Supplementary Information for

Universal Adaptability: Target-Independent Inference that Competes with Propensity Scoring

Michael P. Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, Omer Reingold

Correspondence to Shafi Goldwasser.
E-mail: shafi.goldwasser@berkeley.edu

This PDF file includes:
- Supplementary text
- Figs. S1 to S2 (not allowed for Brief Reports)
- Tables S1 to S3 (not allowed for Brief Reports)
- SI References
Supporting Information Text

1. Universal Adaptability via Multi-Calibration: Formal Analysis

First, we establish useful notation and formal definitions for our study of universal adaptability. For ease of exposition, we recap some of the preliminaries covered in the main manuscript. Recall, we assume a joint distribution over covariates $X$, outcomes $Y$, and source/target membership $Z$. Specifically, $\mathcal{X}$ denotes the space of covariates; $\mathcal{Y}$ denotes the outcome range, and $\mathcal{Z} = \{s,t\}$ denotes sampling in the source or target distribution; for simplicity, we assume that $\mathcal{X}$ and $\mathcal{Y}$ are each discrete domains. We make a key technical assumption that the joint law factorizes so that the distribution of $Y$ given $X$ is independent of source and target membership.

\[
\Pr [ X, Y, Z ] = \Pr [ X ] \cdot \Pr [ Y | X ] \cdot \Pr [ Z | X ]
\]

In other words, the relation between $Y$ and $X$ is preserved in the source and target distributions. This type of conditional independence assumption is standard in work that relies on propensity scoring, and is sometimes referred to as ignorability.

We use $D_s$ and $D_t$ to denote the joint distributions over $X,Y$ pairs, conditioned on $Z$.

\[
D_s \equiv \Pr [ X, Y | Z = s ] \\
D_t \equiv \Pr [ X, Y | Z = t ]
\]

For convenience, we use $U_s$ and $U_t$ to denote the distribution over unlabeled samples (i.e., marginal distribution over covariates $X$), conditioned on $Z$.

\[
U_s \equiv \Pr [ X | Z = s ] \\
U_t \equiv \Pr [ X | Z = t ]
\]

Throughout, we work with a uniform prior over $Z$; that is, the marginal probabilities of the source and target are equal.

\[
\Pr [ Z = s ] = \Pr [ Z = t ]
\]

This assumption is convenient for the analysis, but in principle any prior could be selected. Practically, the choice of prior affects the proportion of source and target data we use when estimating propensity scores.

Formalizing Multi-Calibration. Multi-calibration is a formal guarantee of fairness for prediction functions. We work with a generalized notion of multi-calibration where we parameterize the guarantee in terms of a class of real-valued functions $C$, which take values on the nonnegative real interval $\mathbb{R}^+$. Intuitively, the class of functions can be seen as a collection of “subpopulations” where membership is weighted.

**Definition 1.1** (Multi-Calibration). Suppose $C \subseteq \{ \mathcal{X} \times [0,1] \rightarrow \mathbb{R}^+ \}$ is a collection of functions. For a given distribution $\mathcal{D}$ supported on $\mathcal{X} \times \mathcal{Y}$, a predictor $\hat{p} : \mathcal{X} \rightarrow [0,1]$ is $(\mathcal{C}, \alpha)$-multi-calibrated over $\mathcal{D}$ if for all $c \in \mathcal{C}$:

\[
\left| \mathbb{E}_{(X,Y) \sim \mathcal{D}} [ c(X,\hat{p}(X)) \cdot (\hat{p}(X) - Y) ] \right| \leq \alpha.
\]

In fact, our main result can be obtained from a weaker variant of multi-calibration, where the functions $c \in \mathcal{C}$ ignore the prediction $\hat{p}(X)$. This “weak” variant of multi-calibration has been studied in prior works (1, 2), under the name multi-accuracy. Formally, we consider multi-calibration with respect to a class of functions $C \subseteq \{ \mathcal{X} \rightarrow \mathbb{R}^+ \}$ that only takes an individual from $\mathcal{X}$ as input. Then, the multi-calibration guarantee is as follows: a predictor $\hat{p} : \mathcal{X} \rightarrow [0,1]$ is $(\mathcal{C}, \alpha)$-multi-calibrated over $\mathcal{D}$ if for all $c \in \mathcal{C}$:

\[
\left| \mathbb{E}_{(X,Y) \sim \mathcal{D}} [ c(X) \cdot (\hat{p}(X) - Y) ] \right| \leq \alpha.
\]

We note that the formal definitions presented here generalize the original formulation of multi-calibration and multi-accuracy, as defined by (1). In connecting multi-calibration to propensity scoring, the continuous formulation of multi-calibration will be more convenient.

A. Propensity scores for unbiased inference. Recall, the propensity score gives the relationship between the odds of membership in the source versus target distributions for a given set of covariates.

**Definition 1.2** (Propensity Score). For a given source $D_s$ and target $D_t$, the propensity score $e_{st} : \mathcal{X} \rightarrow [0,1]$ is given as

\[
e_{st}(x) = \Pr [ Z = s | X = x ] \\
1 - e_{st}(x) = \Pr [ Z = t | X = x ]
\]
Under a uniform prior over $Z$, we can use the propensity score to obtain unbiased estimates in $D_t$ using labeled samples from $D_s$, as follows. First, we expand the expectation of $Y$ over $D_t$. 

$$
\mathbb{E}_{D_t} [ Y ] = \sum_{x \in \mathcal{X}} \Pr [ X = x \land Z = t ] \cdot \mathbb{E}_{D_t} [ Y | X = x ]
$$

Then, applying Bayes' rule:

$$
= \sum_{x \in \mathcal{X}} \frac{\Pr [ X = x \land Z = t ]}{\Pr [ Z = t ]} \cdot \mathbb{E}_{D_t} [ Y | X = x ]
$$

Again, applying Bayes' rule and the definition of the propensity score $e_{st}(x)$:

$$
= \sum_{x \in \mathcal{X}} \frac{\Pr [ X = x ] \cdot (1 - e_{st}(x))}{\Pr [ Z = t ]} \cdot \mathbb{E}_{D_t} [ Y | X = x ]
$$

Rearranging by multiplying $e_{st}(x)/e_{st}(x) = 1$, and replacing $\Pr [ Z = t ]$ by $\Pr [ Z = s ]$ under the uniform prior assumption:

$$
\sum_{x \in \mathcal{X}} \frac{\Pr [ X = x ]}{\Pr [ Z = s ]} \cdot \frac{1 - e_{st}(x)}{e_{st}(x)} \cdot \mathbb{E}_{D_t} [ Y | X = x ]
$$

Applying Bayes' rule:

$$
= \sum_{x \in X} \Pr [ X = x | Z = s ] \cdot \frac{1 - e_{st}(x)}{e_{st}(x)} \cdot \mathbb{E}_{D_t} [ Y | X = x ]
$$

And finally collapsing the expectation over $D_s$:

$$
\mathbb{E}_{D_s} \left[ \frac{1 - e_{st}(X)}{e_{st}(X)} \right] \cdot Y
$$

This inverse reweighting using the propensity score is essential for obtaining unbiased target estimates using source data. For a propensity score $\sigma : \mathcal{X} \rightarrow [0, 1]$, we denote the propensity odds ratio $c_\sigma : \mathcal{X} \rightarrow \mathbb{R}^+$ as follows:

$$
c_\sigma(x) = \frac{1 - \sigma(x)}{\sigma(x)}
$$

For a class of propensity scores $\Sigma$, we denote the corresponding class of propensity ratios as $C(\Sigma)$.

$$
C(\Sigma) = \{ c_\sigma : \sigma \in \Sigma \}
$$

**Misspecification of Propensity Scores.** Finally, we introduce the technical aspects of how we measure the quality of specification for a class of propensity scores $\Sigma$. A key step in IPSW inference is estimating the propensity score. In practice, to guarantee statistical validity, we must estimate the propensity score using a bounded class of possible propensity score functions $\Sigma \subseteq \{ \mathcal{X} \rightarrow [0, 1] \}$. With $\Sigma$ fixed, we consider the best-fit approximation $\sigma_{st}^* \in \Sigma$ of the propensity score. Abstractly, we can model this choice best-fit propensity score as a loss optimization.

$$
\sigma_{st}^* = \arg\min_{\sigma \in \Sigma} \mathbb{E}_{D_s} [ \ell(\sigma(X), e_{st}(X)) ]
$$

Our results are robust to the choice of loss function, as they can be stated for all $\sigma \in \Sigma$. In other words, regardless of the estimation procedure, we show that universal adaptability can guarantee performance comparable to propensity scoring. In particular, we measure the *misspecification error* of the propensity score estimation. Denote by $d : (0,1) \times (0,1) \rightarrow \mathbb{R}$ the following difference:

$$
d(p, q) = \left| \frac{p}{1 - p} - \frac{q}{1 - q} \right|
$$

We define the misspecification error of $\sigma$ with respect to the true propensity score $e_{st}$ as the expectation of the difference given by $d$ over $D_s$.

**Definition 1.3.** For any source and target distributions $D_s, D_t$, for any propensity score $\sigma : \mathcal{X} \rightarrow [0, 1]$, the misspecification error $\Delta_{st}(\sigma)$ is the following expectation.

$$
\Delta_{st}(\sigma) = \mathbb{E}_{D_s} \left[ d(\sigma(X), e_{st}(X)) \right]
$$
Concretely, we can think of the misspecification error $\Delta_{st}$ as the $\ell_1$-error measured over $D_t$ of $c_{e_{st}}$ compared to the true propensity score $c_{e_{st}}$. An alternative bound on this misspecification error expresses the distance as an expected distance over $D_t$. In particular, define the symmetric $\chi^2$-distance $d_{\chi^2} : (0, 1) \times (0, 1) \to \mathbb{R}$ as follows. First, define the divergence $\chi^2 : (0, 1) \times (0, 1) \to \mathbb{R}$ as

$$\chi^2(p,q) = \left( \frac{p}{q} \right)^2 + \left( \frac{1 - p}{1 - q} \right)^2 - 1$$

Then, $d_{\chi^2}$ symmetrizes the divergence.

$$d_{\chi^2}(p,q) = \frac{1}{2} \left( \chi^2(p,q) + \chi^2(q,p) \right)$$

We argue that the misspecification error can be bounded in terms of the average symmetric $\chi^2$-distance over $D_t$.

**Lemma 1.4.** For any source and target distributions $D_s, D_t$, and for any propensity score $\sigma : X \to [0, 1]$, the following inequality holds.

$$\Delta_{st}(\sigma) \leq \mathbb{E}_{D_t} \left[ d_{\chi^2}(\sigma(X), e_{st}(X)) \right]$$

**Proof.** We begin by expanding the definition of $\Delta_{st}(\sigma)$.

$$\Delta_{st}(\sigma) = \mathbb{E}_{D_s} \left[ d(\sigma(X), e_{st}(X)) \right]$$

$$= \mathbb{E}_{D_s} \left[ \left| \frac{\sigma(X)}{1 - \sigma(X)} - \frac{e_{st}(X)}{1 - e_{st}(X)} \right| \right]$$

$$= \mathbb{E}_{D_t} \left[ \left| \frac{\sigma(X) \cdot (1 - e_{st}(X))}{e_{st}(X) \cdot (1 - \sigma(X))} - 1 \right| \right]$$

By the inequality $pq \geq \frac{p^2 + q^2}{2}$, we obtain the following bound in terms of the symmetric $\chi^2$ distance.

$$\frac{\sigma(X) \cdot (1 - e_{st}(X))}{e_{st}(X) \cdot (1 - \sigma(X))} - 1 \leq \frac{1}{2} \left( \left( \frac{\sigma(X)}{e_{st}(X)} \right)^2 + \left( \frac{1 - e_{st}(X)}{1 - \sigma(X)} \right)^2 \right) - 1$$

$$\leq \frac{1}{2} \left( \chi^2(\sigma(X)||e_{st}(X)) + \chi^2(e_{st}(X)||\sigma(X)) \right)$$

$$= d_{\chi^2}(\sigma(X), e_{st}(X))$$

Thus, in all, $\Delta(\sigma, e_{st}) \leq \mathbb{E}_{D_t} \left[ d_{\chi^2}(\sigma(X), e_{st}(X)) \right]$.

**B. Proof of universal adaptability theorem.** With these preliminaries in place, we are ready to prove the main theorem. For completeness, we begin by stating the result formally.

**Theorem 1.5.** For source and target distributions $D_s$ and $D_t$, with propensity score $e_{st} : X \to [0, 1]$, suppose $\Sigma \subseteq \{X \to \mathbb{R}^+\}$ be the corresponding collection of propensity odds ratios. Suppose $\tilde{p} : X \to [0, 1]$ is a $(C(\Sigma), \alpha)$-multi-calibrated prediction function over the source $D_s$; then, for any target distribution $D_t$, the estimation error of the universally adaptable estimator $\mu_t(\tilde{p})$ is bounded by the sum of the IPSW estimation error, the misspecification error, and $\alpha$, under the best-fit $\sigma \in \Sigma$. That is,

$$\text{er}_t(\mu_t(\tilde{p})) \leq \min_{\sigma \in \Sigma} \{ \text{er}_t(\mu_t^{ps}(\sigma)) + \Delta_{st}(\sigma) + \alpha \}$$

**Proof.** Suppose $\tilde{p}$ is $(C(\Sigma), \alpha)$-multi-calibrated over the source $D_s$. Let $\sigma \in \Sigma$ be a propensity score. We show that the universal adaptable estimator $\mu_t(\tilde{p})$ and the IPSW estimator $\mu_t^{ps}(\sigma)$ are $\beta$-close in absolute distance for some $\beta \geq 0$.

$$|\mu_t(\tilde{p}) - \mu_t^{ps}(\sigma)| \leq \beta$$

By the triangle inequality, we obtain the theorem.

$$\text{er}_t(\mu_t(\tilde{p})) = |\mu_t(\tilde{p}) - \mu_t^*|$$

$$\leq |\mu_t(\tilde{p}) - \mu_t^{ps}(\sigma)| + |\mu_t^{ps}(\sigma) - \mu_t^*|$$

$$\leq \text{er}_t(\mu_t^{ps}(\sigma)) + \beta$$
It remains to give an upper bound on $\beta$. Beginning with the universal adaptable estimator, we show the following series of inequalities.

\[
E_{D_t} [ \hat{p}(X) ] = E_{D_s} \left[ \frac{1 - e_{st}(X)}{e_{st}(X)} \right] \cdot \hat{p}(X) \\
= E_{D_s} \left[ \frac{1 - \sigma(X)}{\sigma(X)} \right] \cdot \hat{p}(X) + E_{D_s} [ d(e_{st}, \sigma) \cdot \hat{p}(X) ].
\]

By the assumption that $\hat{p}$ is $(C(\Sigma), \alpha)$-multi-calibrated, we have that

\[
\left| E_{D_s} \left[ \frac{1 - \sigma(X)}{\sigma(X)} \right] \cdot \hat{p}(X) - E_{D_s} \left[ \frac{1 - \sigma(X)}{\sigma(X)} \right] \cdot Y \right| \leq \alpha.
\]

Together, we can bound the difference between the estimation under universal adaptability and under IPSW.

\[
|\mu(x, \alpha) - \mu^w(x, \sigma)| = \left| E_{D_s} [ \hat{p}(X) ] - E_{D_s} \left[ \frac{1 - \sigma(X)}{\sigma(X)} \right] \cdot Y \right| + \alpha \leq E_{D_s} [ d(e_{st}, \sigma) \cdot \hat{p}(X) ] + \alpha \leq E_{D_s} [ |d(e_{st}, \sigma)| ] + \alpha = \Delta_{st}(\sigma) + \alpha
\]

Thus, for any $\sigma \in \Sigma$, the estimation error of a universally adaptable estimator based on a $(C(\Sigma), \alpha)$-multi-calibrated prediction function is bounded by $\sigma_t(\alpha) + \Delta_{st}(\sigma) + \alpha$, the IPSW estimation error plus the misspecification error and the multi-calibration error. Thus, the inequality holds for the optimal choice of $\sigma^* \in \Sigma$.

\section*{C. Beyond estimation: prediction universal adaptability.}

In fact, Theorem 1.5 is a special case of a more general phenomenon. We show that when a predictor is multi-calibrated over the source distribution, it retains some degree of multi-calibration over target distributions. Intuitively, if the propensity score can be captured by a reasonably simple function class, but $\hat{p}$ is multi-calibrated with respect to a collection of complex functions, then there will still be some complexity “remaining” after accounting for the shift from $D_s$ to $D_t$. We formalize this intuition in Theorem 1.6.

\begin{theorem}
For source and target distributions $D_s$ and $D_t$ with propensity score $e_{st} : \mathcal{X} \rightarrow [0, 1]$, suppose $\Sigma \subseteq \{ X \rightarrow [0, 1] \}$ is a collection of propensity scores, and let $(C(\Sigma) \subseteq \{ X \rightarrow \mathbb{R} \})$ be the corresponding collection of propensity odds ratios. Let $\delta_{st} = \min_{\sigma^* \in \Sigma} \Delta_{st}(\sigma^*)$ be the optimal misspecification error of $\Sigma$ for $D_s$ and $D_t$. Let $\mathcal{F} \subseteq \{ X \times [0, 1] \rightarrow [0, 1] \}$ be a collection of functions. Suppose $\hat{p} : \mathcal{X} \rightarrow [0, 1]$ is a $(C(\Sigma) \otimes \mathcal{F}, \alpha)$-multi-calibrated prediction function over the source $D_s$, where

\[
C(\Sigma) \otimes \mathcal{F} = \{ e_{\sigma^*} : f : e_{\sigma^*} \in C(\Sigma) \land f \in \mathcal{F} \}, \text{ where } e_{\sigma^*} : f(x, v) = e_{\sigma^*}(x) \cdot f(x, v).
\]

Then, for any target distribution $D_t$, $\hat{p}$ is $(\mathcal{F}, \beta)$-multi-calibrated over $D_t$ for $\beta \leq 2\delta_{st} + \alpha$.

\end{theorem}

\begin{proof}
The proof follows similarly to that of Theorem 1.5. Suppose $\hat{p}$ is $(C(\Sigma) \otimes \mathcal{F}, \alpha)$-multi-calibrated. For any $f \in \mathcal{F}$, we can show an inequality between $E_{D_s} [ f(X, \hat{p}(X)) \cdot Y ]$ and $E_{D_s} [ f(X, \hat{p}(X)) \hat{p}(X) ]$, demonstrating a bound on the degree of miscalibration over $c \in C$, establishing multi-calibration over the target distribution.

\[
E_{D_s} [ f(X, \hat{p}(X)) \cdot Y ] = E_{D_s} \left[ \frac{1 - e_{st}(X)}{e_{st}(X)} \right] \cdot f(X, \hat{p}(X)) \cdot Y \\
= E_{D_s} \left[ \frac{1 - \sigma(X)}{\sigma(X)} \right] \cdot f(X, \hat{p}(X)) \cdot Y + E_{D_s} [ d(e_{st}, \sigma) \cdot f(X, \hat{p}(X)) \cdot \hat{p}(X) ].
\]

By the assumption that $\hat{p}$ is $(C(\Sigma) \otimes \mathcal{F}, \alpha)$-multi-calibrated, we have that

\[
\left| E_{D_s} \left[ \frac{1 - \sigma(X)}{\sigma(X)} \right] \cdot f(X, \hat{p}(X)) \cdot Y - E_{D_s} \left[ \frac{1 - \sigma(X)}{\sigma(X)} \right] \cdot f(X, \hat{p}(X)) \cdot Y \right| \leq \alpha.
\]

In combination, we can show the following bound on the multi-calibration error in $D_t$.

\[
\left| E_{D_t} [ f(X, \hat{p}(X)) \cdot Y ] - E_{D_t} [ f(X, \hat{p}(X)) ] \cdot Y \right| \leq 2 \cdot E_{D_s} [ d(e_{st}, \sigma) \cdot f(X, \hat{p}(X)) \cdot \hat{p}(X) ] + \alpha \leq 2\delta_{st}(\sigma) + \alpha.
\]

Thus, we have that $\hat{p}$ is $(\mathcal{F}, \beta)$-multi-calibrated for $\beta \leq 2 \cdot (\min_{\sigma^* \in \Sigma} \Delta_{st}(\sigma^*)) + \alpha$.

\end{proof}
While the statement of Theorem 1.6 is quite technical, it can be understood simply as providing an adaptive guarantee under different degrees of shift. If $\tilde{p}$ is multi-calibrated on the source distribution, then it will only lose a small amount in the multi-calibration guarantee on targets that are close by. Under extreme shifts—shifts that require the full power of the class to account for $C(\Sigma)$—then, the guarantee bottoms out as in Theorem 1.5. The more general theorem helps to contextual the empirical results demonstrating low bias inferences, even when we restrict our attention to structured subpopulations.

Further, Theorem 1.6 has implications for universal adaptability beyond statistical inference. In particular, multi-calibration is known to provide strong accuracy guarantees in the context of binary prediction (2), as well as many desirable recovery guarantees in terms of fairness (1) and ranking quality (3). Thus, this result establishes that each of these results that derive guarantees from multi-calibration, actually provide a distributionally-robust guarantee as well. That is, the qualitative guarantees of multi-calibration are preserved, even under shift in the underlying distribution of covariates.
2. Data, Software, and Algorithms

A public and anonymous repository contains all methods and materials needed to reproduce the experimental results of this work. It is available on the Open Science Framework at:

https://osf.io/kfpr4/?view_only=adf843b070f54be9f529f10944cd99

All data sources used for experiments are included in the README, and are available to the public. Data preparations are conducted in R (3.6.3) (4), model training and evaluation are implemented in Python (3.6.4), using the scikit-learn library (0.19.1) (5). All code used for cleaning the data, evaluating the inference techniques, and generating plots is included in the repository. In the remaining supplementary materials, we describe the experiments in more detail.

A. The MCBoost Algorithm. For completeness, we describe the MCBoost algorithm, developed theoretically by (1) and studied in a more applied setting by (2). The precise implementation and parameter settings used is available in the OSF repository.

MCBoost is best viewed as an algorithm for post-processing a pre-trained predictor to obtain a $(C, \alpha)$-multi-calibrated prediction function. At a high level, MCBoost works by iteratively searching for a function $c \in C$, on which the current predictor is miscalibrated, and then refining the predictor. A key observation in (1, 2) is that the search for miscalibration can be reduced to a regression task over the class $C$. Finally, (1) analyze the procedure, demonstrating that after a bounded number of iterations (as a function of $\alpha$), the iterative process must terminate, because the miscalibration over all functions $c \in C$ will be less than $\alpha$, and thus return a $(C, \alpha)$-multi-calibrated prediction function.

The algorithm takes a few important parameters.

- Regression algorithm: To search for miscalibrated $c \in C$, we need a regression algorithm $A$, that learns the class $C$. That is, the algorithm $A$ takes labeled data $\{(x_1, \ell_1), \ldots, (x_m, \ell_m)\}$ from $X \times [-1, 1]$ and returns a function $c \in C$ such that $c(x)$ best-fits the labels $\ell$.

- Approximation parameter: To know when to stop updating, MCBoost takes an approximation parameter $\alpha > 0$. The number of iterations is bounded by $O(1/\alpha^2)$, which also bounds the complexity of the resulting multi-calibrated prediction function. In practice, it is also useful to include a fixed upper bound on the number of iterations of auditing.

- Calibration and validation data data sets: MCBoost requires labeled data from the (source) distribution to perform multi-calibration. The first “calibration set” $S$ is used for regression auditing of the predictor. The second “validation set” $V$ is used to test when the predictor is sufficiently multi-calibrated (i.e., to evaluate whether the miscalibration is at most $\alpha$). (1) include sophisticated theoretical tools to guarantee statistical validity of the MCBoost algorithm. In practice, standard cross-validation splits into calibration and validation sets works well.

In Algorithm 1, we give the pseudocode for the basic variant of MCBoost. See the OSF repository for the technical implementation details and code. MCBoost has also been implemented in an R package that is available on CRAN (6).

Algorithm 1 MCBoost

Given:

\[ p_0 : X \to [0, 1] \]
\[ A : (X \times [-1, 1])^m \to C \]
\[ \alpha > 0 \]
\[ S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \]
\[ V = \{(x_1, y_1), (x_2, y_2), \ldots, (x_v, y_v)\} \]

Returns:

$(C, \alpha)$-multi-calibrated predictor $\hat{p}$

Repeat: $t = 0, 1, 2, \ldots$

\[ S_t \leftarrow \{(x_1, y_1 - p_t(x_1)), \ldots, (x_m, y_m - p_t(x_m))\} \]
\[ c \leftarrow A(S_t) \]
\[ \Delta_c \leftarrow \frac{1}{|V|} \sum_{(x,y) \in V} c(x) \cdot (y - p_t(x)) \]

if $|\Delta_c| > \alpha$ then

\[ p_{t+1}(x) \propto e^{-\Delta_c \cdot c(x)/2} \cdot p_t(x) \]

else

return $\hat{p} = p_t$

end if

---

8 If instead of pre-trained predictor, we start with a trivial predictor $p_0(x) = 1/2$ for all $x \in X$, then MCBoost can equivalently be viewed as a learning algorithm.

9 Formally, $A$ should be a regression algorithm whose notion of “best fit” corresponds to the weak agnostic learning property, as explored in (1). Practically, the exact notion of regression does not have a major effect on the algorithmic performance, but can be tuned to the given setting (2).
Regression algorithm selection. Many techniques, ranging in sophistication, can be used to estimate the propensity score (7). The most popular technique for fitting propensity scores involves logistic regression. In the language of the above analysis, for $d$-dimensional covariate vectors, this corresponds to taking

$$\Sigma = \{ \sigma_w : \mathbf{x} \rightarrow [0, 1] : w \in \mathbb{R}^d \}$$

where

$$\sigma_w(x) = \frac{e^{\langle w, x \rangle}}{1 + e^{\langle w, x \rangle}}$$

For such a propensity score $\sigma_w$, the corresponding propensity ratio $c_w$ is given as follows.

$$c_w(x) = \frac{1 - \sigma_w(x)}{\sigma_w(x)} = \frac{1}{e^{\langle w, x \rangle}}$$

Thus, to obtain performance commensurate with logistic regression propensity scoring, we should theoretically choose the corresponding multi-calibration class to be $C(\Sigma) = \{ e^{-\langle w, x \rangle} : w \in \mathbb{R}^d \}$. Note, however, that when the magnitudes of $\langle w, x \rangle$ are upper bounded, this class may be well approximated by $C = \{ 1 - \langle w, x \rangle : w \in \mathbb{R}^d \}$ (by the first-order Taylor approximation). This observation motivates the use of ridge regression auditing (i.e., linear regression with a penalty for large weights) to capture shifts described by logistic propensity scores.
3. Universal Adaptability in Mortality Estimation

We describe the mortality estimation setup, inspired by the data analysis of (8). We provide details about the data sets used, experimental methods employed, and additional results.

A. Data and Experimental Methods. In this case study, data from the third U.S. National Health and Nutrition Examination Survey (NHANES, 1988-1994) serves as the source, while the U.S. National Health Interview Survey (NHIS, 1994) is used as the target. NHANES is a cross-sectional household survey and a medical examination survey of the civilian, non-institutionalized population of the U.S. with 20,050 observations in the adult sample of the household survey (9). NHIS is a cross-sectional household survey of the same target population, from which we use 19,738 observations from the Department of Health and Human Services Year 2000 Health objectives interview (10). While both surveys may be used to obtain US-representative estimates, we use the unweighted NHANES (NHANES) as the source and the weighted NHIS (NHIS) as the target. The source NHANES differs from the target NHIS due to oversampling of, e.g., adults aged 60 and over and non-Hispanic blacks. Both surveys have been linked to death certificate records from the National Death Index (NDI) (11), resulting in 19,598 observations for NHANES and 19,463 observations for NHIS with mortality follow-up information. On this basis, our objective is to estimate 15-year all-cause mortality rates for adults in the US. We use the same set of covariates (age, sex, ethnicity, marital status, education, family income, region, smoking status, health, BMI) for predicting mortality and sample membership (NHANES vs. NHIS). The full NHANES data is used for for obtaining naive estimates, and training propensity score and predictive models. When applying MCBoost, the NHANES data is split into a training set (75%, 14,698 observations), used for training initial predictive models, and an auditing set (25%, 4,900 observations), used for post-processing with MCBoost, at random.

Inference methods. The following methods are used for estimating 15-year all-cause mortality rates in NHIS.

- (Naive) The unweighted mean is estimated in the NHANES source. This estimate is used naively as a proxy for the mean in the NHIS target.
- (IPSW-Overall) The NHIS mean is estimated by the NHANES mean after reweighting samples by the propensity score. The propensity score is estimated using logistic regression to predict membership between NHANES and NHIS.
- (IPSW-Subgroup) A subgroup-specific IPSW estimate. For each demographic group, a subgroup-specific propensity score is learned using logistic regression to predict membership in NHANES and NHIS; then, the IPSW estimate is calculated within the subgroup.
- (RF-Naive) A random forest (RF) is trained for predicting mortality on the NHANES source (naively, without reweighting or multi-calibration). The estimate is derived by averaging the RF predictions over unlabeled samples from NHIS.
- (RF-Hybrid) A random forest is trained for predicting mortality on the inverse propensity score reweighted NHANES source data. The propensity score is estimated as in IPSW-Overall. The estimate is given by the average prediction on NHIS.
- (RF-MC-Ridge) A random forest is trained for predicting mortality on the training set of the NHANES source data. This naive predictive model is post-processed on the auditing set of the source, using MCBoost using ridge regression for auditing. The estimate is given by the average of the ridge-multi-calibrated RF predictions on NHIS.

All propensity scores are fit using the full NHANES data and the (unlabeled) NHIS data, using standard unpenalized logistic regression. RF-Naive and RF-Hybrid are fit using the best hyper-parameter settings after tuning with 5-fold Cross-Validation in the NHANES data. RF-MC-Ridge is fit after tuning the initial random forest with 5-fold Cross-Validation in the training set of the NHANES data and tuning MCBoost with 5-fold Cross-Validation in the auditing set of the NHANES data. The hyper-parameter settings and functions used for implementing the inference techniques are reported in Table S1. For more detail on the computational implementation, we refer the reader to the OSF repository.

Evaluation. We evaluate the performance of the estimates using the following criteria. Let \( \mu^* \in [0, 1] \) denote the true mortality rate in the NHIS target. For an estimate \( \hat{\mu} \), we measure the (signed) estimation error as the difference between the estimated mortality rate and the true mortality rate in the target (scaled by 100), and the percent error which gives the absolute value of the relative error in estimation.

\[
\text{Estimation Error} = (\hat{\mu} - \mu^*) \times 100
\]

\[
\text{Percent Error} = \left| \frac{\hat{\mu} - \mu^*}{\mu^*} \right| \times 100
\]

We evaluate these performance measures overall and restricted to demographic subpopulations (defined by age, sex, ethnicity, marital status, education and region).
Table S1. Hyper-parameter settings and tuning grids of inference methods. scikit-learn (0.19.1) default settings are used for parameters not listed.

| Method       | Function       | Hyper-parameter     | Value          |
|--------------|----------------|--------------------|----------------|
| Logit/ IPSW  | LogisticRegression | penalty            | none           |
| RF           | RandomForestClassifier | max_features       | sqrt, log2     |
|              |                 | min_samples_leaf   | 5, 10          |
|              |                 | n_estimators       | 500            |
| MC-Ridge     | mcboost via Ridge | max_iter           | 30, 40         |
| MC-Tree      | mcboost via DecisionTreeRegressor | max_iter  | 30, 40         |
|              |                 | max_depth          | 2              |

B. Supplemental Results. Expanding on the results of Table 2, we compare the source NHANES and target NHIS compositions with respect to a larger set of subpopulations and the corresponding expected mortality rates in Table S2. Due to the shift in distributions between NHANES and NHIS, naive inference based on the NHANES sample results in considerable error for almost all subpopulations. We also report the sample composition of the inverse propensity reweighted source sample (IPSW NHANES), indicating that the global propensity score model is largely successful in matching the composition of the source sample to the target, according to the selected marginal categories.

We report the error in estimation with additional inference methods and the larger set of subpopulations in Table S3.

- **(RF-MC-Tree)** A random forest is trained for predicting mortality on the training set of the NHANES source data. This naive predictive model is post-processed on the auditing set of the source, using MCBoost using decision-tree regression for auditing. The estimate is given by the average of the tree-multi-calibrated RF predictions on NHIS.

- **(Logit-Naive)** A logistic regression (logit) is trained for predicting mortality on the NHANES source (naively, without reweighting or multi-calibration). The estimate is derived by averaging the logit predictions over unlabeled samples from NHIS.

- **(Logit-Hybrid)** A logistic regression is trained for predicting mortality on the inverse propensity score reweighted NHANES source data. The propensity score is estimated as in IPSW-Overall. The estimate is given by the average prediction on NHIS.

- **(Logit-MC-Ridge)** A logistic regression is trained for predicting mortality on the training set of the NHANES source data. This naive predictive model is post-processed on the auditing set of the source, using MCBoost using ridge regression for auditing. The estimate is given by the average of the ridge-multi-calibrated logit predictions on NHIS.

- **(Logit-MC-Tree)** A logistic regression is trained for predicting mortality on the training set of the NHANES source data. This naive predictive model is post-processed on the auditing set of the source, using MCBoost using decision-tree regression for auditing. The estimate is given by the average of the tree-multi-calibrated logit predictions on NHIS.

Logit-Naive and Logit-Hybrid are based on standard unpenalized logistic regression. RF-MC-Tree, Logit-MC-Ridge and Logit-MC-Tree are implemented as outlined above; a random forest (logistic regression) is tuned (fit) in the training set of the NHANES data and MCBoost is tuned in the auditing set of the NHANES data. See Table S1 again for hyper-parameter settings and modeling functions.

Among the logit-based approaches, both Logit-Naive and Logit-Hybrid perform comparatively well and tend to improve over RF-Naive and RF-Hybrid. The post-processed logit predictions largely result in similarly small errors, with some further improvements overall and for selected subpopulations. For both base predictors, post-processing with ridge regression (RF-MC-Ridge, Logit-MC-Ridge) results in more accurate inference when compared to tree-based post-processing (RF-MC-Tree, Logit-MC-Tree).‡

‡ We also measured run-time as an indicator of computational feasibility. Based on the initial random forest (4.10 sec.), additional 1.02 sec. are needed for post-processing with RF-MC-Ridge (1.17 sec. for RF-MC-Tree). For the initial logistic regression (0.90 sec.), additional 0.43 sec. are needed with Logit-MC-Ridge (0.46 sec. for Logit-MC-Tree). Run-time for IPSW-Overall was 1.81 sec., and 18.59 sec. for IPSW-Subgroup.
Table S2. Composition of unweighted source (NHANES), weighted source (IPSW NHANES), and target (NHIS), with prevalence (%) of subpopulations in the respective distributions. For NHANES and NHIS, average mortality rates (%) are listed.

| Sample Composition | Average Mortality |
|-------------------|------------------|
|                   | NHANES | NHIS | IPSW | NHANES | NHIS |
| Overall           |        |      |      | 27.67  | 17.57 |
| Male              | 46.75  | 44.29| 47.74| 30.56  | 18.77 |
| Female            | 53.25  | 55.71| 52.26| 25.11  | 16.48 |
| Age 18-24         | 13.87  | 11.41| 13.36| 3.81   | 2.23  |
| Age 25-44         | 36.43  | 40.94| 43.61| 5.70   | 3.86  |
| Age 45-64         | 23.11  | 27.85| 26.62| 22.71  | 17.66 |
| Age 65-69         | 6.34   | 5.25 | 5.10 | 48.61  | 45.52 |
| Age 70-74         | 6.57   | 5.63 | 4.57 | 64.24  | 60.03 |
| Age 75+           | 13.69  | 8.92 | 6.75 | 90.47  | 86.25 |
| White             | 42.56  | 75.21| 75.81| 37.25  | 18.70 |
| Black             | 27.30  | 12.79| 11.19| 23.08  | 18.94 |
| Hispanic          | 28.59  | 8.04 | 9.01 | 18.38  | 10.18 |
| Other             | 1.55   | 3.96 | 3.99 | 15.62  | 8.96  |
| Married           | 58.50  | 62.39| 63.54| 24.85  | 15.47 |
| Wid., Div., Sep.  | 22.55  | 20.23| 17.47| 49.46  | 36.93 |
| Never Married     | 18.95  | 17.38| 18.99| 11.10  | 6.78  |
| Elementary        | 24.15  | 8.07 | 7.63 | 45.18  | 43.97 |
| High School       | 17.25  | 11.80| 11.57| 28.43  | 25.84 |
| High School Grad. | 31.19  | 36.36| 37.07| 21.89  | 17.33 |
| College           | 15.42  | 21.71| 22.46| 18.41  | 11.65 |
| College Grad.     | 6.87   | 12.47| 12.32| 16.87  | 10.37 |
| Post College      | 5.12   | 9.59 | 8.95 | 19.78  | 10.12 |
| Northeast         | 14.74  | 23.04| 20.64| 31.99  | 18.12 |
| Midwest           | 19.08  | 23.61| 25.02| 28.54  | 16.75 |
| South             | 42.56  | 33.84| 32.59| 28.21  | 19.36 |
| West              | 23.61  | 19.51| 21.75| 23.24  | 15.30 |
Table S3. Shift-Aware inferences. Estimation Error in inferred mortality rate reported for each technique on each subpopulation (% error in parentheses). For each subgroup, technique achieving (within $2\times$) best performance in bold.

(a) Inference methods: Naive, IPSW

|                  | Naive Overall | Naive Subgroup |
|------------------|---------------|----------------|
| Overall          | 10.10 (57.5%) | 2.37 (13.5%)   | —              |
| Male             | 11.80 (62.9)  | 2.51 (13.4)    | 0.91 (4.9)     |
| Female           | 8.63 (52.4)   | 2.40 (14.6)    | 3.99 (24.2)    |
| Age 18-24        | 1.57 (70.5)   | 0.00 (0.1)     | -0.39 (17.5)   |
| Age 25-44        | 1.84 (47.6)   | -0.20 (5.2)    | -0.41 (10.6)   |
| Age 45-64        | 5.05 (28.6)   | -0.75 (4.2)    | -0.41 (2.3)    |
| Age 65-69        | 3.09 (6.8)    | -4.23 (9.3)    | -5.23 (11.5)   |
| Age 70-74        | 4.21 (7.0)    | -1.36 (2.3)    | 0.47 (0.8)     |
| Age 75+          | 4.22 (4.9)    | 3.53 (4.1)     | 2.85 (3.3)     |
| White            | 18.55 (99.2)  | 3.53 (18.9)    | 0.75 (4.0)     |
| Black            | 4.14 (21.9)   | -4.00 (21.1)   | -0.48 (2.5)    |
| Hispanic         | 8.20 (80.5)   | 1.73 (17.0)    | 0.48 (4.7)     |
| Other            | 6.66 (74.4)   | -0.02 (0.2)    | -3.54 (39.5)   |
| Married          | 9.17 (59.3)   | 1.19 (7.7)     | 0.76 (4.9)     |
| Wid., Div., Sep. | 12.53 (33.9)  | 3.49 (9.5)     | 5.75 (15.6)    |
| Never Married    | 4.32 (63.8)   | 1.10 (16.3)    | 0.87 (12.8)    |
| Elementary       | 1.21 (2.7)    | 2.68 (6.1)     | 4.01 (9.1)     |
| High School      | 2.59 (10.0)   | 1.60 (6.2)     | 0.74 (2.9)     |
| High School Grad.| 4.56 (26.3)   | 1.65 (9.5)     | 1.78 (10.3)    |
| College          | 6.76 (58.1)   | 3.52 (30.3)    | 1.44 (12.4)    |
| College Grad.    | 6.49 (62.6)   | 1.21 (11.7)    | -0.41 (4.0)    |
| Post College     | 9.66 (95.4)   | 3.42 (33.8)    | 1.99 (19.7)    |
| Northeast        | 13.87 (76.6)  | 4.00 (22.1)    | 2.75 (15.2)    |
| Midwest          | 11.80 (70.4)  | 2.73 (16.3)    | 2.41 (14.4)    |
| South            | 8.85 (45.7)   | 0.42 (2.2)     | 1.64 (8.3)     |
| West             | 7.94 (51.9)   | 2.89 (18.9)    | 0.51 (3.3)     |
(b) Inference methods: RF, RF-MCBoost, Logit, Logit-MCBoost

|                  | RF Hybrid | RF Naive | RF MC-Ridge | RF MC-Tree | Logit Hybrid | Logit Naive | Logit MC-Ridge | Logit MC-Tree |
|------------------|-----------|----------|-------------|------------|--------------|-------------|----------------|---------------|
| Overall          | 0.35 (2.0%) | 1.11 (6.3%) | 0.52 (3.0%) | 1.13 (6.4%) | 0.11 (0.6%) | 0.48 (2.7%) | 0.16 (0.9%) | 0.39 (2.2%)  |
| Male             | -1.34 (7.1) | -0.34 (1.8) | 0.11 (0.6) | 0.07 (0.4) | -0.77 (4.1) | 0.24 (1.3) | 0.12 (0.6) | 0.25 (1.3)   |
| Female           | 1.89 (11.5) | 2.43 (14.8) | 0.90 (5.4) | 2.10 (12.7) | 0.91 (5.5) | 0.70 (4.3) | 0.20 (1.2) | 0.52 (3.2)   |
| Age 18-24        | 5.18 (232.1) | 6.03 (270.2) | 1.76 (79.0) | 6.26 (280.3) | -0.14 (6.3) | 0.79 (35.4) | 0.58 (26.0) | 0.97 (43.5)  |
| Age 25-44        | 0.29 (7.6) | 0.82 (21.2) | 0.66 (17.2) | 0.75 (19.4) | -0.18 (4.6) | 0.50 (12.9) | 0.41 (10.5) | 0.39 (10.1)  |
| Age 45-64        | 0.04 (0.2) | 0.86 (4.8) | -0.29 (1.6) | 0.20 (1.1) | -0.03 (0.2) | 0.48 (2.7) | -0.42 (2.4) | 0.56 (3.2)   |
| Age 65-69        | -5.40 (11.9) | -3.52 (7.7) | -1.99 (4.4) | -1.91 (4.2) | -2.47 (5.4) | -2.84 (6.2) | -3.27 (7.2) | -3.17 (7.0)  |
| Age 70-74        | -4.07 (6.8) | -3.02 (5.0) | 0.61 (1.0) | -1.67 (2.8) | 0.53 (0.9) | -0.38 (0.6) | 0.42 (0.7) | -0.09 (0.2)  |
| Age 75+          | -0.25 (0.3) | 0.51 (0.6) | 2.19 (2.5) | 1.31 (1.5) | 4.66 (5.4) | 2.87 (3.3) | 2.39 (2.8) | 1.58 (1.8)   |
| White            | 0.19 (1.0) | 1.03 (5.5) | 0.69 (3.7) | 1.22 (6.5) | 0.17 (0.9) | 0.75 (4.0) | 0.45 (2.4) | 0.75 (4.0)   |
| Black            | -1.30 (6.8) | -0.66 (3.5) | -0.52 (2.7) | -0.92 (4.8) | -0.59 (3.1) | -0.93 (4.9) | -1.03 (5.4) | -1.50 (7.9)  |
| Hispanic         | 2.84 (27.9) | 2.91 (28.6) | 1.55 (15.2) | 2.41 (23.7) | 1.18 (11.6) | 0.89 (8.8) | 0.63 (6.2) | 0.58 (5.7)   |
| Other            | 2.44 (27.3) | 3.52 (39.3) | -2.06 (23.0) | 2.29 (25.5) | -1.60 (17.9) | -1.67 (18.7) | -3.04 (34.0) | -1.60 (17.9) |
| Married          | 0.23 (1.5) | 0.94 (6.1) | 0.40 (2.6) | 1.06 (6.8) | -0.13 (0.9) | 0.34 (2.2) | 0.10 (0.7) | 0.39 (2.5)   |
| Wid., Div., Sep. | 0.44 (1.2) | 1.42 (3.8) | 0.82 (2.2) | 0.85 (2.3) | 1.25 (3.4) | 0.89 (2.4) | 0.36 (1.0) | 0.04 (0.1)   |
| Never Married    | 0.67 (9.9) | 1.40 (20.6) | 0.66 (9.8) | 1.62 (24.0) | -0.14 (2.0) | 0.59 (8.7) | 0.16 (2.4) | 0.71 (10.5)  |
| Elementary       | -0.96 (2.2) | -0.24 (0.5) | 1.29 (2.9) | 0.57 (1.3) | 2.33 (5.3) | 1.29 (2.9) | 1.15 (2.6) | 0.21 (0.5)   |
| High School      | -0.38 (1.5) | 0.31 (1.2) | -0.13 (0.5) | 0.14 (0.5) | -0.92 (3.6) | -0.39 (1.5) | -0.65 (2.5) | -0.91 (3.5)  |
| High School Grad. | 0.31 (1.8) | 1.00 (5.6) | -0.16 (0.9) | 0.91 (5.3) | 0.04 (0.2) | 0.47 (2.7) | -0.27 (1.6) | 0.75 (4.3)   |
| College          | 0.99 (8.5) | 1.77 (15.2) | 1.43 (12.3) | 1.73 (14.9) | 0.51 (4.4) | 0.90 (7.7) | 0.83 (7.1) | 0.72 (6.1)   |
| College Grad.    | -0.19 (1.8) | 0.73 (7.0) | 0.07 (0.7) | 0.88 (8.4) | -1.17 (11.3) | -0.45 (4.4) | -0.41 (4.0) | -0.94 (3.3)  |
| Post College     | 1.73 (17.1) | 2.61 (25.8) | 1.90 (18.8) | 2.62 (25.9) | 0.56 (5.5) | 1.21 (11.9) | 1.25 (12.3) | 0.93 (9.2)   |
| Northeast        | 0.48 (2.6) | 1.16 (6.4) | 1.50 (8.3) | 1.27 (7.0) | 0.76 (4.2) | 0.85 (4.7) | 1.00 (5.5) | 0.57 (3.2)   |
| Midwest          | 0.47 (2.8) | 1.39 (8.3) | 0.43 (2.6) | 1.44 (8.6) | -0.12 (0.7) | 0.54 (3.2) | 0.18 (1.1) | 0.45 (2.7)   |
| South            | -0.21 (1.1) | 0.46 (2.4) | -0.25 (1.3) | 0.44 (2.3) | -0.44 (2.3) | -0.22 (1.2) | -0.61 (3.1) | -0.34 (1.8)  |
| West             | 0.93 (6.1) | 1.71 (11.2) | 0.85 (5.6) | 1.67 (10.9) | 0.57 (3.7) | 1.12 (7.3) | 0.48 (3.1) | 1.24 (8.1)   |
4. Universal Adaptability under Semi-Synthetic Shift

We describe the semi-synthetic experimental setup on estimating voting rates that allows us to measure performance under various degrees and modes of shift. We provide details about the original data, the synthetic-shift generation, experimental methods employed, and additional results.

A. Data and Experimental Methods. We scaffold this experiment around data collected by the Pew Research Center (12). The data consist of a “non-probability” sample, denoted \( S_{np} \), based on 31,319 online opt-in panel interviews (conducted in 2016 by three vendors) and a “reference population”, \( S_p \), with 20,000 observations that combines information from multiple high quality surveys (e.g., the American Community Survey and the Current Population Survey). Our outcome of interest \( y \) is a binary indicator for voting in the 2014 midterm election (1 = “voted”, 0 = “not voted”). We denote \( z \) a sample membership indicator with \( z = 1 \) for a sample from \( S_{np} \) and \( z = 0 \) for a sample from \( S_p \). We use the same vector \( x \) of covariates (age, sex, education, ethnicity, census division) for predicting \( y \) and \( z \).

Rather than using \( S_{np} \) and \( S_p \) as source and target sources directly, we generate new source and target data sets synthetically, based on the original sets. To evaluate varying degrees and modes of shift that mimic realistic shifts, we model the synthetic shifts based on the relationship between \( S_{np} \) and \( S_p \).

**Synthetic Shift.** To begin, we estimate the propensity score \( \sigma(x) \) between \( S_{np} \) and \( S_p \), using unlabeled samples. The fitted propensity score between \( S_{np} \) and \( S_p \) serves as the basis for the synthetic shifts. To model different modes of shift, we use three different regression techniques to fit \( \sigma \).

- *(Logit-linear)* logistic regression (with linear terms only)
- *(Logit-interaction)* logistic regression with linear terms and pairwise interactions (logistic over degree-2 monomials)
- *(Tree)* decision tree regression

To simulate shift of varying intensity, we amplify relation captured by the propensity score \( \sigma \). For each propensity model (Logit-linear, Logit-interaction, Tree), we generate synthetic data of various shift intensity \( q \in \{0, 0.5, \ldots, 8\} \), such that

\[
\frac{e^{(q)}(x)}{1 - e^{(q)}(x)} = \left( \frac{\sigma(x)}{1 - \sigma(x)} \right)^q
\]

Specifically, we use sampling weights for the source \( w_s \) and target \( w_t \) defined as

\[
w_s^{(q)}(x) = \left( \frac{\sigma(x)}{1 - \sigma(x)} \right)^{q/2}
\]

\[
w_t^{(q)}(x) = \left( \frac{\sigma(x)}{1 - \sigma(x)} \right)^{-q/2}
\]

Given the propensity weights, we generate synthetic source and target data sets by sampling each set from the nonprobability Pew data \( S_{np} \). In this way, we model various scenarios where the density underlying covariates \( x \) shift, but the overall relationship between the outcome \( y \) given \( x \) remains the same under shift. For each shift, we sample a fixed-sized source samples of 7,830 observations, then sample without replacement a fixed-size target of 4,698 observations. The source sample is split, where 70% is used for training and propensity scoring, and 30% is used as the auditing set for MCBoost.

**Inference Methods.** The following methods are used for estimating voting rates under each of the generated shifts.

- *(Naive)* Unweighted source mean used as a proxy for target mean.
- *(IPSW)* Propensity score is estimated between source and target using logistic regression. Inverse propensity score weighted mean used to estimate target mean.
- *(IPSW-Trimmed)* Propensity score is estimated between source and target using logistic regression, but to reduce variance under extreme shift, weights are trimmed at the 0.25% and 99.75% quantile of weights. IPSW weighted mean estimated with trimmed propensity scores.
- *(RF-Naive)* Random forest (RF) trained on source distribution. Target mean estimated using the average RF prediction over unlabeled target samples.
- *(RF-Hybrid)* Random forest trained on the inverse propensity score reweighted source distribution. The propensity score is estimated as in IPSW. Target mean estimated using the average prediction over unlabeled target samples.
- *(RF-MC-Ridge)* Naive RF post-processed with MCBoost using ridge regression. Target mean estimated using the average ridge-multi-calibrated RF prediction over unlabeled target samples.
- *(RF-MC-Tree)* Naive RF post-processed with MCBoost using decision-tree regression. Target mean estimated using the average tree-multi-calibrated RF prediction over unlabeled target samples.
All propensity scores are fit using the training set of the source data and the (unlabeled) target data, using unpenalized logistic regression. All random forests are fit using the best hyper-parameter settings after tuning with 5-fold Cross-Validation in the training set of the source data. The MCBoost hyper-parameters are tuned with 5-fold Cross-Validation in the auditing set of the source data. We use the same hyper-parameter settings as outlined in Table S1, with the only deviation being a more restrictive set of grid values for MCBoost’s max_iter (i.e., [5, 10, 25]).

Evaluation. We evaluate the performance of the estimates using the following criteria. Let \( \mu^*_t \in [0, 1] \) denote the true voting rate in the target. For an estimate \( \hat{\mu} \), we measure the relative error.

\[
\text{Relative Error} = \left( \frac{\hat{\mu} - \mu^*_t}{\mu^*_t} \right) \times 100
\]

For each mode of propensity score and shift intensity, sampling from the non-probability sample \( \mathcal{D}_{np} \) (in step 2.) is repeated 100 times, resulting in \( 100 \times 3 \times 17 \) source and target samples that are used for model training and evaluation.

B. Supplemental Results. In addition to the results reported in Figure 2, we report on additional inference techniques under the same experimental conditions. Whereas the base predictor reported in Figure 2 is given by a random forest, here, we explore using logistic regression as the base classifier.

- **(Logit-Naive)** Logistic regression (logit) trained on source distribution. Target mean estimated using the average logit prediction over unlabeled target samples.
- **(Logit-Hybrid)** Logistic regression trained on the inverse propensity score reweighted source distribution. The propensity score is estimated as in IPSW. Target mean estimated using the average prediction over unlabeled target samples.
- **(Logit-MC-Ridge)** Naive logit post-processed with MCBoost using ridge regression. Target mean estimated using the average ridge-multi-calibrated logit prediction over unlabeled target samples.
- **(Logit-MC-Tree)** Naive logit post-processed with MCBoost using decision-tree regression. Target mean estimated using the average tree-multi-calibrated logit prediction over unlabeled target samples.

All logistic regression models are fit using the training set of the source data using standard unpenalized logistic regression. MCBoost hyper-parameters are tuned in the auditing set of the source data as outlined above.

The results are presented in Figure S1. We observe that when the shift is also specified by a logistic-based propensity score, the naive logistic approach actually generalizes as well as the prediction functions that are post-processed for multi-calibration. In this case, the base logistic model may be well-calibrated on subpopulations that arise from logistic-based shifts. When the shift is given by a tree-based model, the base logistic approach incurs more significant error. Overall, the results demonstrate the importance of choosing the underlying predictive model and multi-calibration class to fit the shift well.

Robust prediction. We also validate the theoretical intuition explored in §C that multi-calibrated prediction functions give robust generalization even for classification tasks. Rather than simply predicting the average voting rate, we test the performance of various techniques at classifying, given an individual’s covariates, whether they voted in the midterm election. We perform analogous experiments under the same semi-synthetic shift model, reporting how the classification accuracy degrades with increased shift.

We evaluate the prediction performance of baseline random forests (RF-Naive) and baseline logistic regressions (Logit-Naive) that were trained with the source sample, and multi-calibrated prediction functions after post-processing both baseline models with ridge regression (RF-MC-Ridge, Logit-MC-Ridge) and decision trees (RF-MC-Tree, Logit-MC-Tree). In addition, we compare to hybrid approaches that use labeled data from the source and unlabeled data from the target to train inverse propensity weighted random forests (with trimmed and untrimmed scores; RF-Hybrid, RF-Hybrid-trimmed) and inverse propensity weighted logistic regressions (with trimmed and untrimmed scores; Logit-Hybrid, Logit-Hybrid-trimmed). For all real-valued predictors \( p : \mathcal{X} \rightarrow [0, 1] \), we round to binary predictions in \( \{0, 1\} \) using a fixed threshold at 0.5. We summarize the findings in Figure S2, where we report the classification accuracy. For a test sample \( T \), the accuracy is defined as follows.

\[
\text{Accuracy} = \frac{1}{|T|} \sum_{i \in T} 1(\hat{y}_i = y_i)
\]

As with the statistical inference tasks, multi-calibrated prediction functions generalize well at classification tasks under distributional shift. In particular, the multi-calibrated models perform comparably with the hybrid models that are trained on a propensity weighted data set that explicitly models the shift.
Fig. S1. Relative Error (%) in inferred voting rates under synthetic shift (varying intensity $q$). Shifts given by three modes of propensity score: logistic with linear terms, logistic with pairwise interactions, and a decision tree. Error of (Naive, IPSW and Logit-based) inferences are plotted against unbiased baseline (Relative Error = 0).
Fig. S2. Classification accuracy in target distribution for voting prediction under synthetic shift. Shifts given by three propensity score models: logistic with linear terms, logistic with pairwise interactions, and a decision tree, with exponential amplification $q$. 
References

1. Ú Hébert-Johnson, MP Kim, O Reingold, G Rothblum, Multicalibration: Calibration for the (computationally-identifiable) masses in *International Conference on Machine Learning*. pp. 1939–1948 (2018).

2. MP Kim, A Ghorbani, J Zou, Multiaccuracy: Black-box post-processing for fairness in classification in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. pp. 247–254 (2019).

3. C Dwork, MP Kim, O Reingold, GN Rothblum, G Yona, Learning from outcomes: Evidence-based rankings in *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. (IEEE), pp. 106–125 (2019).

4. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2020).

5. F Pedregosa, et al., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830 (2011).

6. F Pfisterer, et al., mcboost: Multi-Calibration Boosting for R. *J. Open Source Softw.*, **6**, 3453 (2021).

7. C Kern, Y Li, L Wang, Boosted kernel weighting—using statistical learning to improve inference from nonprobability samples (2020).

8. L Wang, BI Graubard, HA Katki, Y Li, Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts (2020).

9. Department of Health and Human Services, Third national health and nutrition examination survey, 1988-1994, NHANES III household adult file (1996).

10. National Center for Health Statistics, Public use data tape documentation, part I, national health interview survey, 1994 (1995).

11. National Center for Health Statistics, Office of analysis and epidemiology, public-use linked mortality file, 2015 (2013).

12. A Mercer, A Lau, C Kennedy, *For Weighting Online Opt-In Samples, What Matters Most?* (Pew Research Center), (2018).