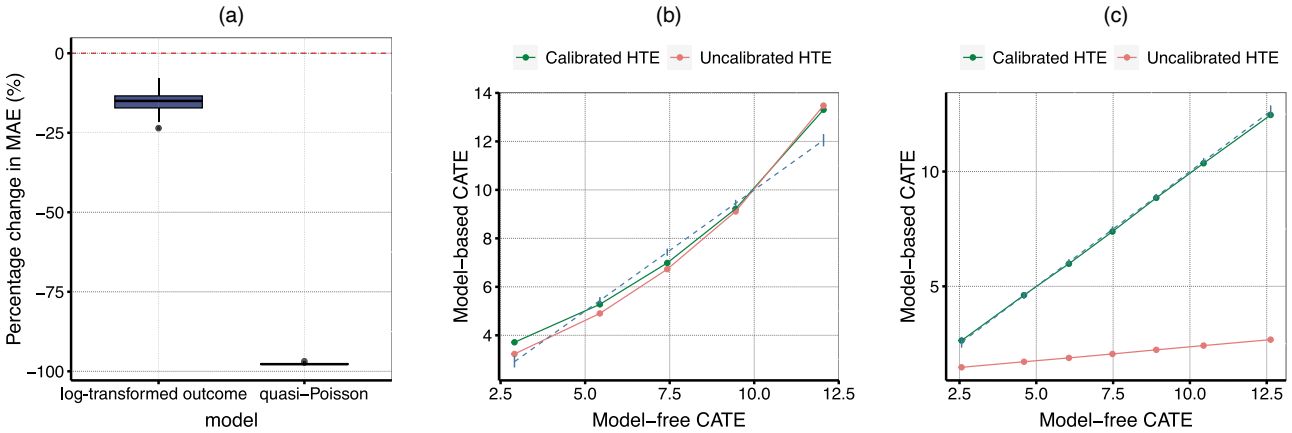
Notes. (a) Equation (11) and equation (28) in Wager and Athey (2018). (b) Q-Q plot (Equation (11)). (c) Equation (12) and equation (29) in Wager and Athey (2018). (d) Q-Q plot (Equation (12)).

Figure 3 shows the benefits of calibration in this setting. As shown in Figure 3, (a) and (c), as the dimensionality of the data increases, the performance of the calibration method improves, reducing the MAE significantly when dimensionality reaches $D = \log^2(N)$, and this benefit continues to grow with further increase in dimensionality. Furthermore, the diagnostic plots in Figure 3, (b) and (d), shed light on why the proposed calibration improves the performance of causal forest. In both DGPs, causal forests tend to overestimate smaller ITEs and underestimate higher ITEs, but the calibration can reduce these inaccuracies, improving the MAE on ITE significantly.

Specifically, in the DGP delineated by Equation (11), causal forests show a systematic tendency to overestimate smaller ITEs and underestimate larger ITEs (Figure 3(b)). This pattern is remedied by splitting the data into five bins and introducing calibration, yielding parameters of $\alpha = -0.67$ and $\beta = 1.49$. These parameters suggest that calibration repositions the uncalibrated curve (displayed in pink) downward by 0.67 units while simultaneously increasing the scale of the curve by 0.49. As a result of the calibration procedure, there is a significant decrease in the MAE of the ITE by 32.47%. In the DGP specified by Equation (12), we observe a similar pattern of causal forests overestimating smaller ITEs and underestimating larger ITEs (Figure 3(d)). Additionally, we observe that causal forests particularly struggles with the subgroup exhibiting the largest CATE. This observation aligns with the argument in Wager and Athey (2018) that causal forests encounter difficulties when handling data with spikes, especially at the boundaries of covariate regions. By applying the calibration procedure and partitioning the data into nine bins, we obtain calibration parameters of $\alpha = -0.28$ and $\beta = 1.58$. The calibration approach effectively shifts the initial estimated curve (in pink) downward by 0.28 units and amplifies its magnitude by 0.58. Consequently, the calibrated estimate curve (in green) displays a noticeable improvement in accuracy, reducing the MAE in ITE by 35.94%. Our results demonstrate that calibration can effectively improve the estimation of ITEs and CATEs, particularly in challenging scenarios where causal forests yield poor performance.

5.1.3. Misspecification-Induced Bias. Misspecification of response functions can introduce bias because the estimation of HTE relies on the predictions of the base learners. This misalignment is prevalent because capturing complex functional forms of the responses is challenging when the ground-truth response surface is

Figure 4. (Color online) Performance of Calibration in Misspecification-Induced Bias



Notes. In (b) and (c), the dashed blue line indicates "perfect" calibration (the bars denote the 95% confidence intervals of $\hat{\tau}^{\mathrm{DM}}$). The closer the estimated effects are to the blue line, the smaller the error in the estimated CATE. (a) Performance. (b) Regression with log-transformed outcome: Q-Q plot. (c) Quasi-Poisson: Q-Q plot.

unknown. In this section, we investigate a DGP where the ITE is a linear function of two-dimensional covariates:

$$\begin{aligned} y(0) &= 1 + \epsilon_0, \\ y(1) &= 1 + 10X_1 + 5X_2 + \epsilon_1, \end{aligned} \tag{13}$$

where $\mathbf{X} \sim \mathrm{Uniform}([0,1]^{N\times 2})$, $N = 3{,}000$, and X_1 and X_2 are the first and second dimensions of the covariates $\mathbf{X}$; ϵ_0 and ϵ_1 follow a normal distribution with parameters $\mathcal{N}(0, 0.1)$, and $w \sim \mathrm{Bernoulli}(0.5)$.

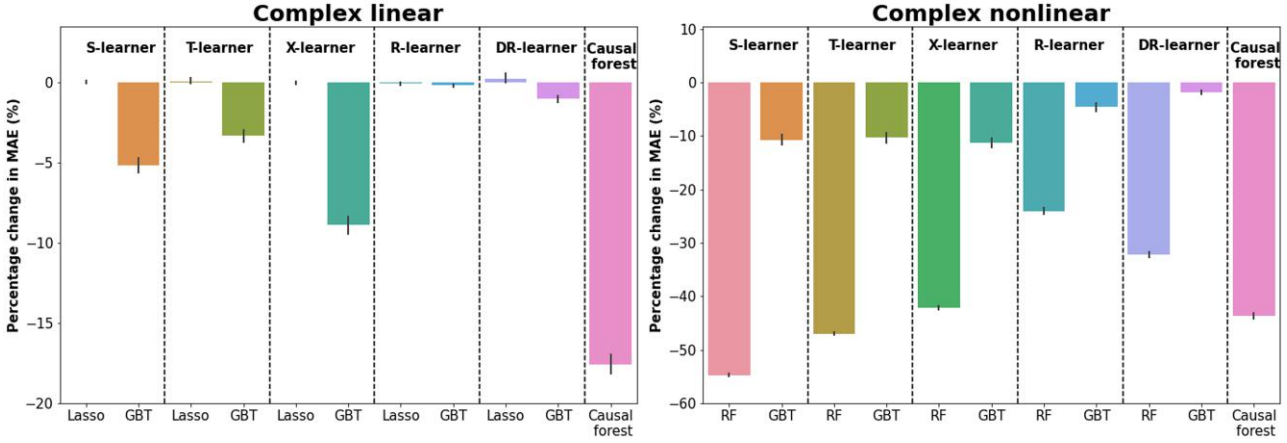
In this DGP, the functions for the potential outcomes are entirely linear and well behaved. However, an analyst may mistakenly model the conditional expectations of the outcome using a nonlinear function of the covariates. We explore two distinct types of such misspecifications in the base learner. The first type frequently occurs when analysts run a regression with a log-transformed outcome, inaccurately assuming that the outcome is an exponential function of the covariates. This model and analysis approach is commonly adopted in social science, especially when dealing with dispersed outcomes and to achieve interpretability of the coefficients (e.g., the coefficient directly signify the treatment effect measured as a semi-elasticity). The second type of misspecification involves the use of a quasi-Poisson generalized linear model (GLM) (Wooldridge 1999) to model the treatment and the outcome conditional expectation. The "quasi" feature of this model adapts Poisson regression to accommodate continuous outcomes or data exhibiting overdisperson. This model is frequently employed in fields of management and economics (Galasso and Simcoe 2011, Oettl 2012, Chatterji and Fabrizio 2014, Kuppuswamy and Bayus 2017, Kuusela et al. 2017). Although both models could be considered good modeling choices if the conditional expectations were correctly specified (Wooldridge 1999), their performance suffers significantly in the case of these misspecifications. These incorrect modeling assumptions result in substantial misspecification, leading to poor modeling of the potential outcome surfaces.

We present the results in Figure 4. The percentage decrease in MAE of ITEs, aggregated across 100 simulations, is demonstrated in Figure 4(a). Collectively, these figures attest to the considerable reduction in MAE achieved through our calibration process compared with uncalibrated model-based ITEs. The average decreases of 15.40% and 97.75% are observed for the two distinct misspecification errors, respectively. Further insights into the efficiency of our proposed calibration methodology can be gained from Figure 4, (b) and (c), which display the Q-Q diagnostic plot from a single simulation. In the case where an log-transformed outcome is used, the uncalibrated estimates succeed in bringing all but the subgroup with the smallest CATE closer to the model-free CATEs. In the case where quasi-Poisson is used, the uncalibrated estimates substantially underestimate the true effects, owing to the misspecification in the response functions. Fortunately, our calibration method corrects this issue, aligning the uncalibrated CATE (shown as the pink line) precisely along the diagonal reference line (representing the gold standard, model-free CATE). In general, the fact that incorrect modeling choices lead to inappropriate statistical tradeoffs should be unsurprising.

5.2. Performance Comparisons in Synthetic Settings

In this section, we evaluate the performance of calibration on eleven CATE estimators using two DGPs. Our analysis looks at the following HTE methods: S-learner (Hill 2011), T-learner (Künzel et al. 2019), X-learner (Künzel et al. 2019), R-learner (Nie and Wager 2021),

Figure 5. (Color online) Performance Evaluations of Calibration on Two DGPs Using S-Learner, T-Learner, X-Learner, R-Learner, DR-Learner, and Causal Forest



Notes. RF, random forest. The y axis shows the percentage change in the MAE, which is computed as $\frac{\text{MAE}(\hat{\tau}^{\text{CAL}})-\text{MAE}(\hat{\tau}^{\text{UNCAL}})}{\text{MAE}(\hat{\tau}^{\text{UNCAL}})}$%, where $\hat{\tau}^{\text{CAL}}$ and $\hat{\tau}^{\text{UNCAL}}$ are used as inputs to Equation (9). A negative value suggests that calibration reduces the MAE of the uncalibrated HTE estimates.

DR-learner (Kennedy 2023), and causal forest (Wager and Athey 2018). More detailed descriptions of these meta-algorithms and the causal forest are provided in Online Appendix A. We evaluate our method on two DGPs with different properties: complex linear and complex nonlinear. In both DGPs, treatments are randomly assigned ($w \sim \text{Bernoulli}(0.5)$). To capture the complexity of the CATE functions, we follow two DGPs used in Künzel et al. (2019), where the treatment effect is as complex as the response functions. These CATE functions do not satisfy regularity conditions such as sparsity or linearity, making them challenging to model accurately.

For both cases, we use $\mathbf{X} \in \mathbb{R}^{N \times D}$, where $N = 5{,}000$, $D = \overline{\log^2(N)}$ and $\overline{(\cdot)}$ is the ceiling operator. We use $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, where Σ is a random correlation matrix (following the vine method in Lewandowski et al. 2009). We follow the complex linear DGP used in Künzel et al. (2019):

$$\begin{aligned} y(0) &= \mathbf{X}^T \boldsymbol{\beta}_0 + \epsilon_0, \text{with } \boldsymbol{\beta}_0 \sim \text{Uniform}([-5,5]^D), \\ y(1) &= \mathbf{X}^T \boldsymbol{\beta}_1 + \epsilon_1, \text{with } \boldsymbol{\beta}_1 \sim \text{Uniform}([-5,5]^D), \end{aligned} \tag{14}$$

where ϵ_0 and ϵ_1 follow a normal distribution with parameters' $\mathcal{N}(0, 0.1)$. For each of the five meta-learners, we separately apply both linear lasso regression and nonlinear gradient-boosted trees (GBT) as the base learner.

We also consider the following complex nonlinear DGP outlined in Künzel et al. (2019):

$$\begin{aligned} y(0) &= -\frac{1}{2}\varsigma(X_1)\varsigma(X_2) + \epsilon_0, \\ y(1) &= \frac{1}{2}\varsigma(X_1)\varsigma(X_2) + \epsilon_1, \end{aligned} \tag{15}$$

where $\varsigma(x) = \frac{1}{1+e^{-12(x-1/2)}}$; X_1 and X_2 are the first and second dimensions of the high-dimensional covariates $\mathbf{X}$; ϵ_0 and ϵ_1 follows a normal distribution with parameters $\mathcal{N}(0, 0.1)$. To account for the nonlinearity of the true outcome functions, we use two nonlinear models, namely random forest and GBT, as the base learners in our meta-learners.

Figure 5 showcases performance comparisons between calibrated and uncalibrated HTE methods across 200 simulations of the aforementioned DGPs. We observe that no single HTE model consistently outperforms others in all scenarios. However, implementing the proposed calibration method often leads to improved performance of existing HTE methods. Considering the complex linear DGP case, S-learner, T-learner, and X-learner using GBT and causal forest exhibit enhanced performance with calibration. This improvement can be attributed to the regularization-induced bias and misspecification-induced bias inherent in the base learners. On the other hand, lasso regression, which aligns well with the true DGP (Equation (14)), performs accurately in capturing the response functions and individual treatment effects. As a result, calibration has minimal impact on the performance of lasso regression-based models, with $\alpha \approx 0$ and $\beta \approx 1$. In the scenario of the complex nonlinear DGP, calibration proves beneficial for all HTE methods. Comparatively, GBTs generally perform better than random forests, with calibration leading to stronger improvements in the latter. The performance improvements are particularly pronounced for the causal forest in both DGPs due to the high-dimensional setting.

Although our analysis demonstrates the effectiveness of our calibration approach, it also highlights that our methodology may not be required in every scenario. For instance, when the underlying DGP is linear and a correctly specified base learner, such as lasso regression, is used, the calibration property is already inherent in the

original HTE method. However, given that no method performs consistently well across all scenarios, our two-stage framework provides practitioners with diagnostic tools to uncover potential calibration issues within any HTE model. If a miscalibration issue is identified, our calibration method can be used to ameliorate such bias.

5.3. Real-World Randomized Experiments

Given that ground-truth ITEs in randomized experiments are unobserved, we use the subgroup CATE as a surrogate measure to assess accuracy of the model-based HTE estimates. We calculate the MAE in the subgroup effects as

$$\mathrm{MAE}_{\mathrm{CATE}}(\hat{\tau}(\mathcal{P})) = \frac{1}{|\mathcal{P}|}\sum_{\mathcal{S}\in\mathcal{P}} |\hat{\tau}(\mathcal{S}) - \hat{\tau}^{DM}(\mathcal{S})|. \quad (16)$$

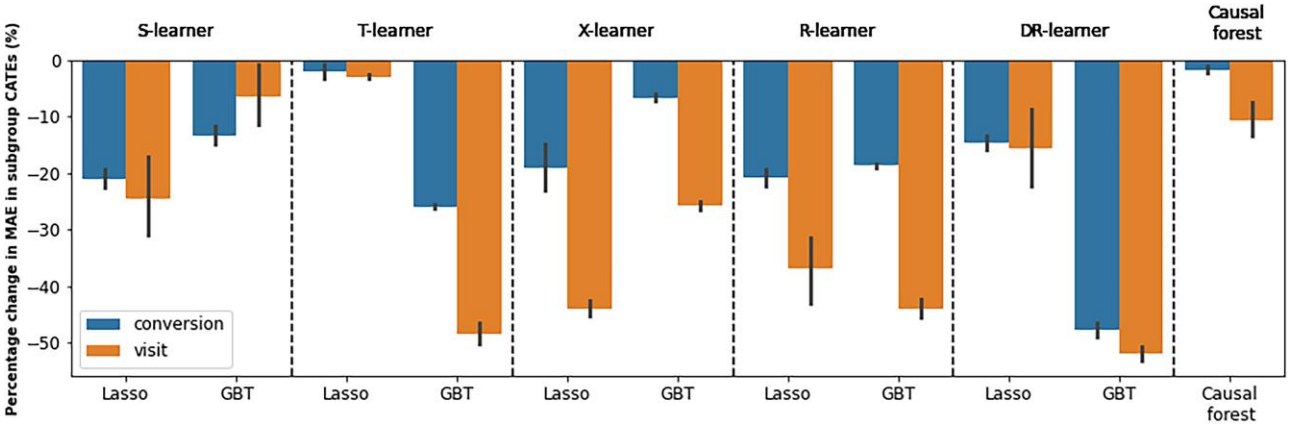
Recall that $\hat{\tau}^{DM}(\mathcal{S})$ is the model-free DM estimates on the CATEs for subgroup $\mathcal{S}$ (treated as the gold standard estimator in RCTs); $\hat{\tau}(\mathcal{S})$ is the model-based CATE, aggregated from the ITEs within subgroup $\mathcal{S}$.

We evaluate calibration on the publicly available advertising campaign data shared by Criteo AI Laboratory (Diemert et al. 2018).[9] This data set is constructed by assembling data from several randomized experiments involving advertising campaigns. It consists of 25M observations, each representing a user with 12 features, a treatment indicator, and two binary outcome labels (i.e., visited/converted or not). A positive label means the user visited or was converted on the advertiser's website during the two-week test period. The feature names have been anonymized for privacy reasons. We train different HTE methods on the training set, perform calibration on the validation set, and then evaluate the MAE on the subgroup CATEs on the test set, comparing it with the model-free estimator. Because of the high computation complexity for the GBTs in our evaluation, we randomly sampled 10% of the data; from this sample, 48% was used for training 11 HTE models, 32% was used for calibration, and 20% was used for testing.

In Figure 6, we showcase the percentage reduction in the MAE of the subgroup CATE (Equation (16)). As illustrated in Figure 6, most HTE methods show a reduction in MAE through calibration. We observe notable improvements in methods such as the T-learner, R-learner, and doubly robust (DR)-learner using GBT as the base learner, and the X-learner and R-learner using lasso regression as the base learner. To better understand the effect of calibration, we examine the Q-Q plot using visits as the outcome in Figure D1 in Online Appendix D and present selected representative Q-Q plots in Figure 7. Broadly speaking, calibration (green curve) brings many of the uncalibrated HTEs (pink curve) closer to the model-free CATEs (dashed black line). For instance, in Figure 7(a), calibration markedly aligns larger CATEs closer to the model-free CATEs using DM estimators by mitigating both underestimations and overestimations.

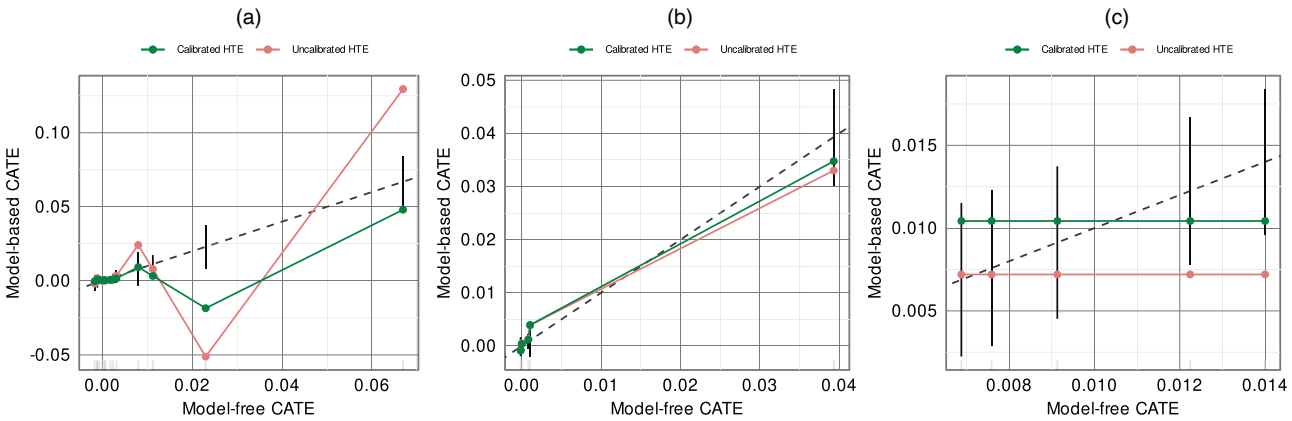
This pattern is also observed in the X-learner with lasso regression as the base learner, as well as the R-learner and DR-learner with GBT base learners (Figure D1 in Online Appendix D). Figure 7(b) demonstrates how calibration can reduce underestimations for the subgroup with the largest CATEs for causal forest, bringing them closer to the model-free CATEs. Similar trends can be seen in S-learner with both base learners (Figure D1). We also observe in Figure 6 that the T-learner with a lasso regression base learner predicts homogeneous ITEs for all individuals. Similarly, the R-learner and DR-learner with lasso regression base learners show no heterogeneity in the predicted ITEs, as depicted in Figure D1. This suggests that these methods do not capture any treatment heterogeneity based on the observed covariates due to the sparsity imposed by the base learner. Because these initial HTE methods lack informative signals, calibration offers minimal benefit beyond aligning the average treatment effect (ATE) with the model-free

Figure 6. (Color online) Performance Evaluation on the RCT by Criteo AI Laboratory



Notes. The test set is bootstrapped with replacements to obtain the confidence intervals. Negative values signify that calibration is able to bring the HTE estimates closer to the model-free CATEs, which is treated as the gold standard estimator for RCTs.

Figure 7. (Color online) Q-Q Plot Comparing Model-Free and Model-Based CATEs for the Visit Outcome on a Large-Scale Randomized Experiment Using Data from Criteo AI Laboratory



Notes. The dashed black line represents a perfect estimate. The black vertical bar is the confidence interval. (a) T-learner (GBT). (b) Causal forest. (c) R-learner (Lasso).

estimator. This analysis, based on a real-world RCT, further emphasizes the broad applicability of our proposed calibration method, showcasing its effectiveness across a range of HTE methods. The impact of calibration on reducing HTE estimate errors or leaving them unchanged depends on various factors, including the underlying CATE function, response surfaces, and the quality of HTEs derived from the machine learning models.

5.4. Discussion: When to Calibrate and When Not to Calibrate

The previous sections provide examples of circumstances where calibration may be useful. In particular, there are a few circumstances that are worth highlighting. Clearly, when regularization is used in a model, as discussed in Section 5.1.1, it can lead to miscalibration, even when the true model is linear and easily expressible by the function class of the regression models. However, this is not necessarily an argument against regularization. In situations with high-dimensional covariates, strong regularization may be necessary to obtain a reasonable model of individual response surfaces. However, this regularization can affect the calibration of treatment effects. When the underlying model does not do a good job of fitting the data, as in Sections 5.1.2 or 5.1.3, it can also result in miscalibration. In summary, when a model very effectively approximates relevant features of the data, then calibration will (typically) be achieved. However, if the model fails to capture relevant features of the data, the calibration property of the method will tend to suffer. When calibration suffers and miscalibration occurs, recalibrating the resulting model with Algorithm 1 can be helpful in improving the overall effectiveness of the HTE model. Conversely, there are cases where the HTE model performs well without requiring the calibration approach. For instance, if treatment effects are constant, many HTE methods may (trivially) provide fairly calibrated estimates of HTEs. Similarly, if treatment effects are approximately linear, a T-learner using linear base learners can perform effectively without the need for calibration.

How should an analyst know whether to ensure calibration by using Algorithm 1? The easiest approach is to simply start with the diagnostic Q-Q plot we suggest. This diagnostic provides an effective way to determine whether there might be a serious problem with an estimated HTE model. Although difference-in-means (DM) estimates of subgroup effects are noisy, they are *unbiased*. If clear and systematic discrepancies emerge between the CATEs derived from DM estimates for subgroups and those obtained from the HTE model, it is reasonable to suspect that the culprit is the HTE model. The examples in the previous section illustrate how this Q-Q plot effectively identifies a poorly fitting HTE model. Figure D3 in Online Appendix D, for example, clearly shows that, under the uncalibrated model, large effects were being systematically underestimated, whereas small effects were being systematically overestimated. In this case, it is clear that a linear rescaling would be very effective in improving the model. In other cases, this might not be true. In Figure 7(c), for example, the model-based HTE estimates are constant across subgroups. Linear calibration would have minimal effect on improving such a model, except for aligning it with the ATE. Instead, the solution is to consider alternative base models (perhaps the true model is nonlinear) or to find more informative covariates to enrich the HTE model's predictive power. Calibration is not the end of the story for improving HTE models, but it is a useful lens through which to examine and diagnose a model. It is recommended that practitioners consistently review the Q-Q diagnostic plot before relying on the outcomes of HTE models.

6. Practical Value of Calibration in Policy Design

This section demonstrates the practical value of the proposed calibration procedure in enhancing policy makers' utility through individual interventions in three representative policy-making scenarios. These applications require navigating and reconciling heterogeneity within and between experiments to achieve optimal policy designs.

6.1. Prescriptive Framework for Budget-Constrained Optimal Policy Deployment

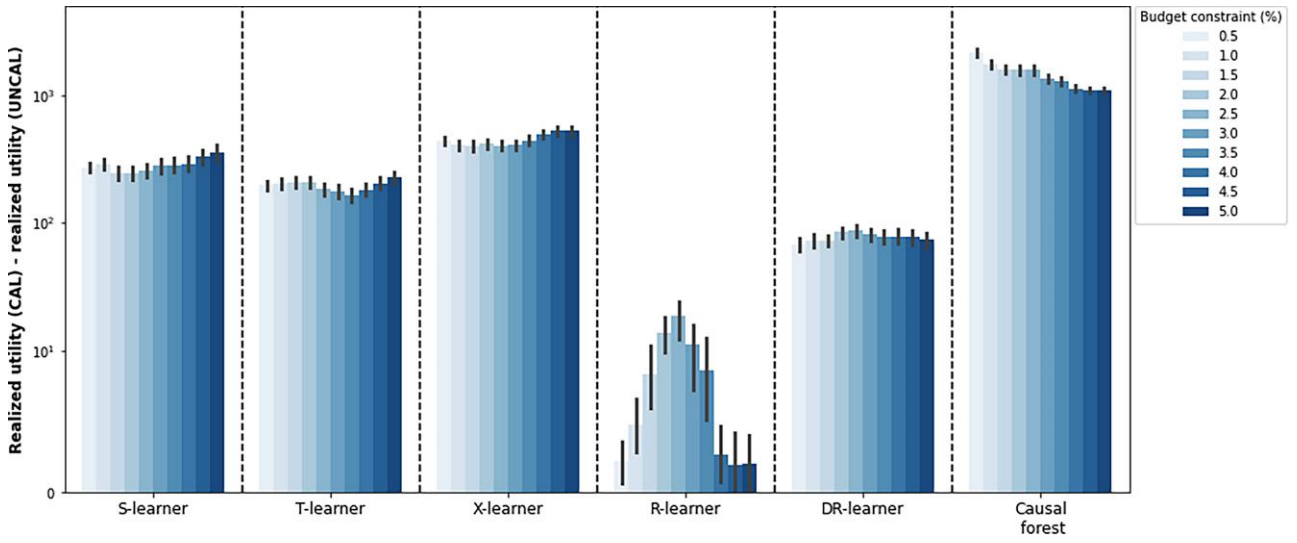
We first investigate the application of calibration in a general budget-constrained environment where multiple policy levers can optimize an organizational objective. This situation involves ex ante unknown individual costs and benefits. This broad framework, originally proposed by McFowland et al. (2021), finds applications in several public or private scenarios where decision makers aim to maximize utility while adhering to budget constraints. It encompasses two notable applications related to referral marketing and public policy, depicted in Example 1. This prescriptive analytics framework involves three steps: conducting multiple RCTs, using machine learning for HTE estimation, and applying constrained optimization. Our calibration method reduces bias in HTE estimates and enables optimization within budget constraints.

The framework operates on a sample $\mathcal{N}_1$, containing independently and identically distributed (i.i.d.) units from the population of interest. For each individual i, we observe a D-dimensional vector of covariates $\mathbf{X}_i \in \mathcal{X}$. The decision maker has a collection of J treatments and must determine the assignment of each binary indicator $W_{ij} \in \{0,1\}$, where $W_{ij}=1$ assigns unit i to treatment j. A sample $\mathcal{N}_2$ of i.i.d. units from the population of interest $\mathcal{P}$ is observed. The sample observed differs from the samples that are used for estimating HTE ($\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$). For each individual i, there exist unknown benefits ($B_i(j)$), costs ($C_i(j)$), and utilities (the difference between the benefit and cost: $U_i(j) = B_i(j) - C_i(j)$) that would be realized if individual i is assigned to treatment j. The optimization can be formulated as follows, assuming a decision maker's budget of M[10]:

$$\begin{aligned} \text{maximize}: \quad & \text{Expected utility} = \sum_{i=1}^{N}\sum_{j=1}^{J} U_i(j) W_{ij}, \\ \text{subject to}: \quad & \sum_{i=1}^{N}\sum_{j=1}^{J} C_i(j) W_{ij} \le M, \\ & \sum_{j=1}^{J} W_{ij} \le 1, \quad \forall i. \text{ and } \quad W_{ij} \in \{0,1\}, \quad \forall i,j. \end{aligned} \tag{17}$$

We evaluate our method using the complex linear DGP (Equation (14) in Section 5.2) to simulate individual cost and benefit from the RCT, generated independently. The models are trained on 50% of the population, with calibration performed on 25% of the population. Using the remaining 25% hold-out test set, we optimize policy assignment based on the estimated HTEs (Equation (17)). The realized utility is then computed using the ground-truth ITEs per the predicted policy. Figure 8[11] shows the performance comparisons based on the realized utility, following the procedure of McFowland et al.

Figure 8. (Color online) Comparison of the Realized Utility from the Prescriptive Framework in a Budget-Constrained Optimization Using Calibrated HTEs vs. Uncalibrated HTEs



Notes. A positive value suggests that calibration can improve the realized utility for the policy makers. For all meta-learners, GBT is the base leaner.

(2021). The budget constraints are set by ranking individuals based on computed utilities and assigning total costs corresponding to the top 0.5%–5% (with 0.5% increments) selected individuals as the budget. Calibration results in noticeable utility improvements across all meta-algorithms and causal forest, particularly for those using GBTs as the base learner. Notably, the causal forest sees the most significant utility enhancement, whereas the R-learner and DR-learner models sees the least. This correlation between the reduction in MAE of ITEs and the increase in policy value underscores calibration's effectiveness and practical significance in budget-constrained optimization. Although the performance of meta-learners using lasso regressions remains unchanged by calibration, our results demonstrate that calibration enhances the realized utility across all other HTE models considered.[12]

6.2. Multi-KPI Optimization with Budget Constraint

Organizations often focus on multiple key performance indicators (KPIs) simultaneously (Diemert et al. 2018). Practical decision-making problems in these domains involve tradeoffs (Deng and Shi 2016, Letham et al. 2019). For instance, evaluating the effectiveness of an advertisement on Facebook or an email marketing campaign requires considering metrics such as reach, fan growth, click-through rates, and conversion rates. These KPIs may not always exhibit strong correlations (Morgan and Rego 2006). To address this, decision makers aim to develop optimal intervention strategies that integrate multiple KPIs into an overall evaluation criterion (OEC) framework (Kohavi et al. 2007), where the relative importance of each KPI is predefined by the policy-maker.

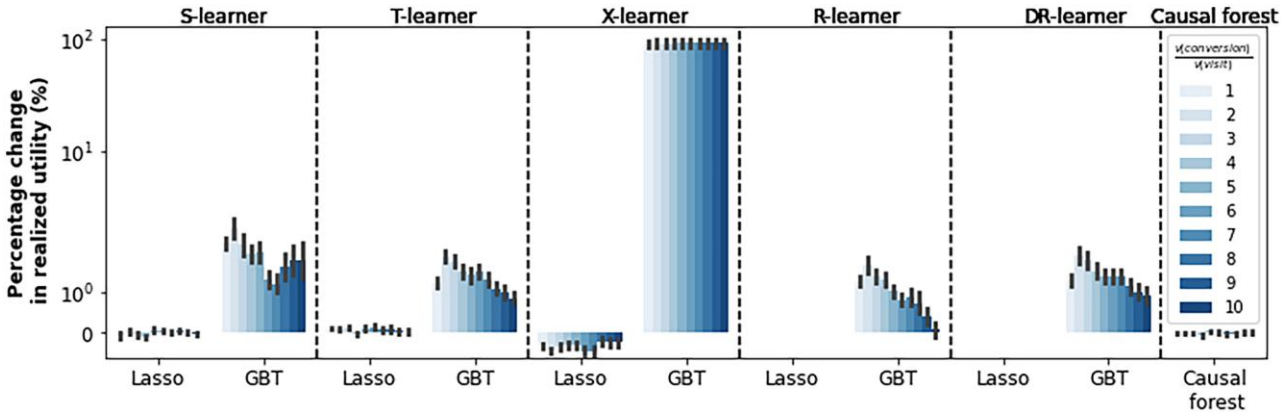
We adapt the framework proposed by McFowland et al. (2021) to this specific application. We estimate the treatment effects on population $\mathcal{N}_1$ and implement the optimal policy on a different population $\mathcal{N}_2$, where $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$. Assuming the decision maker cares about a collection of Q KPIs. The platform's utility from individual i's receiving treatment can be seen as an aggregation of all the KPIs: $U_i = \sum_{q=1}^{Q} \nu_q b_i^q w_i$, where $w_i \in \{0,1\}$ denotes the treatment assignment; ν_q is the importance of KPI indexed by q, determined by the platform. Accurate estimation of the ITEs for each KPI, denoted as $\{b_i^q\}_{q=1}^{Q}$, becomes crucial in calculating the overall utility U_i. Additionally, the optimization problem is subject to a constraint C that limits the number of treated individuals. We formulate the optimization problem as follows:

$$\begin{aligned} &\text{maximize}: \quad \text{Expected utility} = \sum_{i=1}^{N}\sum_{q=1}^{Q} \nu_q b_i^q w_i \\ &\text{subject to}: \quad \sum_{i=1}^{N} w_i \le C, \text{ where } w_i \in \{0,1\}, \quad \forall i. \end{aligned} \tag{18}$$

In this scenario, we use the real-world data set provided by Criteo AI Laboratory (refer to Section 5.3). We explore different levels of relative importance for the two KPIs, with values ranging from $1 \le \frac{\nu_{\text{conversion}}}{\nu_{\text{visit}}} \le 10$. We ensure that the conversion is never less important than visit, as conversion is the more lucrative ad campaign metric. These values are plugged into Equation (18) to optimize estimated HTEs on the test set, with C ranging from 1% to 5% of the test set population over 150 iterations. We then use rejection sampling and off-policy evaluation (OPE)[13] to calculate the realized utility (Dudík et al. 2014), providing an unbiased estimate of the policy value (Equation (18)).

Figure 9 shows calibration's impact on utility improvement across different $\frac{\nu_{\text{conversion}}}{\nu_{\text{visit}}}$ ratios for models

Figure 9. (Color online) Comparison of Realized Utility from Multi-KPI Optimization Using Calibrated HTEs vs. Uncalibrated HTEs with Respect to Ratio $\frac{\nu_{\text{conversion}}}{\nu_{\text{visit}}}$



Notes. The darker blue color of the bars indicates higher importance of conversion in the decision-making process. The y axis represents the improvement in realized utility resulting from calibration, with positive values indicating an enhancement in utility for policy makers.

like T-learner, X-learner, R-learner, and DR-learner using GBT base learners, with X-learner showing the strongest improvement. Yet, subpar performing initial HTE models or those generating homogeneous ITEs, as seen with R-learner and DR-learner models using lasso regression, may not benefit significantly from our calibration method. In examining the X-learner, we found significant variations in utility change based on the base learner used. Although GBT as a base learner led to a significant policy value rise (89.4%), lasso regression led to a small dip (less than 0.3%) compared with the uncalibrated model. Even though calibration improved the estimation of CATEs through X-learner with lasso regression base learners, it did a poor job with a subgroup of high CATE units crucial for optimization, which resulted in slightly decreased utility.

In our analysis, we considered 10 different $\frac{v_{\text{conversion}}}{v_{\text{visit}}}$ values. We observed a decrease in percentage utility improvement as the relative importance of visits declined for T-learner, R-learner, and DR-learner, aligning multi-KPI optimization closer to single-KPI optimization. As focus shifts toward a single KPI, ranking within one KPI gains importance over the accuracy of ITE magnitudes, adding to the academic dialogue of Fernández-Loría and Provost (2022b) that precise HTEs may not be needed in all policy-making situations.

6.3. Multitreatment Uplift Modeling

Our final application pertains to multitreatment uplift modeling (MTUM) (Olaya et al. 2020). The objective of MTUM is to identify the most effective treatment among multiple options in order to achieve the most favorable outcomes. It estimates the conditional probabilities of a positive outcome for each individual under each treatment and then identifies the treatment with the highest ITEs. MTUM is useful in scenarios where decision makers need to choose from multiple treatment options to maximize performance. For example, this could involve a marketer selecting the best personalized promotional message or a doctor choosing among several alternative treatments for a patient. In line with the MTUM framework presented in Olaya et al. (2020), we consider a set of T mutually exclusive treatments, covariates $\mathbf{X}$, and a binary outcome variable $Y \in \{0, 1\}$. The objective is to identify the treatment with the greatest effect. The optimal treatment for individual i is the treatment with the highest uplift. Uplift is computed by estimating the differences in positive response probabilities between each treatment and control group for each individual. We use the DGP in the `causalml` Python package to generate a data set of $N = 3{,}000$ samples with $D = 100$ features, out of which 10 features are informative, whereas the remaining 90 are redundant. We use four metrics: Qini metric and expected response each at 10% and 100% of the population respectively, as outlined in Olaya et al. (2020).

- **Qini Metric:** We calculate the Qini metric by constructing the uplift curve based on the ranking of individuals and measuring the cumulative difference in the response rate to a certain percentage of the population in the test set, relative to the control group (Rzepakowski and Jaroszewicz 2012). The Qini metric quantifies the area between the uplift curve and the curve of a random model. It is adapted for multiple treatments by considering the maximal uplift across all potential treatments (Olaya et al. 2020). A larger Qini index indicates a more significant incremental effect from the predicted optimal treatment.
- **Expected Response:** Designed specifically for MTUM (Zhao et al. 2017b), this metric calculates the expected response $\overline{z}$ using the observed treatment in the test set, the predicted optimal treatment by the uplift model, the prior probability of treatment k as $p_{T=k}$, and the observe outcome Y. Given an uplift model h, the individual expected response is defined as $z_i = \sum_{k=1}^{K} \frac{Y_i}{p_{T=k}} \mathbb{I}\{h(x_i) = k\} \mathbb{I}\{T = k\}$, where $\mathbb{I}(\cdot)$ is the 0/1 indicator function. The expected response is computed as $\overline{z} = \frac{1}{N'} \sum_{i=1}^{N'} z^i$, where N' is chosen as 10% and 100% of the population.

We consider all MTUM approaches outlined in Olaya et al. (2020).[14] We provide a brief overview of these approaches later and additional details are in Online Appendix E. (1) *Dummy and interactions approach* (DIA) constructs a single predictive model using an extended input generated from added treatment dummy variables and interaction between the dummy and pretreatment variables. Random forest and linear regression models with feature selection are used (Olaya et al. 2020). (2) *Separate model approach* (SMA) computes the HTEs for each treatment and uses these estimates to compute the uplift of each treatment. Treatment allocation is based on uplift scores and variability. We use two variations, including random forest and linear regression models with feature selection, following Olaya et al. (2020). (3) *Causal K-nearest neighbor* (CKNN) identifies the optimal treatment for an individual by finding the K most similar individuals (Guelman et al. 2015). (4) *Contextual treatment selection* (CTS) alters the splitting criterion in the decision tree to directly maximize the expected response—an unbiased metric in MTUM (Zhao et al. 2017b). We implement both uplift tree and uplift forest variations of CTS. (5) *Naive uplift approach* (NUA) creates separate binary uplift models to estimate the uplift between treatment and control groups (Olaya et al. 2020). We explore two variants of NUA methods: (i) uplift random forest, which compares the probability distributions of treatment groups using measures such as KL divergence, chi-square, or squared Euclidean distance; and (ii) uplift causal conditional inference forest. This leads to four variations. (6) *Multitreatment modified outcome approach* (MMOA) is

essentially a multiclass classification (Olaya et al. 2020). We use random forests and multinomial regression as the machine learning models for MMOA.

The data are evenly divided into training, validation, and test sets. All methods are trained using the training set. Feature selection and model selection are performed on the validation set for all methods. Calibration is applied to the validation set for methods using uncalibrated estimated ITEs. For the uncalibrated and calibrated HTE models (as similar to all MTUM benchmarks), top-performing models are selected based on validation set performance. The evaluation metrics are subsequently compared on the hold-out test set. Figure 10 demonstrates the varying performance of these methods in the MTUM application. Our results indicate that the calibrated HTE method performs the best in terms of the expected response at both 10% and 100% of the population and in the Qini index at 10%. Meanwhile, the NUA method performs the best in the Qini index at 100%, followed by the calibrated HTE and the DIA approaches. It is worth noting that expected response is an improved metric over Qini index for MTUM (Zhao et al. 2017b). This analysis demonstrates our method's potential to outperform the methods specifically developed for MTUM applications.

In conclusion, we conduct a thorough analysis on three policy applications, showcasing the practical benefits of calibration when dealing with multiple treatment effects. We highlight its usefulness in reconciling heterogeneity within and between experiments. To validate the versatility and reliability of our method, we use both continuous and binary outcomes, and we test it in a controlled synthetic environment and in real-world RCTs. Importantly, we show that calibration performs competitvely when compared with application-specific machine learning techniques developed for MTUM applications. However, it should be noted that we do not claim calibrated effects-based decisions are universally superior to application-specific benchmarks. Rather, we believe that an interesting area for future research is investigating the distinct circumstances under which using calibrated HTEs in downstream applications outperforms or faces challenges when compared with CDM (Fernández-Loría and Provost 2022b).

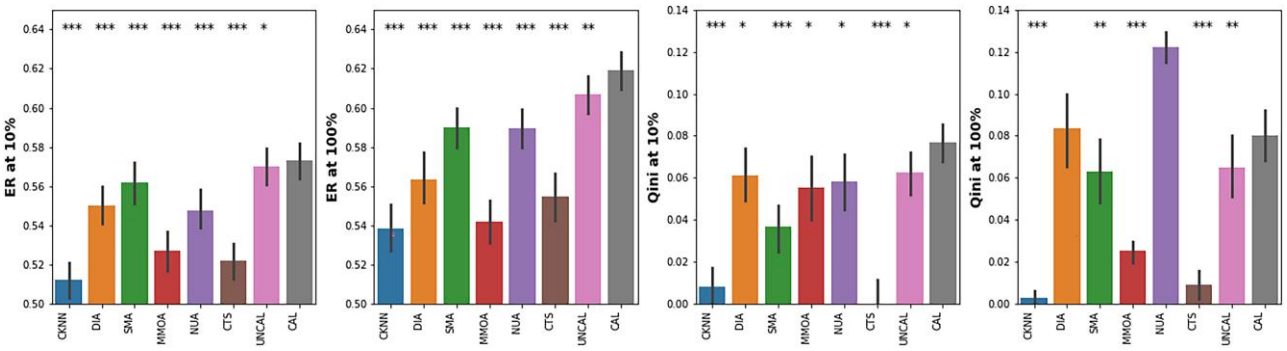
7. Discussion and Conclusion

Understanding HTEs provides a foundation for gaining scientific insights and designing optimal policies in fields like IS, public policy, economics, and healthcare. Machine learning methods, often used for this purpose, promise potential advancements. However, despite their proficiency with supervised learning, machine learning has been less effective in estimating HTEs. This limitation stems from a mismatch between the task of interest (i.e., predicting ITEs) and the task for which machine learning methods excel—response surface modeling.

This paper shows that many existing HTE models do not necessarily yield calibrated causal effects in synthetic scenarios and in two real-world RCTs conducted by Facebook and Criteo AI Laboratory. This result suggests a potential threat that machine learning-based HTE estimates pose both to interpreting experiment findings and to downstream applications. Undoubtedly, policy designs based on these biased estimates are not ideal. We proposed a Q-Q diagnostic plot to help assess whether the CATEs are calibrated. We recommend that practitioners and researchers use this plot to determine whether the HTE estimates are biased before optimizing or deploying any individualized policy in practice.

To rectify the miscalibration issue of HTE models, we develop a simple, scalable, and effective method for calibrating estimates of causal effects to noisy ground-truth benchmarks. Our algorithm is able to provide the

Figure 10. (Color online) Performance Evaluations Using MTUM Methods



Notes. The larger the value, the better the performance. We conduct a two-tailed pairwise t test between a calibrated HTE model and each benchmark method. The alternative hypothesis of this test is that the mean of a metric computed from the calibrated HTE model is different from that of a benchmark. We denote the significance level of this test using asterisks *, with $\{^*, ^{**}, ^{***}\}$ corresponding to $p = \{0.05, 0.01, 0.001\}$, respectively.

BLP of the model-free subgroup effects. Our method is model-agnostic, easy to implement, and can substantially improve the credibility of effect estimates from *any* HTE models. Owing to the simplicity and efficiency of our maximum likelihood estimation process, our method can be easily implemented in production environments. We deploy extensive simulations to elaborate on the underlying mechanisms of biases induced by the combination of response function estimates for ITE estimates. Beyond simulations, we illustrate how our method helps analyze heterogeneity in a large-scale experiment run on Facebook and in advertisement campaigns conducted by Criteo AI Laboratory. We demonstrate the effectiveness and broad applicability of our method in three applications, in relation to various HTE methods.

Despite the promise our study holds, it also faces certain limitations, prompting potential directions for future research. First, using our method hinges on preserving the rank-order of the machine learning-based ITE methods. Fortunately, the diagnostic Q-Q plot can reveal the poor performance of the HTE model in this regard. However, significant differences in the rank order of predicted and ground-truth binned subgroup effects would require researchers and practitioners to resort to alternative models or collect additional pretreatment covariates. In other words, our method requires the initial model to be "good enough" to signal the underlying effect heterogeneity. If the initial model fails to signal underlying effect heterogeneity, our method will not, in general, recover signals from the noise. One potential approach could involve exploring whether stacking multiple underperforming HTE estimators can deliver calibrated causal effects. Second, our calibration method is agnostic to the underlying HTE model. However, evaluating and comparing the performance of different HTE methods across various DGPs and CATE functions falls outside the scope of our current paper. It presents an intriguing avenue for future research to investigate and characterize the performance of different HTE methods under varying DGPs and CATE functions. Third, our calibration method is applicable only to RCTs. An important avenue for future research is to explore the calibration of HTEs in observational studies when unobserved confounders exist. These areas represent promising avenues for future investigation.

Acknowledgments

The authors thank the senior editor, the associate editor, and the anonymous reviewers for their insightful comments.

Endnotes

[1] Model-agnostic here means that our approach can be applied to *any* HTE model.

[2] See https://obamawhitehouse.archives.gov/precision-medicine.

[3] The size of individual leaves refers to a leaf's diameter, which is the length of the longest segment inside the leaf.

[4] Preference over the methods can depend on various factors such as convenience (Fernández-Loría and Provost 2022b), transparency and fairness (McFowland 2022), or reanalysis of existing RCTs (Eckles 2022). CDM is ideal for convenience, whereas CEE suits the other preferences. With rapidly evolving literature on both CDM and CEE, a definitive comparison is beyond our paper's scope. We refer readers to ongoing discussions on these methods' merits for data-driven decision-making (Eckles 2022, Fernández-Loría and Provost 2022c, McFowland 2022, Shalit 2022).

[5] We overload the notation of $\hat{\tau}(\cdot)$. Its meaning depends on the type of input it receives. When $\hat{\tau}(\cdot)$ is a function of a set, it defines an estimate of a subgroup effect. However, when $\hat{\tau}(\cdot)$ is a function of a vector, it defines an estimator of the conditional treatment effect at a specific value of covariates.

[6] A T-learner involves training separate models for the treatment and control groups and calculating the treatment effect as the difference between their predictions (Künzel et al. 2019).

[7] The Q-Q diagnostic plot helps to discern potential biases within HTE estimates based on our subgrouping method (Algorithm 1). In practice, practitioners can diagnose their chosen subgrouping with our Q-Q plot. It is important, however, to underscore that an unproblematic Q-Q plot does not guarantee the absence of miscalibration in alternative subgroupings.

[8] We adopt the T-learner to elucidate bias mechanisms for several reasons: First, because of its simplicity and interpretability, the T-learner allows a more transparent display of the mechanism. Secondly, recent meta-algorithms enhance the T-learner procedure by incorporating nuisance functions related to treatment assignments. Methods such as the X-learner (Künzel et al. 2019) and DR-learners (Kennedy 2023) use plug-in estimators for the treatment effect, which are based on a T-learner. Many methods employ plugin estimation of the response functions; therefore, the bias mechanisms we discuss for the T-learners are also relevant for these meta-learners and tree-based methods. Lastly, T-learners are frequently integrated into the experimentation infrastructure of numerous industry firms due to their scalability, interpretability, and convenience (Markov et al. 2021).

[9] We retrieved the unbiased version of the Criteo data set via https://ailab.criteo.com/criteo-uplift-prediction-dataset/.

[10] We can formalize Facebook's decision in Example 3 on whether to implement the policy for each user as a special case of this optimization problem. There are two simplifications. First, in this particular application, the cost of implementing the policy for each user on Facebook can be considered negligible, denoted as $C_i = 0$ for all users i. Second, Facebook is only evaluating one treatment option. As a result, the objective function can be adapted as $\sum_{i=1}^{N} B_i w_i$ where B_i is the treatment effect estimated from RCTs for i. The constraint remains to be the same.

[11] In most of our figures, we used relative terms to report the results, with this figure as an exception. The reason for reporting utility in absolute term is that the utility for the uncalibrated model may be negative, which renders it insensible to compute a relative change.

[12] We did not compare our results with uplift models and contextual bandit approaches as McFowland et al. (2021) demonstrated their prescriptive framework using causal forest outperforms these benchmarks.

[13] OPE involves excluding units for which the observed treatment does not align with the proposed treatment under a given policy, and then calculating the policy value based on the remaining units. Thus, OPE provides an unbiased estimate of the policy value in Equation (18).

[14] The repository can be accessed through https://github.com/vub-dl/MTUM.

References

Aronow P, Robins JM, Saarinen T, Sävje F, Sekhon J (2021) Nonparametric identification is not enough, but randomized controlled trials are. Preprint, submitted September 27, https://arxiv.org/abs/2108.11342.

Athey S, Imbens GW (2015) Machine learning methods for estimating heterogeneous causal effects. *Statistics* 1050(5):1–26.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* 113(27):7353–7360.

Breiman L (2001) Random forests. *Machine Learn.* 45(1):5–32.

Casella G, Berger RL (2002) *Statistical Inference*, vol. 2 (Duxbury, Pacific Grove, CA).

Chatterji AK, Fabrizio KR (2014) Using users: When does external knowledge enhance corporate product innovation? *Strategic Management J.* 35(10):1427–1445.

Chernozhukov V, Fernández-Val I, Luo Y (2018) The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica* 86(6):1911–1938.

Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I (2023) Generic machine learning inference on heterogenous treatment effects in randomized experiments. NBER Working Paper No. 24678, National Bureau of Economic Research, Cambridge, MA.

Deng A, Shi X (2016) Data-driven metric development for online controlled experiments: Seven lessons learned. *Proc. 22nd ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining* (ACM, New York), 77–86.

Diemert E, Betlei A, Renaudin C, Amini MR (2018) A large scale benchmark for uplift modeling. *Proc. AdKDD & TargetAd (ADKDD'18)* (ACM, New York), 1–6.

Dudík M, Erhan D, Langford J, Li L (2014) Doubly robust policy evaluation and optimization. *Statist. Sci.* 29(4):485–511.

Dwivedi R, Tan YS, Park B, Wei M, Horgan K, Madigan D, Yu B (2020) Stable discovery of interpretable subgroups via calibration in causal studies. *Internat. Statist. Rev.* 88:S135–S178.

Eckles D (2022) Commentary on "Causal decision making and causal effect estimation are not the same… and why it matters": On loss functions and bias–variance tradeoffs in causal estimation and decisions. *INFORMS J. Data Sci.* 1(1):17–18.

Fernández-Loría C, Provost F (2022a) Causal classification: Treatment effect estimation vs. outcome prediction. *J. Machine Learn. Res.* 23(59):1–35.

Fernández-Loría C, Provost F (2022b) Causal decision making and causal effect estimation are not the same… and why it matters. *INFORMS J. Data Sci.* 1(1):4–16.

Fernández-Loría C, Provost F (2022c) Rejoinder to "causal decision making and causal effect estimation are not the same… and why it matters". *INFORMS J. Data Sci.* 1(1):23–26.

Fernández-Loría C, Provost F, Anderton J, Carterette B, Chandar P (2023) A comparison of methods for treatment assignment with an application to playlist generation. *Inform. Systems Res.* 34(2):786–803.

Galasso A, Simcoe TS (2011) CEO overconfidence and innovation. *Management Sci.* 57(8):1469–1484.

Green DP, Kern HL (2012) Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quart.* 76(3):491–511.

Greenfeld D, Shalit U (2020) Robust learning with the Hilbert-Schmidt independence criterion. *Proc. Internat. Conf. on Machine Learn.* (PMLR, New York), 3759–3768.

Grimmer J, Messing S, Westwood SJ (2017) Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis* 25(4):413–434.

Guelman L, Guillén M, Pérez-Marín AM (2015) A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems* 72:24–32.

Hahn PR, Carvalho CM, Puelz D, He J, et al (2018) Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* 13(1):163–182.

Hill JL (2011) Bayesian nonparametric modeling for causal inference. *J. Comput. Graphical Statist.* 20(1):217–240.

Holland PW (1986) Statistics and causal inference. *J. Amer. Statist. Assoc.* 81(396):945–960.

Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Statist.* 7(1):443–470.

Imai K, Strauss A (2011) Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Anal.* 19(1):1–19.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, UK).

Jacob D (2020) Cross-fitting and averaging for machine learning estimation of heterogeneous treatment effects. Preprint, submitted July 6, https://arxiv.org/abs/2007.02852.

Jung J, Bapna R, Golden JM, Sun T (2020) Words matter! Toward a prosocial call-to-action for online referral: Evidence from two field experiments. *Inform. Systems Res.* 31(1):16–36.

Kennedy EH (2023) Toward optimal doubly robust estimation of heterogeneous causal effects. *Electronic J. Statis.* 17(2):3008–3049.

Kohavi R, Henne RM, Sommerfield D (2007) Practical guide to controlled experiments on the web: Listen to your customers not to the hippo. *Proc. 13th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining* (ACM, New York), 959–967.

Kuleshov V, Fenner N, Ermon S (2018) Accurate uncertainties for deep learning using calibrated regression. *Proc. Internat. Conf. on Machine Learn.* (ACM, New York), 2801–2809.

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* 116(10):4156–4165.

Kuppuswamy V, Bayus BL (2017) Does my contribution to your crowdfunding project matter? *J. Bus. Venturing* 32(1):72–89.

Kuusela P, Keil T, Maula M (2017) Driven by aspirations, but in what direction? Performance shortfalls, slack resources, and resource-consuming vs. resource-freeing organizational change. *Strategic Management J.* 38(5):1101–1120.

Letham B, Karrer B, Ottoni G, Bakshy E (2019) Constrained bayesian optimization with noisy experiments. *Bayesian Anal.* 14(2):495–519.

Lewandowski D, Kurowicka D, Joe H (2009) Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* 100(9):1989–2001.

Lin W (2013) Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Ann. Appl. Statist.* 7(1):295–318.

Markov IL, Wang H, Kasturi N, Singh S, Yuen SW, Garrard M, Tran S, et al. (2021) Looper: An end-to-end ml platform for product decisions. Preprint, submitted October 14, https://arxiv.org/abs/2110.07554.

McFowland E III (2022) Commentary on "Causal decision making and causal effect estimation are not the same… and why it matters". *INFORMS J. Data Sci.* 1(1):21–22.

McFowland E, Gangarapu S, Bapna R, Sun T (2021) A prescriptive analytics framework for optimal policy deployment using heterogeneous treatment effects. *MIS Quart.* 45(4):1807–1832.

Morgan NA, Rego LL (2006) The value of different customer satisfaction and loyalty metrics in predicting business performance. *Marketing Sci.* 25(5):426–439.

Nie X, Wager S (2021) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108(2):299–319.

Oettl A (2012) Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Sci.* 58(6):1122–1140.

Olaya D, Coussement K, Verbeke W (2020) A survey and benchmarking study of multitreatment uplift modeling. *Data Mining Knowledge Discovery* 34(2):273–308.

Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classifiers* 10(3):61–74.

Prais SJ, Aitchison J (1954) The grouping of observations in regression analysis. *Rev. Inst. Internat. Statist.* 22(1/3):1–22.

Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, Rich S, et al. (2020) Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2(7):369–375.

Rzepakowski P, Jaroszewicz S (2012) Decision trees for uplift modeling with single and multiple treatments. *Knowledge Inform. Systems* 32(2):303–327.

Schuler A, Baiocchi M, Tibshirani R, Shah N (2018) A comparison of methods for model selection when estimating individual treatment effects. Preprint, submitted June 13, https://arxiv.org/abs/1804.05146.

Shalit U (2022) Commentary on "Causal decision making and causal effect estimation are not the same… and why it matters". *INFORMS J. Data Sci.* 1(1):19–20.

Sun T, Gao G, Jin GZ (2019) Mobile messaging for offline group formation in prosocial activities: A large field experiment. *Management Sci.* 65(6):2717–2736.

Van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. *Statist. Appl. Genetic Molecular Biology*, vol. 6 (De Gruyter, Berlin), 1–23.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113(523):1228–1242.

Wolpert DH (1992) Stacked generalization. *Neural Networks* 5(2): 241–259.

Wooldridge JM (1999) Distribution-free estimation of some nonlinear panel data models. *J. Econometrics* 90(1):77–97.

Wu H, Tan S, Li W, Garrard M, Obeng A, Dimmery D, Singh S, et al. (2022) Interpretable personalized experimentation. *Proc. 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining* (ACM, New York), 4173–4183.

Xie Y (2007) Otis dudley duncan's legacy: The demographic approach to quantitative reasoning in social science. *Res. Soc. Stratification Mobility* 25(2):141–156.

Xie Y, Brand JE, Jann B (2012) Estimating heterogeneous treatment effects with observational data. *Sociol. Methodology* 42(1): 314–347.

Zhang M, Tsiatis AA, Davidian M (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* 64(3):707–715.

Zhao Q, Small DS, Ertefaie A (2017a) Selective inference for effect modification via the lasso. Preprint, submitted May 22, https://arxiv.org/abs/1705.08020.

Zhao Y, Fang X, Simchi-Levi D (2017b) Uplift modeling with multiple treatments and general response types. *Proc. SIAM Internat. Conf. on Data Mining* (SIAM, Philadelphia), 588–596.