Universal adaptability: Target-independent inference that competes with propensity scoring

Michael P. Kim[a,b,1], Christoph Kern[c,1], Shafi Goldwasser[a,d,2], Frauke Kreuter[e,f], and Omer Reingold[g]

[a]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720; [b]Miller Institute for Basic Research in Science, Berkeley, CA 94720; [c]School of Social Sciences, University of Mannheim, 68159 Mannheim, Germany; [d]Simons Institute for the Theory of Computation, Berkeley, CA 94720; [e]Joint Program in Survey Methodology, University of Maryland, College Park, MD 20742; [f]Department of Statistics, Ludwig-Maximilians-Universität München, 80539 München, Germany; and [g]Department of Computer Science, Stanford University, Stanford, CA 94305

Contributed by Shafi Goldwasser; received April 29, 2021; accepted December 2, 2021; reviewed by Toniann Pitassi and Pragya Sur

The gold-standard approaches for gleaning statistically valid conclusions from data involve random sampling from the population. Collecting properly randomized data, however, can be challenging, so modern statistical methods, including propensity score reweighting, aim to enable valid inferences when random sampling is not feasible. We put forth an approach for making inferences based on available data from a source population that may differ in composition in unknown ways from an eventual target population. Whereas propensity scoring requires a separate estimation procedure for each different target population, we show how to build a single estimator, based on source data alone, that allows for efficient and accurate estimates on any downstream target data. We demonstrate, theoretically and empirically, that our target-independent approach to inference, which we dub "universal adaptability," is competitive with target-specific approaches that rely on propensity scoring. Our approach builds on a surprising connection between the problem of inferences in unspecified target populations and the multicalibration problem, studied in the burgeoning field of algorithmic fairness. We show how the multicalibration framework can be employed to yield valid inferences from a single source population across a diverse set of target populations.

statistical validity | propensity scoring | algorithmic fairness

Across the world, there is a growing push to leverage data about populations to inform effective data-driven policy. In the United States, the Evidence-Based Policymaking Act of 2018 and the US Federal Data Strategy (1) established government-wide reforms for making data accessible and useful for decision-making; globally, in the Post-2015 Development Agenda, the High Level Panel articulated the need for a "data revolution" to promote evidence-based decisions and to strengthen accountability (2). Answering key questions like "Will this policy work in our context?" or "How will this disease variant spread in our country?" can be very challenging, because the composition of populations varies considerably across regions, as is acutely apparent during the COVID-19 crisis. To make progress, systematic methodology for collecting and processing data is needed.

The gold-standard approaches for gleaning insights from data involve proper random sampling. Classic experimental methods estimate statistics (e.g., population averages, or causal effects) by randomly sampling individuals to participate in trial groups (interviewed, treatment/control). The statistical validity of the conclusions—and the methods' effectiveness as part of a policy platform—depends crucially on the quality of randomness. Collecting data with proper randomization, however, is often difficult and costly.

For instance, to understand medical trends across the United States, traditional statistical methods might require analysts to coordinate with hospitals across the country to collect random samples throughout the general population. Comparatively, it would be cheap and easy to collect data from a single hospital, but samples from a given hospital may not be representative of the US population at large. As such, a huge body of modern quantitative statistical research focuses on methods that, given access to observational data from a "source" population, enable valid inferences for "target" populations, even when proper randomization over the target is not possible (3, 4). Today, developing such robust approaches to statistical inference is a challenging and active research area, spanning areas including domain adaptation (5), "quasi-experimental" methods in causal inference (6–8), and prediction and inference under distributional shift (9, 10), with particular emphasis in public health studies (11).

Within this research, a major paradigm for obtaining valid statistical inferences involves propensity score reweighting (3, 12). The propensity score between a source and target population relates the likelihood of observing data under the two populations. As such, the propensity score can be used to reweight samples of data from an observational source to "look like" a sample from a randomly sampled target population. Given access to labeled data from a source population—where we observe a variable of interest Y associated with covariates X—and unlabeled data from a target population—where we observe only the covariates X—we can use the propensity score to obtain valid statistical inferences about Y within the target. Performing inference via propensity score reweighting involves two steps: first, estimating

Significance

We revisit the problem of ensuring statistically valid inferences across diverse target populations from a single source of training data. Our approach builds a surprising technical connection between the inference problem and a technique developed for algorithmic fairness, called "multicalibration." We derive a correspondence between the fairness goal, to protect subpopulations from miscalibrated predictions, and the statistical goal, to ensure unbiased estimates on target populations. We derive a single-source estimator that provides inferences in any downstream target population, whose performance is comparable to the popular target-specific approach of propensity score reweighting. Our approach can extend the benefits of evidence-based decision-making to communities that do not have the resources to collect high-quality data on their own.

Author contributions: M.P.K., C.K., S.G., F.K., and O.R. designed research; M.P.K. and C.K. performed research; M.P.K. and C.K. analyzed data; and M.P.K., C.K., S.G., F.K., and O.R. wrote the paper.

Reviewers: T.P., University of Toronto; and P.S., Harvard University.

The authors declare no competing interest.

[1]M.P.K. and C.K. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: shafi.goldwasser@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2108097119/-/DCSupplemental.

Published January 19, 2022.

the propensity score using unlabeled samples of data from the source and the target; second, evaluating the statistic of interest using labeled samples from the source population that have been reweighted by the propensity score.

In this way, propensity score reweighting provides a pragmatic method for gleaning insights about populations of interest (targets) from plentiful but nonrandomized observational data (sources). The paradigm has been successfully applied across a vast array of scientific settings, including estimating the effects of training programs on later earnings (13), the relationship between postmenopausal hormone therapy and coronary heart disease (14), and the effectiveness of HIV therapy (15). In each of these examples, the propensity reweighted estimates demonstrated differences in efficacy across populations significant enough to change policy for treatment (16).

Despite the remarkable success of propensity score reweighting for performing inferences in diverse applications, the approach has some critical limitations. Crucially, to obtain accurate inferences in a given target population, we must first estimate the propensity score from the source to the target. In this way, propensity score reweighting is best suited for making inferences in a single fixed target population. Often, however, it may be useful to make inferences in many target populations. Continuing the earlier example, rather than making a broad-strokes medical inference about the entire US population, hospitals across the country may benefit from findings tailored to their region and patient demographic. In such a setting, analysts may wish to transfer insights from data collected at a single source hospital to many target hospitals across the country.

When performing inferences in multiple possibly evolving target populations, the need to estimate target-specific propensity scores presents challenges. In particular, in addition to the labeled source data, for every new target population, we must obtain a random sample of unlabeled data from the target and perform regression to estimate the propensity score from the source population. This framework for target-specific inferences is demanding in terms of data and computation, and may be prohibitive for making inferences in resource-limited target populations.

An Approach for Inference across Targets

In settings where we want to perform inferences on many downstream target populations, ideally, we would eliminate the need to estimate a separate propensity score model for each target. In this work, we explore the possibility of such an ideal framework: Rather than estimating target-specific reweighting functions, we aim to learn a single estimator that automatically adapts to shifts from source to target populations. In particular, we study how to use labeled source data to build a single prediction function $\tilde{p}$ that, given unlabeled samples from any target population t, allows for efficient estimation of the statistic of interest over t. We formalize the goal of our approach through a criterion for prediction functions, which we call "universal adaptability." Informally, for a given statistic, we say a prediction function $\tilde{p}$ is universally adaptable from a source s, if, for any target population t, the error in estimation using $\tilde{p}$ is comparable to the error obtained via target-specific propensity score reweighting from s to t.

We make progress on the problem of universal adaptability by demonstrating a surprising connection to a problem studied in the burgeoning field of algorithmic fairness (17). One serious fairness concern with using algorithmic predictions is that they may be miscalibrated—or systematically biased—on important but historically marginalized groups of people. A recent study suggests, for instance, that predictive algorithms used within the health care system can exhibit miscalibration across racial groups, contributing to significant disparity in health outcomes between Black and White patients (18). Introduced recently in ref. 19, multicalibration provides a formal and constructive framework for mitigating such forms of systematic bias.

In this work, we reinterpret the fairness guarantees of multicalibration to obtain universal adaptability: We derive a direct correspondence between protecting a vast collection of subpopulations from miscalibration and ensuring unbiased statistical estimates over a vast collection of target populations. Technically, we show how multicalibration implicitly anticipates propensity shifts to potential target populations. In turn, we leverage multicalibration to obtain estimates in the target that achieve comparable accuracy with estimators that use propensity scoring. Whereas modeling the propensity score explicitly requires performing regression over source and target data for every new target, our approach for universal adaptability allows the analyst to learn a single estimator (based on a multicalibrated prediction function) that can be evaluated efficiently on any downstream target population.

Formal Preliminaries

Let $\mathcal{X}$ denote the space of covariates representing individual records and $\mathcal{Y}$ denote the outcome range, for example, $\mathcal{Y} = \{0, 1\}$ for binary outcomes. For each individual x, the associated y may represent a variable of interest, for instance, a health outcome after treatment. Let $\mathcal{Z} = \{\mathrm{s}, \mathrm{t}\}$ denote sampling in the source or target distribution. Formally, we assume a joint distribution over X, Y, Z triples, for covariates X, outcome Y, and source vs. target indicator Z. We use $\mathcal{D}_\mathrm{s}$ and $\mathcal{D}_\mathrm{t}$ to denote the joint distributions over X, Y pairs, conditioned on $Z = \mathrm{s}$ and $Z = \mathrm{t}$, respectively. For convenience, we use $\mathcal{U}_\mathrm{s}$ and $\mathcal{U}_\mathrm{t}$ to denote the distribution over unlabeled samples (i.e., marginal distribution over covariates X), conditioned on source or target. Importantly, we assume that the relationship between X and Y does not change from $\mathcal{D}_\mathrm{s}$ to $\mathcal{D}_\mathrm{t}$; formally, we assume the joint law factorizes as $\mathbf{Pr}[\, X, Y, Z \,] = \mathbf{Pr}[\, X \,] \cdot \mathbf{Pr}[\, Y \mid X \,] \cdot \mathbf{Pr}[\, Z \mid X \,]$, sometimes called "ignorability."

Inference Task

We aim for accurate statistical inferences over a target distribution. For concreteness, we focus on the task of estimating the average value of the outcome of interest in the target population, denoted as

$$\mu^*_\mathrm{t} = \mathop{\mathbf{E}}_{(X,Y)\approx\mathcal{D}_\mathrm{t}}[\, Y \,].$$

For any estimate of the statistic, $\tilde{\mu}$, we define the estimation error on the target er_t to be the absolute deviation from the true statistic.

$$\mathrm{er}_\mathrm{t}(\tilde{\mu}) = |\, \tilde{\mu} - \mu^*_\mathrm{t} \,|.$$

Given direct access to labeled samples from $\mathcal{D}_\mathrm{t}$, the empirical estimator gives a good approximation to μ^*_t. When our access to labeled samples from the target distribution is limited, more-sophisticated techniques are necessary to obtain unbiased estimates.

Propensity Score Reweighting

For given source and target, the propensity score allows us to relate the odds, given a set of covariates $X = x$, of being sampled from $\mathcal{D}_\mathrm{s}$ and $\mathcal{D}_\mathrm{t}$ (20, 21). Specifically, for a given source $\mathcal{D}_\mathrm{s}$ and target $\mathcal{D}_\mathrm{t}$, the propensity score $e_\mathrm{st} : \mathcal{X} \to [0, 1]$ is defined as the following probability:

$$e_\mathrm{st}(x) = \mathbf{Pr}[\, Z = \mathrm{s} \mid X = x \,].$$

Correspondingly, $1 - e_{\mathrm{st}}(x) = \mathbf{Pr}[\, Z = \mathrm{t} \mid X = x \,]$. Given the propensity score, we can obtain an unbiased estimate of the expectation of Y in the target by reweighting $\mathcal{D}_{\mathrm{s}}$ accordingly.*

$$\mathop{\mathbf{E}}_{(X,Y)\approx\mathcal{D}_{\mathrm{s}}}\left[\frac{1 - e_{\mathrm{st}}(X)}{e_{\mathrm{st}}(X)} \cdot Y\right] = \mu_{\mathrm{t}}^{*}.$$

The approach of inverse propensity score weighting (IPSW) follows from this observation. First, the analyst chooses a class of propensity scoring functions Σ. With Σ fixed, the analyst uses unlabeled samples from $\mathcal{U}_{\mathrm{s}}$ and $\mathcal{U}_{\mathrm{t}}$ to find the best fit approximation $\sigma \in \Sigma$ of the propensity score. Then, to obtain an inference of the target mean, the analyst reweights labeled samples from $\mathcal{D}_{\mathrm{s}}$ according to the (best-fit) propensity odds $1 - \sigma(X)/\sigma(X)$.

$$\mu_{\mathrm{t}}^{\mathrm{ps}}(\sigma) = \mathop{\mathbf{E}}_{(X,Y)\approx\mathcal{D}_{\mathrm{s}}}\left[\frac{1 - \sigma(X)}{\sigma(X)} \cdot Y\right].$$

Many techniques, varying in sophistication, can be used to estimate the propensity score (22). Concretely, logistic regression is the most commonly used method for fitting the propensity score; in this case, Σ is taken to be the class of linear functions passed through the logistic activation.

For any method of fitting a best-fit propensity score $\sigma \in \Sigma$ to the true propensity score e_{st}, we define the misspecification error, denoted $\Delta_{\mathrm{st}}(\sigma)$, as the following expected distance:

$$\Delta_{\mathrm{st}}(\sigma) = \mathop{\mathbf{E}}_{\mathcal{D}_{\mathrm{s}}}[\, d(\sigma(X), e_{\mathrm{st}}(X))\,], \qquad [1]$$

where $d(p, q) = |p/1 - p - q/1 - q|$ is the absolute difference in the corresponding propensity ratios. Intuitively, if Σ fits the shift from $\mathcal{D}_{\mathrm{s}}$ to $\mathcal{D}_{\mathrm{t}}$ accurately, then the misspecification error $\Delta_{\mathrm{st}}(\sigma)$ will be small for some $\sigma \in \Sigma$. Importantly, if Σ correctly specifies the true propensity score (i.e., $e_{\mathrm{st}} \in \Sigma$), then the propensity-based estimator is unbiased.

Imputation and Universal Adaptability

An alternative strategy for inference, known as imputation, involves learning a prediction function that estimates the relationship of the variable of interest given the covariates of an individual, then averaging this prediction over the target distribution. Specifically, given a prediction function $\tilde{p}$ and an unlabeled sample from the target distribution, we can use $\tilde{p}(X)$ as a surrogate estimate for Y, and estimate $\mu_{\mathrm{t}}(\tilde{p})$ as follows:

$$\mu_{\mathrm{t}}(\tilde{p}) = \mathop{\mathbf{E}}_{X\approx\mathcal{U}_{\mathrm{t}}}[\, \tilde{p}(X)\,].$$

Our goal will be to learn $\tilde{p}$ using samples from the source distribution $\mathcal{D}_{\mathrm{s}}$ in order to guarantee the imputation inference is competitive with the propensity score estimate, regardless of the eventual target population. We formalize this goal through a notion we call "universal adaptability."

Definition: (Universal Adaptability). For a source distribution $\mathcal{D}_{\mathrm{s}}$ and a class of propensity scores Σ, a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ is (Σ, β) universally adaptable if for any target distribution $\mathcal{D}_{\mathrm{t}}$,

$$\mathrm{er}_{\mathrm{t}}(\mu_{\mathrm{t}}(\tilde{p})) \le \mathrm{er}_{\mathrm{t}}(\mu_{\mathrm{t}}^{\mathrm{ps}}(\sigma_{\mathrm{st}}^{*})) + \beta.$$

Table 1 summarizes the differences in data and estimation requirements under propensity scoring and universal adaptability inferences. In general, prediction accuracy on the source population $\mathcal{D}_{\mathrm{s}}$ will not imply that the downstream estimates $\mu_{\mathrm{t}}(\tilde{p})$ will be universally adaptable. Our goal is to characterize sufficient conditions on $\tilde{p}$ such that universal adaptability is guaranteed.

Table 1. Comparing propensity scoring and universal adaptability

Method	Required estimation	Inference procedure
Propensity scoring	Estimate target-specific propensity score e_t using unlabeled source/target data	Evaluate statistic of Y using labeled source data reweighted by e_t
Universal adaptability	Estimate target-independent prediction function $\tilde{p}$ using labeled source data	Evaluate statistic of $\tilde{p}(X)$ using unlabeled target data

Required estimation: Propensity scoring (PS) requires estimating a target-specific propensity score using unlabeled samples from $\mathcal{U}_{\mathrm{s}}$ and $\mathcal{U}_{\mathrm{t}}$; universal adaptability (UA) estimates a multicalibrated prediction function $\tilde{p}$ using labeled samples from the source $\mathcal{D}_{\mathrm{s}}$. Inference procedure: For each method, the inferences consist of empirical expectations of different variables over different distribution. To obtain the PS estimates, labeled samples from $\mathcal{D}_{\mathrm{s}}$ are reweighted by the propensity score and then the variable of interest is averaged; to obtain the UA estimates, the prediction $\tilde{p}(X)$ is used in place of the variable of interest, and is averaged across unlabeled samples from $\mathcal{U}_{\mathrm{t}}$. Importantly, universal adaptability via multicalibration requires access to $\mathcal{U}_{\mathrm{t}}$ only at inference time, so a given prediction function $\tilde{p}$ can imply efficient inferences simultaneously for many target populations.

Multicalibrated Predictions

Multicalibration is a property of prediction functions, initially studied in the context of algorithmic fairness. Intuitively, multicalibrated predictions mitigate subpopulation bias, by ensuring that the prediction function appears well calibrated, not simply overall but even when we restrict our attention to structured subpopulations of interest. In the original formulation of ref. 19, the subpopulations of interest are defined in terms of a class of Boolean functions $\mathcal{C}$. Here, we work with a generalization of where we parameterize multicalibration in terms of a collection of real-valued functions $\mathcal{C}$.†

Definition: (Multicalibration). For a given distribution $\mathcal{D}$ and class of functions $\mathcal{C}$, a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ *is* $(\mathcal{C}, \alpha)$ multicalibrated if

$$\left|\mathop{\mathbf{E}}_{(X,Y)\approx\mathcal{D}}[\, c(X) \cdot (Y - \tilde{p}(X))\,]\right| \le \alpha.$$

Multicalibration ensures that predictions are unbiased across every (weighted) subpopulation defined by $c \in \mathcal{C}$. Importantly, it is always feasible (e.g., perfect predictions are multicalibrated) and—unlike many notions of fairness—exhibits no fairness–accuracy tradeoff. Further, the framework is constructive: There is a boosting-style algorithm—MCBoost—that, given a small sample of labeled data, produces a multicalibrated prediction function (19, 23). See *SI Appendix*, section 2.A for a formal description of the definition and algorithm.

At a high level, the multicalibration algorithm works by iteratively identifying a function $c \in \mathcal{C}$ (auditing via regression) under which the current predictions violate multicalibration. Then, the algorithm updates the predictions to improve the calibration over the weighted subpopulation defined by c. This process repeats until $\tilde{p}$ is multicalibrated. The approach, and how it differs from IPSW, is depicted in Fig. 1.

Multicalibration Guarantees Universal Adaptability

Our main theorem relates the estimation error obtained using multicalibration to that of propensity score reweighting for any source and target populations. Specifically, given a class of

*By convention, we assume a uniform prior over $Z \in \{\mathrm{s}, \mathrm{t}\}$. The choice of uniform prior is arbitrary and only affects the proportions of source and target samples used in learning the propensity score, as well as the constant factor for reweighting samples. See *SI Appendix*, section 1.A for a formal derivation.

†Ref. 19 introduces two variants of multicalibration. The variant presented in this manuscript is the weaker variant, often called "multiaccuracy." This notion is sufficient to obtain our main result, but working with the stronger version of multicalibration yields an even stronger guarantee of universal adaptability. We explore these extensions in *SI Appendix*, section 2.A.

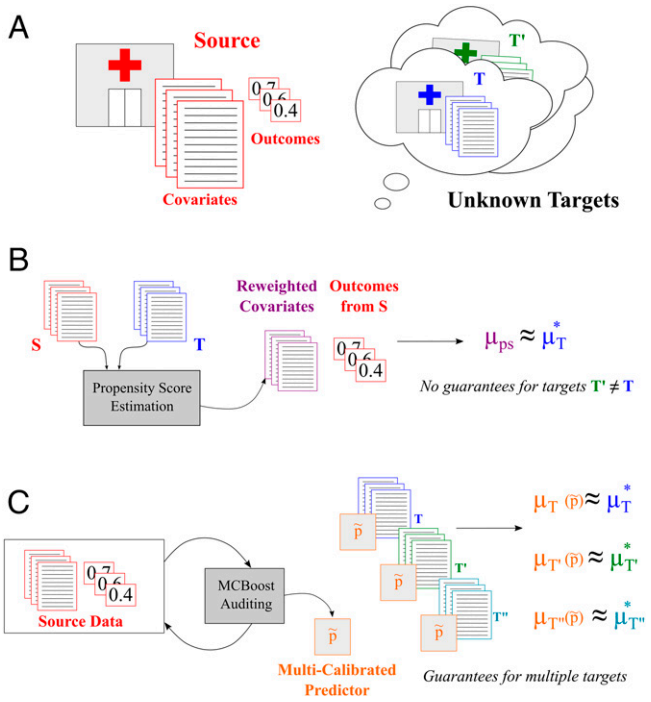

Fig. 1. (*A*) Setting. We consider a single source of labeled data (covariates and outcomes), for example, from a hospital study. Our goal is to make inferences that generalize to different target distributions, for example, to inform patient care at other hospitals. (*B*) Propensity scoring. First, unlabeled samples from the source and target are employed to learn a propensity score. Then, target-specific estimates are computed on the reweighted (labeled) source samples. (*C*) Universal adaptability via multicalibration. The MCBoost algorithm iteratively performs regression over the source data, updating the prediction function, and returning a multicalibrated $\tilde{p}$. The output predictor can be used to make estimates in any target distribution, with performance similar to that of the target-specific propensity score estimators.

propensity scores Σ, we define a corresponding class of functions $\mathcal{C}(\Sigma) = \{c_\sigma : \sigma \in \Sigma\}$, where c_σ is the likelihood ratio under σ, defined as follows:

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

With this class of functions in place, we can state the theorem, which establishes universal adaptability from multicalibration. The guaranteed estimation error depends directly on the misspecification error $\Delta_{\text{st}}(\sigma)$ for any propensity score $\sigma \in \Sigma$.

Theorem. *Suppose $\tilde{p} : \mathcal{X} \to [0,1]$ is a $(\mathcal{C}(\Sigma), \alpha)$-multicalibrated prediction function over source distribution $\mathcal{D}_{\text{s}}$. Then, for any target distribution $\mathcal{D}_{\text{t}}$, and for any $\sigma \in \Sigma$, the estimator $\mu_{\text{t}}(\tilde{p})$ is $(\Sigma, \alpha + \Delta_{\text{st}}(\sigma))$ universally adaptable.*

Note that, in the case where Σ is well specified (i.e., $\Delta_{\text{st}}(\sigma) = 0$), then, as with the propensity scoring approach, the multicalibrated estimator will be nearly unbiased (up to the multicalibration error α). In other words, even though a multicalibrated prediction function can be learned using only samples from the source, when evaluated on the target, the estimation error is nearly as good as the IPSW inferences, which explicitly model the shift. Thus, appealing to the learning algorithm of ref. 19, it is possible to obtain universally adaptable estimators that perform as well as the shift-specific propensity-based estimates. Further technical details and a proof of the theorem are included in *SI Appendix*, section 1.B.

Multicalibrated Prediction Functions Adapt to Subpopulation Shifts

We consider an inference task from epidemiology. To model distributional shift, we use data from two US household surveys. As source, we use the third US National Health and Nutrition Examination Survey (NHANES), with 20,050 observations in the adult sample (24); as target, we use the (weighted) US National Health Interview Survey (NHIS), with 19,738 observations from the Department of Health and Human Services Year 2000 Health objectives interview (25). NHANES differs from NHIS in composition, for example, in sampling rates of demographic groups. Both surveys are linked to death certificate records from the National Death Index (26). We infer 15-y mortality rates, using covariates (age, sex, ethnicity, marital status, education, family income, region, smoking status, health, BMI) for estimating both propensity scores (IPSW) and the multicalibrated predictor of mortality.

To evaluate the methods, we measure the estimation error on the overall target distribution, and also on demographic subpopulations. We can view each subpopulation G as its own extreme shift, where $\mathbf{Pr}[\, Z = \mathrm{t} \mid X = x \,] = 0$ for any $x \notin G$. In this way, the experiment measures adaptability across many different shifts simultaneously. For a subgroup G, we denote, by NHANES(G) and NHIS(G), the restrictions to individuals in the subgroup.

Methods. For each group G, as a naive baseline, we estimate the target mean over NHIS(G) using the source mean over NHANES(G). We evaluate two IPSW approaches: First, we run IPSW with a global propensity score between NHANES and NHIS, reporting the propensity-weighted average over NHANES(G); second, we run a stronger subgroup-specific IPSW, where we learn a separate propensity score for each group G between NHANES(G) and NHIS(G).

We evaluate the adaptability properties of naive and multicalibrated prediction functions. To start, we evaluate estimates derived using a random forest (RF), trained on the source data.[‡] Then, we evaluate the performance of an RF that is postprocessed for multicalibration using MCBoost, auditing with ridge regression, using samples from NHANES (MCRidge). In all predictor-based methods, we estimate the target expectation using unlabeled samples from NHIS(G). Finally, we run a hybrid predictor-based method, where we estimate a propensity score between NHANES and NHIS, then learn RF to predict outcomes over propensity-weighted samples from NHANES(G), providing a strong benchmark. [§] For a detailed description of the methods and results for additional techniques, see *SI Appendix*, section 3.

Results. We report the estimation error for each technique in Tables 2 and 3. First, we observe that the source and target compositions differ in significant ways: The distribution of covariates shifts nontrivially, resulting in different expected mortality rates across groups. As such, the naive inference suffers considerable estimation error. The techniques that account for this shift—through propensity scoring or universal adaptability—incur significantly smaller errors. Among the propensity scoring techniques, the overall IPSW, subgroup-specific IPSW, and hybrid approaches perform similarly overall; on the race-based demographic groups, the subgroup-specific IPSW model performs better than the others. Among the RF-based inferences, the naive approach exhibits nontrivial estimation error overall and on many subpopulations. The

[‡]The naive predictor approach is a simple variant of the mass imputation strategy explored in ref. 27.

[§]This hybrid strategy is developed extensively in the study of "doubly robust" estimation (28, 29).

Table 2. Source and target composition

	Sample composition		Average mortality	
	NHANES	NHIS	NHANES	NHIS
Overall			27.67	17.57
Male	46.75	47.74	30.56	18.77
Female	53.25	52.26	25.11	16.48
Age 18 y to 24 y	13.87	13.36	3.81	2.23
Age 25 y to 44 y	36.43	43.61	5.70	3.86
Age 45 y to 64 y	23.11	26.62	22.71	17.66
Age 65 y to 69 y	6.34	5.10	48.61	45.52
Age 70 y to 74 y	6.57	4.57	64.24	60.03
Age 75+ y	13.69	6.75	90.47	86.25
White	42.56	75.81	37.25	18.70
Black	27.30	11.19	23.08	18.94
Hispanic	28.59	9.01	18.38	10.18
Other	1.55	3.99	15.62	8.96

For NHANES and NHIS, subpopulations are listed with prevalence (percent) in the distributions, and average mortality rate (percent) in NHANES and NHIS.

RF that has been postprocessed to be multicalibrated has consistently smaller errors, and obtains estimation performance comparable or better than the IPSW approaches.

The performance obtained via multicalibration highlights how a single multicalibrated prediction function can actually generalize to many different target distributions, competitive with shift-specific inference techniques. Intuitively, the strong performance across subgroups highlights the connection between universal adaptability and multicalibration as a notion of fairness: To be multicalibrated, the predictor must model the variation in outcomes robustly—not just overall, but simultaneously across many subpopulations.

Universal Adaptability Maintains Small Error under Extreme Shift

To push the limits of universal adaptability, we design a semisynthetic experimental setup to model extreme shift (i.e., strong differences between source and target distribution). We use data collected by the Pew Research Center (30), with a source of 31,319 online opt-in interviews (OPT) and a reference target of 20,000 observations from high-quality surveys (REF). Our outcome of interest indicates whether an individual voted in the 2014 midterm election. Our inference methods use covariates (age, sex, education, ethnicity, census division). Using OPT and REF, we construct various semisynthetic target distributions. At a high level, we estimate a propensity score σ between OPT and REF that we amplify exponentially according to varying intensities q. Technically, the qth semisynthetic shift is implemented using a propensity score $e^{(q)}$ with odds ratio given as $e^{(q)}(x)/1-e^{(q)}(x) = (\sigma(x)/1-\sigma(x))^q$. See *SI Appendix*, section 4.A for a detailed description of the sampling procedure. We also track how techniques' performance changes based on the mode of shift, based on the model type and specification—logistic regression with linear terms, logistic regression with linear terms and pairwise interactions, decision-tree regression—used to fit the initial propensity score σ.

Methods and Results. We perform target inference using various methods. Fig. 2 shows the (signed) estimation error resulting under different modes and shift intensities for the different methods. Using the source mean as an estimate for the target mean (Naive) incurs significant bias, increasing linearly with the shift exponent in all three modes. We evaluate IPSW, where we fit propensity scores using logistic regression. The standard approach (IPSW) achieves nearly unbiased estimates even under extreme shifts, but has large variance (especially on the logistic shift models). We also evaluate the IPSW approach after trimming the propensity scores (IPSW-trimmed); this results in considerably lower variance estimator, but incurs increased estimation error at extreme shifts. We evaluate four predictor-based approaches. The bias for a baseline RF (RF-Naive) estimate also increases linearly with q, albeit slowly under the tree-based shift. Training a propensity score–weighted RF (RF-Hybrid) leads to improved estimates compared to RF-Naive but still results in considerable bias under logit-based shifts. We explore two variants of multicalibration: one auditing with ridge regression (RF-MC-Ridge) and one with decision tree regression (RF-MC-Tree). The ridge-multicalibrated predictor obtains low overall estimation error, competitive with IPSW, while maintaining reasonable variance. The tree-multicalibrated predictor incurs considerable error on the logistic-based shifts, but maintains low error on the tree-based shift, highlighting how the choice of multicalibration functions $\mathcal{C}$ relative to the true shift affects its adaptability. In *SI Appendix*, section 4, we report and discuss additional inference techniques.

Conclusion

In all, the theory and experiments validate the conclusion that it is possible to achieve universal adaptability in diverse contexts. By training a single multicalibrated prediction function on source data, the analyst can guarantee estimation error on any target population, comparable to the performance achieved by explicitly modeling the propensity score. In this way, universal

Table 3. Comparison of inference methods

		IPSW		RF	RF	
	Naive	Overall	Subgroup	Hybrid	Naive	MC-Ridge
Overall	10.10 (57.5)	2.37 (13.5)	—	**0.35 (2.0)**	1.11 (6.3)	**0.52 (3.0)**
Male	11.80 (62.9)	2.51 (13.4)	0.91 (4.9)	−1.34 (7.1)	−0.34 (1.8)	**0.11 (0.6)**
Female	8.63 (52.4)	2.40 (14.6)	3.99 (24.2)	1.89 (11.5)	2.43 (14.8)	**0.90 (5.4)**
Age 18 y to 24 y	1.57 (70.5)	**0.00 (0.1)**	−0.39 (17.5)	5.18 (232.1)	6.03 (270.2)	1.76 (79.0)
Age 25 y to 44 y	1.84 (47.6)	**−0.20 (5.2)**	−0.41 (10.6)	**0.29 (7.6)**	0.82 (21.2)	0.66 (17.2)
Age 45 y to 64 y	5.05 (28.6)	−0.75 (4.2)	−0.41 (2.3)	**0.04 (0.2)**	0.86 (4.8)	−0.29 (1.6)
Age 65 y to 69 y	3.09 (6.8)	−4.23 (9.3)	−5.23 (11.5)	−5.40 (11.9)	**−3.52 (7.7)**	**−1.99 (4.4)**
Age 70 y to 74 y	4.21 (7.0)	−1.36 (2.3)	**0.47 (0.8)**	−4.07 (6.8)	−3.02 (5.0)	**0.61 (1.0)**
Age 75+ y	4.22 (4.9)	3.53 (4.1)	2.85 (3.3)	**−0.25 (0.3)**	0.51 (0.6)	2.19 (2.5)
White	18.55 (99.2)	3.53 (18.9)	0.75 (4.0)	**0.19 (1.0)**	1.03 (5.5)	0.69 (3.7)
Black	4.14 (21.9)	−4.00 (21.1)	**−0.48 (2.5)**	−1.30 (6.8)	**−0.66 (3.5)**	**−0.52 (2.7)**
Hispanic	8.20 (80.5)	1.73 (17.0)	**0.48 (4.7)**	2.84 (27.9)	2.91 (28.6)	1.55 (15.2)
Other	6.66 (74.4)	**−0.02 (0.2)**	−3.54 (39.5)	2.44 (27.3)	3.52 (39.3)	−2.06 (23.0)

Shift-aware inferences: Estimation error in inferred mortality rate for each technique on each subpopulation is shown (percent error in parentheses). For each subgroup, the technique achieving (within 2×) best performance is in bold. Results highlight the universal adaptability of the multicalibrated prediction function (MC-Ridge).

Fig. 2. Relative error (percent) in inferred voting rates under synthetic shift (varying intensity q). Shifts are given by three modes of propensity score: logistic with linear terms (Logit-linear), logistic with linear terms and pairwise interactions (Logit-interaction), and decision tree (Tree). Error of (naive, IPSW, and RF-based) inferences plotted against unbiased baseline (relative error = 0).

adaptability suggests a pathway for rapid and accessible dissemination of statistical inferences to many target populations: After running a study, a research organization can publish a multi-calibrated prediction function based on their findings without the need to reweight data for novel targets. This strategy may be particularly effective in communities that want to implement evidence-based decision-making but do not have the resources to collect high-quality data or perform propensity score estimation on their own.

Data Availability. Previously published data were used for this work. Data and code are linked within the following anonymous repository: https://osf.io/kfpr4/?view_only=adf843b070f54bde9f529f910944cd99. Previously published data are from refs. 24–26 and 30.

ACKNOWLEDGMENTS. M.P.K. is supported by the Miller Institute for Basic Research in Science. Part of this work was completed at Stanford University while supported by NSF Award IIS-1908774. C.K. is supported by the University of Mannheim, Germany. Part of this work was completed while at the University of Maryland and at the Ludwig Maximilian University of Munich. F.K.'s research is partially supported by the NSF's Rapid Response Research Grant 2028683, "Evaluating the Impact of COVID-19 on Labor Market, Social, and Mental Health Outcomes," the German Research Foundation, Collaborative Research Center SFB 884 "Political Economy of Reforms" (Project A8, Number 139943784), and the Mannheim Center for European Social Research. Part of this work was completed during the semester on privacy at Simons Institute for the Theory of Computing, Berkeley. M.P.K., S.G., and O.R. are supported, in part, by the Simons Collaboration on the Theory of Algorithmic Fairness. O.R. is supported by NSF Award IIS-1908774 and Sloan Foundation Grant 2020-13941.

1. N. Potok, Deep policy learning: Opportunities and challenges from the evidence act. *Harv. Data Sci. Rev.* **1**, e63f8f (2019).
2. High-Level Panel of Eminent Persons on the Post-2015 Development Agenda, "A new global partnership: Eradicate poverty and transform economies through sustainable development" (Tech. Rep., United Nations, 2013).
3. D. B. Rubin, Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.* **66**, 688–701 (1974).
4. J. D. Angrist, J. S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, NJ, 2009).
5. K. Zhou, Z. Liu, Y. Qiao, T. Xiang, C. C. Loy, Domain generalization in vision: A survey. arXiv [Preprint] (2021). https://arxiv.org/abs/2103.02503v4 (Accessed 1 August 2021).
6. J. Peters, P. Bühlmann, N. Meinshausen, Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Series B Stat. Methodol.* **78**, 947–1012 (2016).
7. P. Bühlmann, Invariance, causality and robustness. *Stat. Sci.* **35**, 404–426 (2020).
8. J. Pearl, E. Bareinboim, External validity: From do-calculus to transportability across populations. *Stat. Sci.* **29**, 579–595 (2014).
9. P. Patil, G. Parmigiani, Training replicable predictors in multiple studies. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2578–2583 (2018).
10. I. Gibbs, E. Candès, Adaptive conformal inference under distribution shift. arXiv [Preprint] (2021). https://arxiv.org/abs/2106.00170v2 (Accessed 1 August 2021).
11. B. Ackerman, C. R. Lesko, J. Siddique, R. Susukida, E. A. Stuart, Generalizing randomized trial findings to a target population using complex survey population data. *Stat. Med.* **40**, 1101–1120 (2021).
12. H. I. Weisberg, V. C. Hayden, V. P. Pontes, Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? *Clin. Trials* **6**, 109–118 (2009).
13. R. H. Dehejia, S. Wahba, Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Am. Stat. Assoc.* **94**, 1053–1062 (1999).
14. M. A. Hernán *et al.*, Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* **19**, 766–779 (2008).
15. S. R. Cole, E. A. Stuart, Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* **172**, 107–115 (2010).
16. C. Frangakis, The calibration of treatment effects from clinical trials to target populations. *Clin. Trials* **6**, 136–140 (2009).
17. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, "Fairness through awareness" in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Association for Computing Machinery, 2012), pp. 214–226.
18. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
19. Ú. Hébert-Johnson, M. P. Kim, O. Reingold, G. Rothblum, "Multicalibration: Calibration for the (computationally-identifiable) masses" in *Proceedings of the 2018 10th International Conference on Machine Learning*, J. G. Dy, A. Krause, Eds. (Association for Computing Machinery, 2018), pp. 1939–1948.
20. P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
21. M. R. Elliott, R. Valliant, Inference for nonprobability samples. *Stat. Sci.* **32**, 249–264 (2017).
22. C. Kern, Y. Li, L. Wang, Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. *J. Surv. Stat. Methodol.* **9**, 1088–1113 (2020).
23. M. P. 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* (AAAI Press, 2019), pp. 247–254.
24. Department of Health and Human Services, Third national health and nutrition examination survey, 1988–1994, NHANES III household adult file (National Center for Health Statistics, 1996).
25. National Center for Health Statistics, Public use data tape documentation, part I, national health interview survey, 1994 (National Center for Health Statistics, 1995).
26. National Center for Health Statistics, Office of analysis and epidemiology, public-use linked mortality file, 2015 (2013). https://www.cdc.gov/nchs/data-linkage/mortality-public.htm (Accessed 1 September 2021)
27. S. Chen, S. Yang, J. K. Kim, Nonparametric mass imputation for data integration. *J. Surv. Stat. Methodol.*, https://doi.org/10.1093/jssam/smaa036 (2020).
28. R. Valliant, Comparing alternatives for estimation from nonprobability samples. *J. Surv. Stat. Methodol.* **8**, 231–263 (2019).
29. H. Bang, J. M. Robins, Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973 (2005).
30. A. Mercer, A. Lau, C. Kennedy, For weighting online opt-in samples, what matters most? (Pew Research Center, 2018). https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/ (Accessed 7 June 2020).