HHS Public Access

Author manuscript

Proc Mach Learn Res. Author manuscript; available in PMC 2023 August 11.

Published in final edited form as:

Proc Mach Learn Res. 2023 July ; 202: 34831–34854.

Causal isotonic calibration for heterogeneous treatment effects

Lars van der Laan[*,1], **Ernesto Ulloa-Pérez**[*,2], **Marco Carone**[3,1], **Alex Luedtke**[1,3]

[1]Department of Statistics, University of Washington, USA

[2]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, USA

[3]Department of Biostatistics, University of Washington, USA

Abstract

We propose causal isotonic calibration, a novel nonparametric method for calibrating predictors of heterogeneous treatment effects. In addition, we introduce a novel data-efficient variant of calibration that avoids the need for hold-out calibration sets, which we refer to as cross-calibration. Causal isotonic cross-calibration takes cross-fitted predictors and outputs a single calibrated predictor obtained using all available data. We establish under weak conditions that causal isotonic calibration and cross-calibration both achieve fast doubly-robust calibration rates so long as either the propensity score or outcome regression is estimated well in an appropriate sense. The proposed causal isotonic calibrator can be wrapped around any black-box learning algorithm to provide strong distribution-free calibration guarantees while preserving predictive performance.

1. Introduction

Estimation of causal effects via both randomized experiments and observational studies is critical to understanding the effects of interventions and informing policy. Moreover, it is often the case that understanding treatment effect heterogeneity can provide more insights than overall population effects (Obermeyer & Emanuel, 2016; Athey, 2017). For instance, a study of treatment effect heterogeneity can help elucidate the mechanism of an intervention, design policies targeted to subpopulations who can most benefit (Imbens & Wooldridge, 2009), and predict the effect of interventions in populations other than the ones in which they were developed. These necessities have arisen in a wide range of fields, such as marketing (Devriendt et al., 2018), the social sciences (Imbens & Wooldridge, 2009), and the health sciences (Kent et al., 2018). For example, in the health sciences, heterogeneous treatment effects (HTEs) are of high importance to understanding and quantifying how certain exposures or interventions affect the health of various subpopulations (Dahabreh et al., 2016; Lee et al., 2020). Potential applications include prioritizing treatment to certain sub-populations when treatment resources are scarce, or individualizing treatment assignments when the treatment can have no effect (or even be harmful) in certain subpopulations (Dahabreh et al., 2016). As an example, treatment assignment based on risk scores has been used to provide clinical guidance in cardiovascular disease prevention

Corresponding author is Lars van der Laan (lvdlaan@uw.edu).

*These authors contributed equally to this work.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(Lloyd-Jones et al., 2019) and to improve decision-making in oncology (Collins & Varmus, 2015; Cucchiara et al., 2018).

A wide range of statistical methods are available for assessing HTEs, with recent examples including Wager & Athey (2018), Carnegie et al. (2019), Lee et al. (2020), Yadlowsky et al. (2021), and Nie & Wager (2021), among others. In particular, many methods, including Imbens & Wooldridge (2009) and Dominici et al. (2020), scrutinize HTEs via conditional average treatment effects (CATEs). The CATE is the difference in the conditional mean of the counterfactual outcome corresponding to treatment versus control given covariates, which can be defined at a group or individual level. When interest lies in predicting treatment effect, the CATE can be viewed as the oracle predictor of the individual treatment effect (ITE) that can feasibly be learned from data. Optimal treatment rules have been derived based on the sign of the CATE estimator (Murphy, 2003; Robins, 2004), with more recent works incorporating the use of flexible CATE estimators (Luedtke & van der Laan, 2016). Thus, due to its wide applicability and scientific relevance, CATE estimation has been of great interest in statistics and data science.

Regardless of its quality as a proxy for the true CATE, it is generally accepted that predictions from a given treatment effect predictor can still be useful for decision-making. However, theoretical guarantees for rational decision-making using a given treatment effect predictor typically hinge on the predictor being a good approximation of the true CATE. Accurate CATE estimation can be challenging because the nuisance parameters involved can be non-smooth, high-dimensional, or otherwise difficult to model correctly. Additionally, a CATE estimator obtained from samples of one population, regardless of its quality, may not generalize well to different target populations (Frangakis, 2009). Usually, CATE estimators (often referred to as learners) build upon estimators of the conditional mean outcome given covariates and treatment level (i.e., outcome regression), the probability of treatment given covariates (i.e., propensity score), or both. For instance, plug-in estimators such as those studied in Künzel et al. (2019) — so-called T-learners — are obtained by taking the difference between estimators of the outcome regression obtained separately for each treatment level. T-learners can suffer in performance because they rely on estimation of nuisance parameters that are at least as non-smooth or high-dimensional as the CATE, and are prone to the misspecification of involved outcome regression models; these issues can result in slow convergence or inconsistency of the CATE estimator. Doubly-robust and Neyman-orthogonal CATE estimation strategies like the DR-learner and R-learner (Wager & Athey, 2018; Foster & Syrgkanis, 2019; Nie & Wager, 2021; Kennedy, 2020) mitigate some of these issues by allowing for comparatively fast CATE estimation rates even when nuisance parameters are estimated at slow rates. However, while less sensitive to the learning complexity of the nuisance parameters, their predictive accuracy in finite-samples still relies on potentially strong smoothness assumptions on the CATE. Even when the CATE is estimated consistently, predictions based on statistical learning methods often produce biased predictions that overestimate or underestimate the true CATE in the extremes of the predicted values (van Klaveren et al., 2019; Dwivedi et al., 2020). For example, the ‘pooled cohort equations’ (Goff et al., 2014) risk model used to predict cardiovascular disease has been found to underestimate risk in patients with lower socioeconomic status or chronic inflammatory diseases (Lloyd-Jones et al., 2019). The implications of biased treatment effect

predictors are profound when used to guide treatment decisions and can range from harmful use to withholding of treatment (van Calster et al., 2019).

Due to the consequence of treatment decision-making, it is essential to guarantee, under minimal assumptions, that treatment effect predictions are representative in magnitude and sign of the actual effects, even when the predictor is a poor approximation of the CATE. In prediction settings, the aim of bestowing these properties on a given predictor is commonly called *calibration*. A calibrated treatment effect predictor has the property that the *average* treatment effect among individuals with identical predictions is close to their shared prediction value. Such a predictor is more robust against over-or-under estimation of the CATE in extremes of predicted values. It also has the property that the best predictor of the ITE given the predictor is the predictor itself, which facilitates transparent treatment decision-making. In particular, the optimal treatment rule (Murphy, 2003) given only information provided by the predictor is the one that assigns the treatment predicted to be most beneficial. Consequently, the rule implied by a perfectly calibrated predictor is at least as favorable as the best possible static treatment rule that ignores HTEs. While complementing one another, the aims of calibration and prediction are fundamentally different. For instance, a constant treatment effect predictor can be well-calibrated even though it is a poor predictor of treatment effect heterogeneity (Gupta et al., 2020). In view of this, calibration methods are typically designed to be wrapped around a given black-box prediction pipeline to provide strong calibration guarantees while preserving predictive performance, thereby mitigating several prediction challenges mentioned previously.

In the machine learning literature, calibration has been widely used to enhance prediction models for classification and regression (Bella et al., 2010). However, due to the comparatively little research on calibration of treatment effect predictors, such benefits have not been realized to the same extent in the context of heterogeneous treatment effect prediction. Several works have contributed to addressing this gap in the literature. Brooks et al. (2012) propose a targeted (or debiased) machine learning framework (van der Laan & Rose, 2011) for within-bins calibration that could be applied to the CATE setting. Zhang et al. (2016) and Josey et al. (2022) consider calibration of marginal treatment effect estimates for new populations but do not consider CATEs. Dwivedi et al. (2020) consider estimating calibration error of CATE predictors for subgroup discovery using randomized experimental data. Chernozhukov et al. (2018) and Leng & Dimmery (2021) propose CATE methods for linear calibration, a weaker form of calibration, in randomized experiments. For causal forests, Athey & Wager (2019) evaluate model calibration using a doubly-robust estimator of the ATE among observations above or below the median predicted CATE. Lei & Candès (2021) propose conformal inference methods for constructing calibrated prediction intervals for the ITE from a given predictor but do not consider calibration of the predictor itself. Xu & Yadlowsky (2022) propose a nonparametric doubly-robust estimator of the calibration error of a given treatment effect predictor, which could be used to detect uncalibrated predictors. Our work builds upon the above works by providing a nonparametric doubly-robust method for calibrating treatment effect predictors in general settings.

This paper is organized as follows. In Section 2, we introduce our notation and formally define calibration. There we also provide an overview of traditional calibration methods. In

Section 3, we outline our proposed approach, and we describe its theoretical properties in Section 4. In Section 5, we examine the performance of our method in simulation studies. We conclude with a discussion of our proposed approach in Section 6.

2. Statistical setup

2.1. Notation and definitions

Suppose we observe n independent and identically distributed realizations of data unit $O := (W, A, Y)$ drawn from a distribution P, where $W \in \mathscr{W} \subset \mathbb{R}^d$ is a vector of baseline covariates, $A \in \{0, 1\}$ is a binary indicator of treatment, and $Y \in \mathscr{Y} \subset \mathbb{R}$ is an outcome. For instance, W can include a patient's demographic characteristics and medical history, A can indicate whether an individual is treated (1) or not (0), and Y could be a binary indicator of a successful clinical outcome. We denote by $\mathscr{D}_n := \{O_1, O_2, \ldots, O_n\}$ the observed dataset, with $O_i := (W_i, A_i, Y_i)$ representing the observation on the ith study unit.

For covariate value $w \in \mathscr{W}$ and treatment level $a \in \{0, 1\}$, we denote by $\pi_0(w) := P(A = 1 \mid W = w)$ the propensity score and by $\mu_0(a, w) := E(Y \mid A = a, W = w)$ the outcome regression. The individual treatment effect is $Y_1 - Y_0$, where Y_a represents the potential outcome obtained by setting $A = a$. Without loss of generality, we assume that higher values of $Y_1 - Y_0$ are desirable. We also assume that the contrast $\tau_0(w) := \mu_0(1, w) - \mu_0(0, w)$ equals the true CATE, $E(Y_1 - Y_0 \mid W = w)$, which holds under certain causal assumptions (Rubin, 1974). Throughout, we denote by $\| \cdot \|$ the $L^2(P)$ norm, that is, $\|f\|^2 = \int [f(w)]^2 dP_W(w)$ for any given P_W-square integrable function $f : \mathscr{W} \in \mathbb{R}$, where P_W is the marginal distribution of W implied by P. We deliberately take as convention that the median median $\{x_1, x_2, \ldots, x_k\}$ of a set $\{x_1, x_2, \ldots, x_k\}$ equals the $\lfloor k / 2 \rfloor^{th}$ order statistic of this set, where $\lfloor k / 2 \rfloor := \max\{z \in \mathbb{N} : z \leq k / 2\}$.

Let $\tau : \mathscr{W} \in \mathbb{R}$ be a treatment effect predictor, that is, a function that maps a realization w of W to a treatment effect prediction $\tau(w)$. In practice, τ can be obtained using any black-box algorithm. Below, we first consider τ to be fixed, though we later address situations in which τ is learned from the data used for subsequent calibration. We define the calibration function $\gamma_0(\tau, w) := E[Y_1 - Y_0 \mid \tau(W) = \tau(w)]$ as the conditional mean of the individual treatment effect given treatment effect score value $\tau(w)$. By the tower property, $\gamma_0(\tau, w) := E[\tau_0(W) \mid \tau(W) = \tau(w)]$, and so, expectations only involving $\gamma_0(\tau, W)$ and other functions of W can be taken with respect to P_W.

The solution to an isotonic regression problem is typically nonunique. Throughout this text, we follow Groeneboom & Lopuhaa (1993) in taking the unique càdlàg piece-wise constant solution of the isotonic regression problem that can only take jumps at observed values of the predictor.

2.2. Measuring calibration and the calibration-distortion decomposition

Various definitions of risk predictor calibration have been proposed in the literature — see Gupta & Ramdas (2021) and Gupta et al. (2020) for a review. Here, we outline our definition

of calibration and its rationale. Given a treatment effect predictor τ, the best predictor of the individual treatment effect in terms of MSE is $w \mapsto \gamma_0(\tau, w) := E[Y_1 - Y_0 \mid \tau(W) = \tau(w)]$. By the law of total expectation, this predictor has the property that, for any interval $[a, b)$,

$$E\{[\tau_0(W) - \gamma_0(\tau, W)]I(\gamma_0(\tau, W) \in [a, b))\} = 0\,. \tag{1}$$

Equation 1 indicates that $\gamma_0(\tau, \cdot)$ is perfectly calibrated on $[a, b)$. Therefore, when a given predictor τ is such that $\tau(W) = \gamma_0(\tau, W)$ with P-probability one, τ is said to be perfectly calibrated (Gupta et al., 2020) for the CATE — for brevity, we omit "for the CATE" hereafter when the type of calibration being referred to is clear from context.

In general, perfect calibration cannot realistically be achieved in finite samples. A more modest goal is for the predictor τ to be approximately calibrated in that $\tau(w)$ is close to $\gamma_0(\tau, w)$ across all covariate values $w \in \mathscr{W}$. This naturally suggests the calibration measure:

$$CAL(\tau) := \int [\gamma_0(\tau, w) - \tau(w)]^2 dP_w(w)\,. \tag{2}$$

This measure, referred to as the ℓ_2-expected calibration error, arises both in prediction (Gupta et al., 2020) and in assessment of treatment effect heterogeneity (Xu & Yadlowsky, 2022). We note that $\mathrm{CAL}(\tau)$ is zero if $\tau(w)$ is perfectly calibrated. Additionally, averaging in $\mathrm{CAL}(\tau)$ with respect to measures other than P_W could be more relevant in certain applications; such cases can occur, for instance, when there is a change of population that results in covariate shift and we are interested in measuring how well τ is calibrated in the new population.

Interestingly, the above calibration measure plays a role in a decomposition of the mean squared error (MSE) between the treatment predictor and the true CATE, in that

$$\mathrm{MSE}(\tau) := \|\tau_0 - \tau\|^2 = \mathrm{CAL}(\tau) + \mathrm{DIS}(\tau), \tag{3}$$

with $\mathrm{DIS}(\tau) := E\{var[\tau_0(W) \mid \tau(W)]\}$ a quantity we term the distortion of τ. We refer to the above as a *calibration-distortion* decomposition of the MSE. A consequence of the calibration-distortion decomposition is that MSE-consistent CATE estimators are also calibrated asymptotically. However, particularly in settings where the covariates are high-dimensional or the CATE is nonsmooth, the calibration error rate for such predictors can be arbitrarily slow — this is discussed further after Theorem 4.6.

To interpret $\mathrm{DIS}(\tau)$, we find it helpful to envision a scenario in which a distorted message is passed between two persons. The goal is for Person 2 to discern the value of $\tau_0(w)$, where the value of $w \in \mathscr{W}$ is only known to Person 1. Person 1 transmits w, which is then distorted through a function τ and received by Person 2. Person 2 knows the functions τ and τ_0, and may use this information to try to discern $\tau_0(w)$. If τ is one-to-one, $\tau_0(w)$ can be discerned by simply applying $\tau_0 \circ \tau^{-1}$ to the received message $\tau(w)$. More generally, whenever there exists a function f such that $\tau_0 = f \circ \tau$, Person 2 can recover the value of

$\tau_0(w)$. For example, if $\tau = \tau_0$ then f is the identity function. If no such function f exists, it may not be possible for Person 2 to recover the value of $\tau_0(w)$. Instead, they may predict $\tau_0(w)$ based on $\tau(w)$ via $\gamma_0(\tau, w)$. Averaged over $W \sim P_W$, the MSE of this approach is precisely $\mathrm{DIS}(\tau)$. See Equation 3 in Kuleshov & Liang (2015) for a related decomposition of $E[\{Y - \tau(X)\}^2] = \mathrm{MSE}(\tau) + E[\{Y - \tau_0(X)\}^2]$ derived in the context of probability forecasting.

The calibration-distortion decomposition shows that, at a given level of distortion, better-calibrated treatment effect predictors have lower MSE for the true CATE function. We will explore this fact later in this work when showing that, in addition to improving calibration, our proposed calibration procedure can improve the MSE of CATE predictors.

2.3. Calibrating predictors: desiderata and classical methods

In most calibration methods, the key goal is to find a function $\theta : \mathbb{R} \to \mathbb{R}$ of a given predictor τ such that $\mathrm{CAL}(\theta \circ \tau) < \mathrm{CAL}(\tau)$, where $\theta \circ \tau$ refers to the composed predictor $w \mapsto \theta(\tau(w))$. A mapping θ that pursues this objective is referred to as a *calibrator*. Ideally, a calibrator θ_n for τ constructed from the dataset $\mathcal{D}_n$ should satisfy the following desiderata:

Property 1: $\mathrm{CAL}(\theta_n \circ \tau)$ tends to zero quickly as n grows;

Property 2: $\theta_n \circ \tau$ and τ are comparably predictive of τ_0.

Property 1 states the primary objective of a calibrator, that is, to yield a well-calibrated predictor. Property 2 requires that the calibrator not destroy the predictive power of the initial predictor in the pursuit of Property 1, which would occur if the calibration term in decomposition (3) were made small at the cost of dramatic inflation of the distortion term.

In the traditional setting of classification and regression, a natural aim is to learn, for $a \in \{0, 1\}$, a predictor $w \mapsto \nu^a(w)$ of the outcome Y among individuals with treatment $A = a$. The best possible such predictor is given by the treatment-specific outcome regression $w \mapsto \mu_0(a, w)$. For $a \in \{0, 1\}$, $\nu^{(a)}$ is said to be calibrated for the outcome regression if $\nu^{(a)}(w) \approx E(Y \mid \nu^{(a)}(W) = \nu^{(a)}(w)), A = a$ for P_0-almost every w. Such a calibrated predictor can be obtained using existing calibration methods for regression (Huang et al., 2020), which we review in the next paragraph. It is natural to wonder, then, whether existing calibration approaches can be directly used to calibrate for the CATE. As a concrete example, given predictors $\nu^{(1)}$ and $\nu^{(0)}$ of $\mu_0(1, \cdot)$ and $\mu_0(0, \cdot)$, a natural CATE predictor is the T-learner $\tau := \nu^{(1)} - \nu^{(0)}$. However, even if $\nu^{(1)}$ and $\nu^{(0)}$ are calibrated for their respective outcome regressions, the predictor τ can still be poorly calibrated for the CATE. Indeed, in settings with treatment-outcome confounding, T-learners can be poorly calibrated when the calibrated predictors $\nu^{(1)}$ and $\nu^{(0)}$ are poor approximations of their respective outcome regressions. As an extreme example, suppose that $\nu^{(a)}$ equals the constant predictor $w \mapsto E(Y \mid A = a)$ for $a \in \{0, 1\}$, which is perfectly calibrated for the outcome regression. Then, the corresponding T-learner $\tau(\cdot) = E(Y \mid A = 1) - E(Y \mid A = 0)$ typically has poor calibration for the CATE in observational settings.

In classification and regression settings (Huang et al., 2020), the most commonly used calibration methods include Platt's scaling (Platt et al., 1999), histogram binning (Zadrozny & Elkan, 2001), Bayesian binning into quantiles (Naeini et al., 2015), and isotonic calibration (Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005). Broadly, Platt's scaling is designed for binary outcomes and uses the estimated values of the predictor to fit the logistic regression model

$$\text{logit} P(Y = 1 \mid \tau(W) = t) = \alpha + \beta t$$

with $\alpha, \beta \in \mathbb{R}$. While it typically satisfies Property 2, Platt's scaling is based on strong parametric assumptions and, as a consequence, may lead to predictions with significant calibration error, even asymptotically (Gupta et al., 2020). Nevertheless, Platt's scaling may be preferred when limited data is available. Histogram binning, also known as quantile binning, involves partitioning the sorted values of the predictor into a fixed number of bins. Given an initial prediction, the calibrated prediction is given by the empirical mean of the observed outcome values within the corresponding prediction bin. A significant limitation of histogram binning is that it requires a priori specification of the number of bins. Selecting too few bins can significantly degrade the predictive power of the calibrated predictor, whereas selecting too many bins can lead to poor calibration. Bayesian binning improves upon histogram binning by considering multiple binning models and their combinations; nevertheless, it still suffers from the need to pre-specify binning models and prior distributions.

Isotonic calibration is a histogram binning method that learns the bins from data using isotonic regression, a nonparametric method traditionally used for estimating monotone functions (Barlow & Brunk, 1972; Martino et al., 2019; Huang et al., 2020). Specifically, the bins are selected by minimizing an empirical MSE criterion under the constraint that the calibrated predictor is a nondecreasing monotone transformation of the original predictor. Isotonic calibration is motivated by the heuristic that, for a good predictor τ, the calibration function $\gamma_0(\tau, \cdot)$ should be approximately monotone as a function of τ. For instance, when $\tau = \tau_0$, the map $\tau_0 \mapsto \gamma_0(\tau_0, \cdot) = \tau_0$ is the identity function. Despite its popularity and strong performance in practice (Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005; Gupta & Ramdas, 2021), to date, whether isotonic calibration satisfies distribution-free calibration guarantees remains an open question (Gupta, 2022). In this work, we will show that isotonic calibration satisfies a distribution-free calibration guarantee in the sense of Property 1. We further establish that Property 2 holds, in that the isotonic selection criterion ensures that the calibrated predictor is at least as predictive as the original predictor up to negligible error.

3. Causal isotonic calibration

In real-world experiments, Dwivedi et al. (2020) found empirically that state-of-the-art CATE estimators tend to be poorly calibrated. However, strikingly, the authors found that such CATE predictors can often still correctly rank the average treatment effect among subgroups defined by bins of the predicted effects. These findings support the heuristic that the calibration function $\gamma_0(\tau_0, \cdot)$ is often approximately monotone as a function of

the predictor τ. This heuristic makes extending isotonic calibration to the CATE setting especially appealing since the monotonicity constraint ensures that the calibrated predictions preserve the (non-strict) ranking of the original predictions.

Inspired by isotonic calibration, we propose a doubly-robust calibration method for treatment effects, which we refer to as *causal isotonic calibration*. Causal isotonic calibration takes a given predictor trained on some dataset and performs calibration using an independent (or hold-out) dataset. Mechanistically, causal isotonic calibration first automatically learns uncalibrated regions of the given predictor. Calibrated predictions are then obtained by consolidating individual predictions within each region into a single value using a doubly-robust estimator of the ATE. In addition, we introduce a novel data-efficient variant of calibration which we refer to as cross-calibration. In contrast with the standard calibration approach, *causal isotonic cross-calibration* takes cross-fitted predictors and outputs a single calibrated predictor obtained using all available data. Our methods can be implemented using standard isotonic regression software.

Let τ be a given treatment effect predictor assumed, for now, to have been built using an external dataset, and suppose that $\mathscr{D}_n$ is the available calibration dataset. In general, we can calibrate the predictor τ using regression-based calibration methods by employing an appropriate surrogate outcome for the CATE. For both experimental and observational settings, a surrogate outcome with favorable efficiency and robustness properties is the pseudo-outcome $\chi_0(O)$ defined via the mapping

$$\chi_0 : o \mapsto \tau_0(w) + \frac{a - \pi_0(w)}{\pi_0(w)[1 - \pi_0(w)]}[y - \mu_0(a, w)], \tag{4}$$

with $o := (w, a, y)$ representing a realization of the data unit. This pseudo-outcome has been used as surrogate for the CATE in previous methods for estimating τ_0, including the DR-learner (Luedtke & van der Laan, 2016; Kennedy, 2020). If χ_0 were known, an external predictor τ could be calibrated using $\mathscr{D}_n$ by isotonic regression of the pseudo-outcomes $\chi_0(O_1), \chi_0(O_2), \ldots, \chi_0(O_n)$ onto the calibration sample predictions $\tau(W_1), \tau(W_2), \ldots, \tau(W_n)$. However, χ_0 depends on π_0 and μ_0, which are usually unknown and must be estimated.

A natural approach for calibrating treatment effect predictors using isotonic regression is as follows. First, define χ_n as the estimated pseudo-outcome function based on estimates μ_n and π_n derived from $\mathscr{D}_n$. Then, a calibrated predictor is given by $\theta_n \circ \tau$, where the calibrator θ_n is found via isotonic regression as a minimizer over $\mathscr{F}_{iso} := \{\theta : \mathbb{R} \to \mathbb{R};\ \theta \text{ is monotone nondecreasing}\}$ of the empirical least-squares risk function

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} [\chi_n(O_i) - \theta \circ \tau(W_i)]^2 .$$

However, this optimization problem requires a double use of $\mathscr{D}_n$: once, for creating the pseudo-outcomes $\chi_n(O_i)$, and a second time, in the calibration step. This double usage could lead to over-fitting (Kennedy, 2020), and so we recommend obtaining pseudo-outcomes via

sample splitting or cross-fitting. Sample splitting involves randomly partitioning $\mathscr{D}_n$ into $\mathscr{E}_m \cup \mathscr{C}_\ell$, with $\mathscr{E}_m$ used to estimate μ_0 and π_0, and $\mathscr{C}_\ell$ used to carry out the calibration step — see Algorithm 1 for details. Cross-fitting improves upon sample splitting by using all available data to estimate μ_0 and π_0 as well as to carry out the calibration step. Algorithm 4, outlined in Appendix B, is the cross-fitted variant of Algorithm 1.

Algorithm 1 Causal isotonic calibration

Require: predictor τ, training data $\mathscr{E}_m$, calibration data $\mathscr{C}_\ell$

1: obtain estimate χ_m of χ_0 using $\mathscr{E}_m$;

2: perform isotonic regression to find

$$\theta_n^* = \underset{\theta \in \mathscr{F}_{iso}}{\operatorname{argmin}} \sum_{i \in \mathscr{I}_{ell}} [\chi_m(O_i) - \theta \circ \tau(W_i)]^2$$

with $\mathscr{I}_{ell}$ the set of indices for observations in $\mathscr{C}_\ell \subset \mathscr{D}_n$;

3: set $\tau_n^* = \theta_n^* \circ \tau$.

Ensure: τ_n^*

In practice, the external dataset used to construct τ for input into Algorithm 1 is likely to arise from a sample splitting approach wherein a large dataset is split in two, with one half used to estimate τ and the other to calibrate it. This naturally leads to the question of whether there is an approach that fully utilizes the entire dataset for both fitting an initial estimate of τ_0 and calibration. Algorithm 2 describes causal isotonic cross-calibration, which provides a means to accomplish precisely this. In brief, this approach applies Algorithm 1 a total of k times on different splits of the data, where for each split an initial predictor of τ_0 is fitted based on the first subset of the data and this predictor is calibrated using the second subset. These k calibrated predictors are then aggregated via a pointwise median. Interestingly, other aggregation strategies, such as pointwise averaging, can lead to uncalibrated predictions (Gneiting & Ranjan, 2013; Rahaman & Thiery, 2020). A computationally simpler variant of Algorithm 2 is given by Algorithm 3. In this implementation, a single isotonic regression is performed using the pooled out-of-fold predictions; this variant may also yield more stable performance in finite-samples than Algorithm 2 — see Section 2.1.2 of Xu & Yadlowsky (2022) for a related discussion in the context of debiased machine learning.

Algorithm 2 Causal isotonic cross-calibration (unpooled)

Require: dataset $\mathscr{D}_n$, # of cross-fitting splits k

1: partition $\mathscr{D}_n$ into datasets $\mathscr{E}^{(1)}, \mathscr{E}^{(2)}, \ldots, \mathscr{E}^{(k)}$;
2: **for** $s = 1, 2, \ldots, k$ **do**
3: set $\mathscr{E}^{(s)} := \mathscr{D}_n \backslash \mathscr{E}^{(s)}$;
4: get initial predictor $\tau_{n,s}$ of τ_0 using $\mathscr{E}^{(s)}$;
5: get calibrated predictor $\tau^*_{n,s}$ via Alg. 1 using predictor $\tau_{n,s}$, training data $\mathscr{E}^{(s)}$, and calibration data $\mathscr{E}^{(s)}$;
6: **end for**
7: set $\tau^*_n : w \mapsto \text{median}\{\tau^*_{n,1}(w), \tau^*_{n,2}(w), \ldots, \tau^*_{n,k}(w)\}$.

Ensure: τ^*_n

4. Large-sample theoretical properties

We now present theoretical results for causal isotonic calibration. We first obtain results for causal isotonic calibration described by Algorithm 1 applied to a fixed predictor τ. We also establish MSE guarantees for the calibrated predictor and argue that the proposed calibrator satisfies Properties 1 and 2. We then extend our results to the procedure described by Algorithm 2.

For ease of presentation, we only establish theoretical results for the case where the nuisance estimators are obtained using sample splitting. With minor modifications, our results can be readily extended to cross-fitting by arguing along the lines of Newey & Robins (2018). In that spirit, we assume that the available data $\mathscr{D}_n$ is the union of a training dataset $\mathscr{E}_m$ and a calibration dataset $\mathscr{C}_\ell$ of sizes m and ℓ, respectively, with $n = m + \ell$ and $\min\{m, \ell\} \to \infty$ as

Algorithm 3 Causal isotonic cross-calibration (pooled)

Require: dataset $\mathscr{D}_n$, # of cross-fitting splits k

1: partition $\mathscr{D}_n$ into datasets $\mathscr{E}^{(1)}, \mathscr{E}^{(2)}, \ldots, \mathscr{E}^{(k)}$;
2: **for** $s = 1, 2, \ldots, k$ **do**
3: let $j(i) = s$ for each $i \in \mathscr{E}^{(s)}$;
4: set $\mathscr{E}^{(s)} := \mathscr{D}_n \backslash \mathscr{E}^{(s)}$;
5: get estimate predictor $\chi_{n,s}$ of χ_0 from $\mathscr{E}^{(s)}$;
6: get initial predictor $\tau_{n,s}$ of τ_0 using $\mathscr{E}^{(s)}$;
7: **end for**
8: perform isotonic regression using pooled out-of-fold predictions to find

$$\theta_n^* = \underset{\theta \in \mathscr{F}_{iso}}{\operatorname{argmin}} \sum_{i=1}^{n} [\chi_{n,j(i)}(O_i) - \theta \circ \tau_{n,j(i)}(W_i)]^2$$

9: set $\tau_{n,s}^* := \theta_n^\theta \circ \tau_{n,s}$ for $s = 1, 2, \ldots, k$;
10: set $\tau_n^* : w \mapsto \text{median}\{\tau_{n,1}^*(w), \tau_{n,2}^*(w), \ldots, \tau_{n,k}^*(w)\}$.
Ensure: τ_n^*

$n \to \infty$. Let τ_n^* be the calibrated predictor obtained from Algorithm 1 using τ, $\mathscr{E}_m$ and $\mathscr{C}_\ell$ where the estimated pseudo-outcome χ_m is obtained by substituting estimates π_m and μ_m of π_0 and μ_0 into (4).

Condition 4.1 (bounded outcome support). The P-support $\mathscr{Y}$ of Y is a uniformly bounded subset of $\mathbb{R}$.

Condition 4.2 (positivity). There exists $\epsilon > 0$ such that $P(\epsilon < \pi_0 W) < 1 - \epsilon) = 1$.

Condition 4.3 (independence). Estimators π_m and μ_m do not use any data in $\mathscr{C}_\ell$.

Condition 4.4 (bounded range of π_m, μ_m, τ). There exist $0 < \eta$, $\alpha < \infty$ such that $P(\eta < \pi_m(W) < 1 - \eta) = P(\mid \mu_m(A, W) \mid < \alpha) = P(\mid \tau(W) \mid < \alpha) = 1$ for $m = 1, 2, \ldots$

Condition 4.5 (bounded variation of best predictor). The function $\theta_0 : \mathbb{R} \mapsto \mathbb{R}$ such that $\theta_0 \circ \tau = \gamma_0(\tau, \cdot)$ is of bounded total variation.

It is worth noting that the initial predictor and its best monotone transformation can be arbitrarily poor CATE predictors. Condition 4.1 holds trivially when outcomes are binary, but even continuous outcomes are often known to satisfy fixed bounds (e.g., physiologic bound, limit of detection of instrument) in applications. Condition 4.2 is standard in causal inference and requires that all individuals have a positive probability of being assigned to either treatment or control. Condition 4.3 follows as a direct consequence of the sample splitting approach, because the estimators are obtained from an independent sample from the data used to carry the calibration step. Condition 4.4 requires that the estimators of the outcome regression and propensity score be bounded; this can be enforced, for example,

by threshholding when estimating these regression functions. Condition 4.5 excludes cases in which the best possible predictor of the CATE given only the initial predictor τ has pathological behavior, in the sense that it has infinite variation norm as a (univariate) mapping of τ. We stress here that isotonic regression is used only as a tool for calibration, and our theoretical guarantees do not require any monotonicity on components of the data-generating mechanism — for example, $\gamma_0(\tau, w)$ need not be monotone as a function of $\tau(w)$.

The following theorem establishes the calibration rate of the predictor τ_n^* obtained using causal isotonic calibration.

Theorem 4.6 (τ_n^* is well-calibrated). Under Conditions 4.1–4.5, as $n \to \infty$, it holds that

$$\mathrm{CAL}(\tau_n^*) = O_P\left(\ell^{-2/3} + \|(\pi_m - \pi_0)(\mu_m - \mu_0)\|^2\right).$$

The calibration rate can be expressed as the sum of an oracle calibration rate and the rate of a second-order cross-product bias term involving nuisance estimators. Notably, the causal isotonic calibrator rate can satisfy Property 1 at the oracle rate $\ell^{-2/3}$ so long as $\|(\pi_m - \pi_0)(\mu_m - \mu_0)\|$ shrinks no slower than $\ell^{-1/3}$, which requires that one or both of π_0 and μ_0 is estimated well in an appropriate sense. If π_0 is known, as in most randomized experiments, the fast calibration rate of $\ell^{-2/3}$ can be achieved even when μ_m is inconsistent, thereby providing distribution-free calibration guarantees irrespective of the smoothness of the outcome regression or dimension of the covariate vector. When $\boldsymbol{\pi}0$ is unknown, the oracle rate of $\ell^{-2/3}$ may not be achievable if the propensity score and outcome regression are insufficiently smooth relative to the dimension of the covariate vector (Kennedy, 2020; Kennedy et al., 2022).

It is interesting to contrast the calibration guarantee in Theorem 4.6 with existing MSE guarantees for DR-learners (Kennedy, 2020) since, in view of (3), they also provide calibration guarantees. While the MSE estimation rates for the CATE depend on the dimension and smoothness of τ_0, the curse of dimensionality for our calibration rates only manifests itself in the doubly-robust cross-remainder term that involves nuisance estimation rates. For instance, when $\ell = m = n/2$, if π_0 and μ_0 are known to be Hölder smooth with exponent $\alpha \geq 1$, the calibration rate implied by Theorem 4.6 with minimax optimal nuisance estimators is, up to logarithmic factors, $\ell^{-2/3} + \ell^{-4\alpha/(2\alpha+d)}$. In contrast, if τ_0 is known to be Hölder smooth with exponent $\beta \geq 1$, a minimax optimal estimator of τ_0 is only guaranteed to achieve an MSE, and therefore calibration, rate of $\ell^{-2\beta/(2\beta+d)} + \ell^{-4\alpha/(2\alpha+d)}$ (Kennedy et al., 2022). Moreover, when the nuisance smoothness satisfies $\alpha \geq d/4$, causal isotonic calibration can achieve the oracle calibration rate of $\ell^{-2/3}$, whereas a minimax optimal CATE estimator is only guaranteed to achieve the same calibration rate under the stringent condition that the smoothness of τ_0 satisfies $\beta \geq d$.

The following theorem states that the predictor obtained by taking pointwise medians of calibrated predictors is also calibrated.

Theorem 4.7 (Pointwise median preserves calibration). Let $\tau^*_{n,1}, \tau^*_{n,2}, \ldots, \tau^*_{n,k}$ be predictors, and define pointwise $\tau^*_{n,k}(w) \coloneqq \text{median}\{\tau^*_{n,1}(w), \tau^*_{n,2}(w), \ldots, \tau^*_{n,k}(w)\}$. Then

$$\text{CAL}(\tau^*_n) \leq k \sum_{s=1}^{k} \text{CAL}(\tau^*_{n,s}),$$

where the median operation is defined as in Section 2.1.

Under similar conditions, Theorem 4.7 combined with a generalization of Theorem 4.6 that handles random τ (see Theorem C.5 in Appendix C.4) establishes that a predictor $\tau^*_{n,k}$ obtained using causal isotonic cross-calibration (Algorithm 2) has calibration error $\text{CAL}(\tau^*_{n,k})$ of order

$$O_P\left(n^{-2/3} + \max_{1 \leq s \leq k} \|(\pi_{n,s} - \pi_0)(\mu_{n,s} - \mu_0)\|^2\right)$$

as $n \to \infty$, where $\mu_{n,s}$ and $\pi_{n,s}$ are the outcome regression and propensity score estimators obtained after excluding the s^{th} fold of the full dataset. In fact, Theorem 4.7 is valid for any calibrator of the form $\tau^*_n : w \mapsto \tau^*_{n,s_n(w)}(w)$, where $s_n(w)$ is any random selector that may depend on the covariate value w. This suggests that the calibration rate for the median-aggregated calibrator implied by Theorem 4.7 is conservative as it also holds for the worst-case oracle selector that maximizes calibration error.

We now establish that causal isotonic calibration satisfies Property 2, that is, it maintains the predictive accuracy of the initial predictor τ. In what follows, predictive accuracy is quantified in terms of MSE. At first glance, the calibration-distortion decomposition appears to raise concerns that causal isotonic calibration may distort τ so much that the predictive accuracy of τ^*_n may be worse than that of τ. This possibility may seem especially concerning given that the ouput of isotonic regression is a step function, so that there could be many $w, w' \in \mathscr{W}$ such that $\tau(w) \neq \tau(w')$ but $\tau^*_n(w) = \tau^*_n(w')$. The following theorem alleviates this concern by establishing that, up to a remainder term that decays with sample size, the MSE of τ^*_n is no larger than the MSE of the initial CATE predictor τ. A consequence of this theorem is that causal isotonic calibration does not distort τ so much as to destroy its predictive performance. To derive this result, we leverage that τ^*_n is in fact a misspecified DR-learner of the univariate CATE function $\gamma_0(\tau, \cdot)$. While isotonic calibrated predictors are calibrated even when $\gamma_0(\tau, \cdot)$ is not a monotone function of τ, we stress that misspecified DR-learners for $\gamma_0(\tau, \cdot)$ are typically uncalibrated.

In the theorem below, we define the best isotonic approximation of the CATE given the initial predictor τ as

$$\tau_0^* := \underset{\theta \circ \tau : \theta \in \mathscr{F}_{iso}}{\operatorname{argmin}} \|\tau_0 - \theta \circ \tau\|.$$

Theorem 4.8 (Causal isotonic calibration does not inflate MSE much). *Under* Conditions 4.1—4.5,

$$\|\tau_n^* - \tau_0^*\| = O_P\left(\ell^{-1/3} + \|(\pi_m - \pi_0)(\mu_m - \mu_0)\|\right)$$

as $n \to \infty$. *As such, as* $n \to \infty$, *the inflation in root MSE from causal isotonic calibration satisfies*

$$\sqrt{\text{MSE}\tau_n^*} - \sqrt{\text{MSE}(\tau)} \leq O_P\left(\ell^{-1/3} + \|(\pi_m - \pi_0)(\mu_m - \mu_0)\|\right).$$

A similar MSE bound can be established for causal isotonic cross-calibration as defined in Algorithm 2.

5. Simulation studies

5.1. Data-generating mechanisms

We examined the behavior of our proposal under two data-generating mechanisms. The first mechanism (Scenario 1) includes a binary outcome whose conditional mean is an additive function (on the logit scale) of non-linear transformations of four confounders with treatment interactions. The second mechanism (Scenario 2) includes instead a continuous outcome with conditional mean linear on covariates and treatment interactions, with more than 100 covariates of which only 20 are true confounders. In both scenarios, the propensity score follows a logistic regression model. All covariates were independent and uniformly distributed on (−1, +1). Sample sizes 1,000, 2,000, 5,000 and 10,000 were considered. Further details are given in Appendix D.1.

5.2. CATE estimation

In Scenario 1, to estimate the CATE, we implemented gradient-boosted regression trees (GBRT) with maximum depths equal to 2, 5, and 8 (Chen & Guestrin, 2016), random forests (RF) (Breiman, 2001), generalized linear models with lasso regularization (GLMnet) (Friedman et al., 2010), generalized additive models (GAM) (Wood, 2017), and multivariate adaptive regression splines (MARS) (Friedman, 1991). In Scenario 2, we implemented RF, GLMnet, and a combination of variable screening with lasso regularization followed by GBRT with maximum depth determined via cross-validation. We used the implementation of these estimators found in R package `sl3` (Coyle et al., 2021). Causal isotonic cross-calibration was implemented using the variant outlined in Algorithm 3. Further details are given in Appendix D.2.

5.3. Performance metrics

We compared the performance of the causal isotonic calibrator to its uncalibrated version in terms of three metrics: the calibration measure defined in (1), MSE, and the calibration bias within bins defined by the first and last prediction deciles. The calibration bias within bins is given by the measure in (2) standardized by the probability of falling within each bin. For each simulation iteration, the metric was estimated empirically using an independent sample $\mathscr{V}$ of size $n_{\mathscr{V}} = 10^4$. These metric estimates were then averaged across 500 simulations. Additional details on these metrics is provided in Appendix D.3.

5.4. Simulation results

Results from Scenario 1 are summarized in Figure 3. The predictors based on GLMnet, RF, GAM, and MARS happened to be well-calibrated, and so, causal isotonic calibration did not lead to noticeable improvements in calibration error. In contrast, causal isotonic calibration of GBRT substantially decreased its calibration error, regardless of tree depth and sample size. In terms of MSE, calibration improved the predictive performance of GBRT and GAM, and preserved the performance of GLMnet and MARS. The calibration bias within bins of prediction was generally smaller after calibration, with a more notable improvement on GBRT — see Table 2 in Appendix E.

Results from Scenario 2 are summarized in Figure 2. The predictors based on RF and GBRT with GLMnet screening were poorly calibrated, and causal isotonic calibration substantially reduced their calibration error. Calibration did not noticeably change the already small calibration error of the GLMnet predictions; however, calibration substantially reduced the calibration error within quantile bins of its predictions — see Table 3 in Appendix E. Finally, with respect to MSE, causal isotonic calibration improved the performance of RF and GBRT with variable screening, and yielded similar performance to GLMnet.

In Figure 4 of Appendix E, we compared calibration performance using hold-out sets to cross-calibration. We found substantial improvements in MSE and calibration by using cross-calibration over conventional calibration.

6. Conclusion

In this work, we proposed causal isotonic calibration as a novel method to calibrate treatment effect predictors. In addition, we established that the pointwise median of calibrated predictors is also calibrated. This allowed us to develop a data-efficient variant of causal isotonic calibration using cross-fitted predictors, thereby avoiding the need for a hold-out calibration dataset. Our proposed methods guarantee that, under minimal assumptions, the calibration error defined in (2) vanishes at a fast rate of $\ell^{-2/3}$ with little or no loss in predictive power, where ℓ denotes the number of observations used for calibration. This property holds regardless of how well the initial predictor τ approximates the true CATE function. To our knowledge, our method is the first in the literature to directly calibrate CATE predictors without requiring trial data or parametric assumptions. Potential applications of our method include data-driven decision-making with strong robustness guarantees. In future work, it would be interesting to study whether pairing causal isotonic

cross-calibration with conformal inference (Lei & Candès, 2021) leads to improved ITE prediction intervals, and whether causal isotonic calibration and shape-constrained inference methods (Westling & Carone, 2020) can be used to construct confidence intervals for the calibration function $\gamma_0(\tau_n^*, \cdot)$.

Our method has limitations. Its calibration guarantees require that either μ_0 or π_0 be estimated sufficiently well. Flexible learning methods can be used to satisfy this condition. If π_0 is known, this condition can be trivially met. Hence, our method can be readily used to calibrate CATE predictors and characterize HTEs in clinical trials. For proper calibration, our method requires all confounders to be measured and adjusted for. In future work, it will be important to study CATE calibration in the context of unmeasured confounding. Our strategy could be adapted to construct calibrators for general learning tasks, including E-learning of the conditional relative risk (Jiang et al., 2019; Qiu et al., 2019), proximal causal learning (Tchetgen et al., 2020; Sverdrup & Cui, 2023), and instrumental variable-based learning (Okui et al., 2012; Syrgkanis et al., 2019).

In simulations, we found that causal isotonic cross-calibration led to well-calibrated predictors without sacrificing predictive performance; benefits were especially prominent in high-dimensional settings and for tree-based methods. This is of particularly high relevance given that regression trees have become popular for CATE estimation due to their flexibility (Athey & Imbens, 2016) and interpretability (Lee et al., 2020). We also found that cross-calibration substantially improved the MSE of the calibrated predictor relative to hold-out set calibration approaches and can even improve the MSE of the original predictor.

Though our focus was on treatment effect estimation, our theoretical arguments can be readily adapted to provide guarantees for isotonic calibration in regression and classification problems. Hence, we have provided an affirmative answer to the open question of whether it is possible to establish distribution-free calibration guarantees for isotonic calibration (Gupta, 2022).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements.

Research reported in this publication was supported by NIH grants DP2-LM013340 and R01-HL137808. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Athey S Beyond prediction: Using big data for policy problems. Science, 355(6324):483–485, 2017. [PubMed: 28154050]

Athey S and Imbens G Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27):7353–7360, 2016.

Athey S and Wager S Estimating treatment effects with causal forests: An application. Observational Studies, 5 (2):37–51, 2019.

Barlow RE and Brunk HD The isotonic regression problem and its dual. Journal of the American Statistical Association, 67(337):140–147, 1972.

Bella A, Ferri C, Hernández-Orallo J, and Ramírez-Quintana MJ Calibration of machine learning models. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pp. 128–146. IGI Global, 2010.

Breiman L Random forests. Machine learning, 45(1):5–32, 2001.

Brooks J, van der Laan MJ, and Go AS Targeted maximum likelihood estimation for prediction calibration. The international journal of biostatistics, 8(1), 2012.

Carnegie N, Dorie V, and Hill JL Examining treatment effect heterogeneity using bart. Observational Studies, 5 (2):52–70, 2019.

Chen T and Guestrin C Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, 2016.

Chernozhukov V, Demirer M, Duflo E, and Fernandez-Val I Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

Collins FS and Varmus H A new initiative on precision medicine. New England journal of medicine, 372(9):793–795, 2015. [PubMed: 25635347]

Coyle JR, Hejazi NS, Malenica I, Phillips RV, and Sofrygin O sl3: Modern pipelines for machine learning and Super Learning. https://github.com/tlverse/sl3, 2021. URL 10.5281/zenodo.1342293. R package version 1.4.2.

Cucchiara V, Cooperberg MR, Dall'Era M, Lin DW, Montorsi F, Schalken JA, and Evans CP Genomic markers in prostate cancer decision making. European urology, 73(4):572–582, 2018. [PubMed: 29129398]

Dahabreh IJ, Hayward R, and Kent DM Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. International journal of epidemiology, 45(6):2184–2193, 2016. [PubMed: 27864403]

Devriendt F, Moldovan D, and Verbeke W A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big data, 6(1):13–41, 2018. [PubMed: 29570415]

Dominici F, Bargagli-Stoffi FJ, and Mealli F From controlled to undisciplined data: estimating causal effects in the era of data science using a potential outcome framework. arXiv preprint arXiv:2012.06865, 2020.

Dwivedi R, Tan YS, Park B, Wei M, Horgan K, Madigan D, and Yu B Stable discovery of interpretable subgroups via calibration in causal studies. International Statistical Review, 88:S135–S178, 2020.

Foster DJ and Syrgkanis V Orthogonal statistical learning. arXiv preprint arXiv:1901.09036, 2019.

Frangakis C The calibration of treatment effects from clinical trials to target populations, 2009.

Friedman J, Hastie T, and Tibshirani R Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1, 2010. [PubMed: 20808728]

Friedman JH Multivariate adaptive regression splines. The annals of statistics, 19(1):1–67, 1991.

Gneiting T and Ranjan R Combining predictive distributions. Electronic Journal of Statistics, 7:1747–1782, 2013.

Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'donnell CJ, et al. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. Journal of the American College of Cardiology, 63(25 Part B):2935–2959, 2014. [PubMed: 24239921]

Groeneboom P and Lopuhaa H Isotonic estimators of monotone densities and distribution functions: basic facts. Statistica Neerlandica, 47(3):175–183, 1993.

Gupta C Post-hoc calibration without distributional assumption, 2022. (Ph.D. proposal).

Gupta C and Ramdas AK Distribution-free calibration guarantees for histogram binning without sample splitting. arXiv preprint arXiv:2105.04656, 2021.

Gupta C, Podkopaev A, and Ramdas A Distribution-free binary classification: prediction sets, confidence intervals and calibration. In Larochelle H, Ranzato M, Hadsell R, Balcan MF, and

Lin H (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 3711–3723. Curran Associates, Inc., 2020.

Huang Y, Li W, Macheret F, Gabriel RA, and Ohno-Machado L A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association, 27(4):621–633, 2020. [PubMed: 32106284]

Imbens GW and Wooldridge JM Recent developments in the econometrics of program evaluation. Journal of economic literature, 47(1):5–86, 2009.

Jiang B, Song R, Li J, and Zeng D Entropy learning for dynamic treatment regimes. Statistica Sinica, 29:1633–1710, 01 2019. doi: 10.5705/ss.202018.0076. [PubMed: 31534307]

Josey KP, Yang F, Ghosh D, and Raghavan S A calibration approach to transportability and data-fusion with observational data. Statistics in Medicine, 41(23):4511–4531, 2022. [PubMed: 35848098]

Kennedy E, Balakrishnan S, and Wasserman L Minimax rates for heterogeneous causal effect estimation. 03 2022.

Kennedy EH Optimal doubly robust estimation of heterogeneous causal effects. arXiv preprint arXiv:2004.14497, 2020.

Kent DM, Steyerberg E, and van Klaveren D Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. Bmj, 363, 2018.

Kuleshov V and Liang PS Calibrated structured prediction. Advances in Neural Information Processing Systems, 28, 2015.

Künzel SR, Sekhon JS, Bickel PJ, and Yu B Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10):4156–4165, 2019.

Lee K, Bargagli-Stoffi FJ, and Dominici F Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. arXiv preprint arXiv:2009.09036, 2020.

Lei L and Candès EJ Conformal inference of counter-factuals and individual treatment effects. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83(5):911–938, 2021.

Leng Y and Dimmery D Calibration of heterogeneous treatment effects in random experiments. Available at SSRN 3875850, 2021.

Lloyd-Jones DM, Braun LT, Ndumele CE, Smith SC Jr, Sperling LS, Virani SS, and Blumenthal RS Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the american heart association and american college of cardiology. Circulation, 139(25):e1162–e1177, 2019. [PubMed: 30586766]

Luedtke AR and van der Laan MJ Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. Annals of statistics, 44(2):713, 2016. [PubMed: 30662101]

Martino A, De Santis E, Baldini L, Rizzi A, et al. Calibration techniques for binary classification problems: A comparative analysis. In IJCCI, pp. 487–495, 2019.

Murphy SA Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2):331–355, 2003.

Naeini MP, Cooper G, and Hauskrecht M Obtaining well calibrated probabilities using bayesian binning. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

Newey WK and Robins JR Cross-fitting and fast remainder rates for semiparametric estimation. arXiv preprint arXiv:1801.09138, 2018.

Niculescu-Mizil A and Caruana R Obtaining calibrated probabilities from boosting. In UAI, volume 5, pp. 413–20, 2005.

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

Obermeyer Z and Emanuel EJ Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine, 375(13):1216, 2016. [PubMed: 27682033]

Okui R, Small DS, Tan Z, and Robins JM Doubly robust instrumental variable regression. Statistica Sinica, pp. 173–205, 2012.

Platt J et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3):61–74, 1999.

Qiu H, Luedtke A, and van der Laan M Contribution to discussion of “entropy learning for dynamic treatment regimes” by jiang b, song r, li j, zeng d. statistica sinica. Statistica Sinica, 29(4):1666–1678, 2019.

Rahaman R and Thiery AH Uncertainty quantification and deep ensembles. arXiv preprint arXiv:2007.08792, 2020.

Robins JM Optimal structural nested models for optimal sequential decisions. In Proceedings of the second seattle Symposium in Biostatistics, pp. 189–326. Springer, 2004.

Royden H Real analysis, the macmillan company. New York, 1963.

Rubin DB Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688, 1974.

Sverdrup E and Cui Y Proximal causal learning of heterogeneous treatment effects. arXiv preprint arXiv:2301.10913, 2023.

Syrgkanis V, Lei V, Oprescu M, Hei M, Battocchi K, and Lewis G Machine learning estimation of heterogeneous treatment effects with instruments. Advances in Neural Information Processing Systems, 32, 2019.

Tchetgen EJT, Ying A, Cui Y, Shi X, and Miao W An introduction to proximal causal learning. arXiv preprint arXiv:2009.10982, 2020.

van Calster B, McLernon DJ, van Smeden M, Wynants L, and Steyerberg EW Calibration: the achilles heel of predictive analytics. BMC medicine, 17(1):1–7, 2019. [PubMed: 30651111]

van der Laan MJ and Rose S Targeted Learning: Causal Inference for Observational and Experimental Data. Springer, New York, 2011.

van der Laan MJ, Polley EC, and Hubbard AE Super learner. Statistical applications in genetics and molecular biology, 6(1), 2007.

van der Vaart AW and Wellner J Weak convergence and empirical processes: with applications to statistics. Springer Science & Business Media, 1996.

van Klaveren D, Balan TA, Steyerberg EW, and Kent DM Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. Journal of clinical epidemiology, 114:72–83, 2019. [PubMed: 31195109]

Wager S and Athey S Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018.

Westling T and Carone M A unified study of nonparametric inference for monotone functions. Annals of statistics, 48(2):1001, 2020. [PubMed: 32704192]

Wood SN Introducing gams. In Generalized additive models, pp. 161–194. Chapman and Hall/CRC, 2017.

Xu Y and Yadlowsky S Calibration error for heterogeneous treatment effects. arXiv preprint arXiv:2203.13364, 2022.

Yadlowsky S, Fleming S, Shah N, Brunskill E, and Wager S Evaluating treatment prioritization rules via rank-weighted average treatment effects. arXiv preprint arXiv:2111.07966, 2021.

Zadrozny B and Elkan C Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In Icml, volume 1, pp. 609–616. Citeseer, 2001.

Zadrozny B and Elkan C Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 694–699, 2002.

Zhang Z, Nie L, Soon G, and Hu Z New methods for treatment effect calibration, with applications to noninferiority trials. Biometrics, 72(1):20–29, 2016. [PubMed: 26363775]

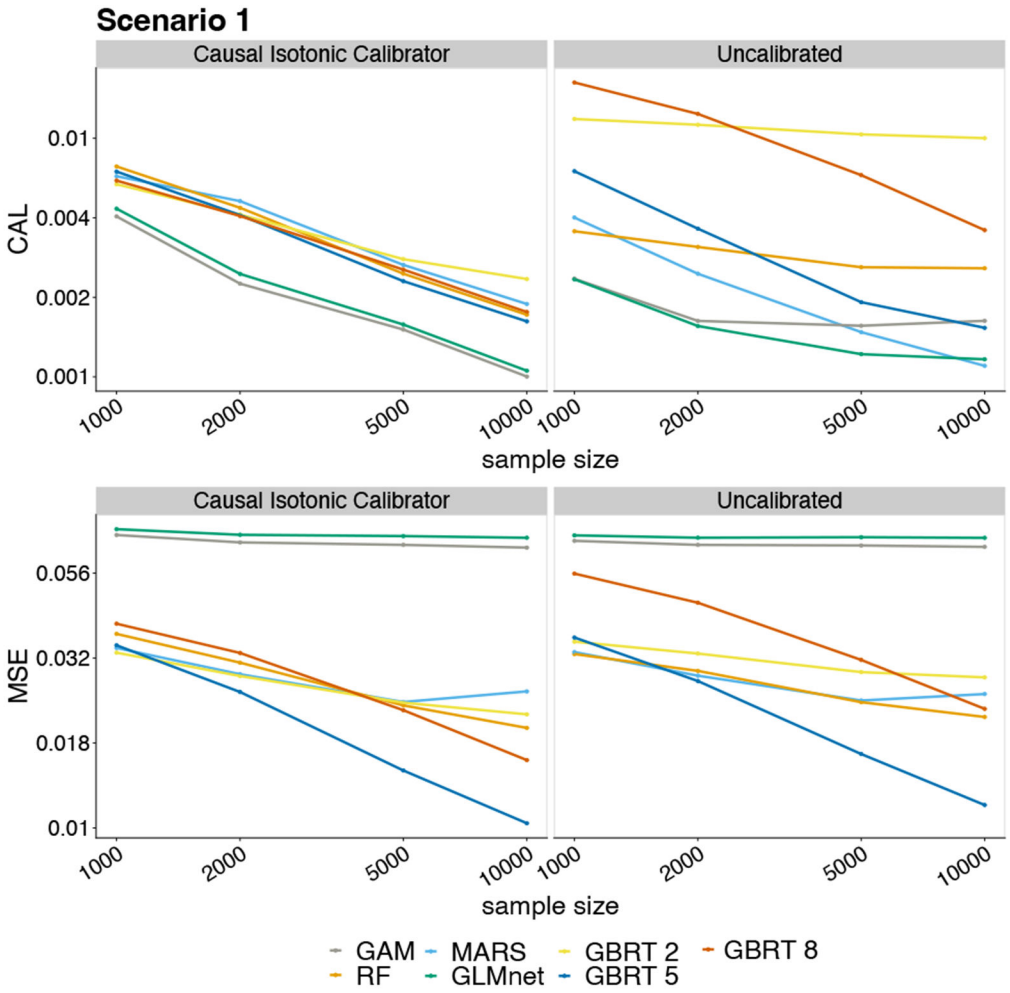

Figure 1.

Calibration error and MSE in Scenario 1. The panels show the calibration error (top) and MSE (bottom) using the calibrated (left) and uncalibrated (right) predictors as a function of sample size. The y-axes are on a log scale.

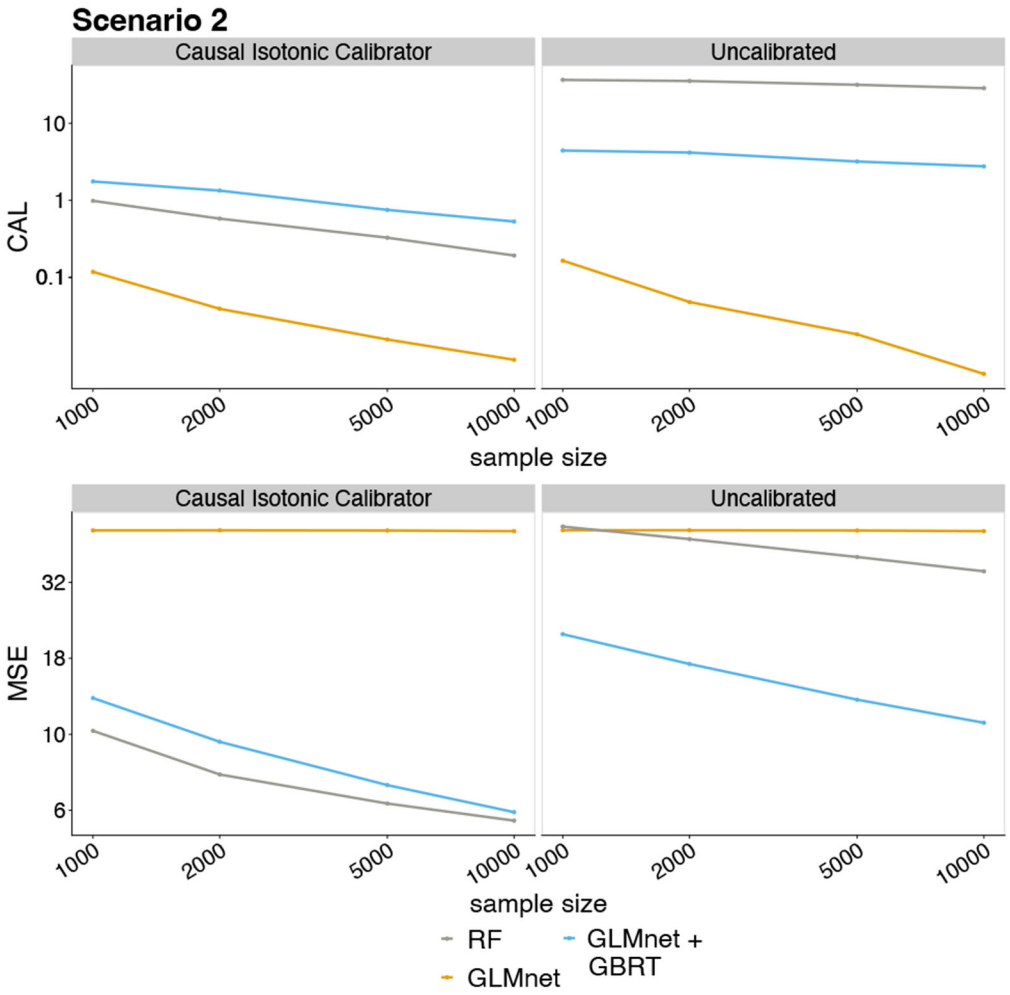

Figure 2.

Calibration error and MSE in Scenario 2. The panels show the calibration error (top) and MSE (bottom) using the calibrated (left) and uncalibrated (right) predictors as a function of sample size. The y-axes are on a log scale.