Connections between weighting-based and imputation-based domain adaptation methods

Nina Katz-Christy
ninakatzchristy@college.harvard.edu

Abstract

Imputation-based and weighting-based approaches to domain imputation are often viewed as contrasting frameworks, yet recent empirical and theoretical results suggest they may more similar than previously thought, and even equivalent in some cases. In this work, I review some of the most recent theoretical results connecting these frameworks and prove a bound for the relative error of a propensity score-weighted estimator with respect to the optimal imputation-based estimator corresponding to a class of outcome functions. An epidemiological task is used to empirically compare methods from each framework.

1 Introduction

In many applications, we only have access to unlabeled data from the population on which we wish to perform inference, but we have labeled data from another population. For example, suppose we are interested in determining the percentage of individuals in a state who have been infected with COVID-19. As it would be infeasible to test the entire population, one strategy would be to test all individuals who attend an in-person doctor's appointment. However, the overall population may differ in important ways from the individuals who come to the doctor's, so it is insufficient to simply apply the rate of COVID-19 cases of those who are tested to the overall population. We would need to account for the difference in population in order to transfer the results from the tested population to the overall population. The task of leveraging labelled data from one population to perform inference on another, compositionally distinct one is known as *domain adaptation*.

A related task is that of the estimation of counterfactual outcomes, usually for the purpose of estimating the *average treatment effect on the treated (ATT)*. In these applications, the limitation is not that the outcome of interest is infeasible to measure directly, but that the outcome of interest is itself hypothetical. For example, consider an educational program that enrolls students based on demonstrated academic achievement. To estimate the effect of the program, it is necessary to estimate the expected performance of the students in the program had they not been in the program, referred to as the *counterfactual*. A naive approach would be to assume the performance of the students in the program would have been the same as those not given the opportunity to enroll. However, the students in the program might differ in important ways from the other students, and this must be taken into account.

There are two major approaches to domain adaptation, imputation-based estimation and propensity score-weighted estimation. In the context of the educational example, an imputation-based approach would first learn how the observed performance of students not in the program differs by individual characteristics, such as their parents' educational background. Then, this estimated function would be applied to the observed characteristics of students in the program to estimate their counterfactual performance. Conversely, a propensity score-based approach would first learn how the composition of students in the program differed from those not selected to enroll. Then, to estimate the counterfactual performance of students who were in the program, the true performances of students not in the program would be re-weighted to approximate the composition of students in the program. For

example, if the proportion of students in the program with college-educated parents was higher than of those not in the program, the performance of students not in the program with college-educated parents would be given more weight in this estimate. These two approaches are illustrated in 2.

Clearly, imputation-based methods rely on accurate estimation of the outcome function and propensity score-based methods rely on accurate estimation of the true distributional shift. While these learned functions are typically evaluated via their l_1 error (i.e. the expectation of the absolute error over the source distribution), the loss resulting from their respective methods is not directly proportional to this estimation error. In fact, the error for imputation-based and propensity score-based estimation depends, respectively, on the true distribution shift and the true outcome function.

This observation has led to a number of methods that aim to minimize the estimation error while invoking as few assumptions as possible. In this paper, we review such methods, summarise connections between the two broad approaches, and propose a method based on a notion introduced in the algorithmic fairness literature. Our proposed method can be implemented prior to observing the outcome data, yet yields results that are provably comparable to the outcome-specific method of imputation-based estimation. We apply our proposed method to data from two U.S. household surveys.

2 Setup and Assumptions

We adopt the setup and notation in Kim et al. [2022] to emphasize the connections to their approach. In particular, we define a joint distribution over (X, Y, Z) triples, for covariates $X \in \mathcal{X}$, outcome $Y \in \mathcal{Y}$, and source vs. target indicator $Z \in \{s, t\}$. We use D_s and D_t to denote the joint distributions over X, Y pairs conditioned on $Z = s$ and $Z = t$, respectively. Similarly, we let U_s and U_t denote the marginal distribution over X, conditioned on $Z = s$ and $Z = t$, respectively. We use $D^*(X)$ to denote the conditional distribution of $Y|X$ and let $f(x) = E(Y|X)$ denote the conditional expectation of the outcome (also referred to as the *outcome function*).

Our estimand is $\mu_t^* = E_{D_t}[Y]$. However, we only observe $(X_1, Y_1), \dots (X_n, Y_n) \sim D_s$ and $X_1, \dots X_n \sim U_t$, as illustrated by the blue blocks in Figure 1. In the *domain adaptation* setting, the distributions D_s and D_t are both assumed to be unknown. In the *non-probability sampling* literature, the source and target distributions are referred to as *non-probability* and *probability* distributions, respectively, and the probability (target) distribution is typically assumed to be known.

In the causal inference context, the estimand can be interpreted as the average expected outcome of the treated units, had they not been treated. When combined with the observed treated outcomes, this can be used to estimate the ATT. Analogously, the *average treatment effect on the untreated (ATU)* can be estimated by estimating the average expected outcome of the untreated units, had they been treated.

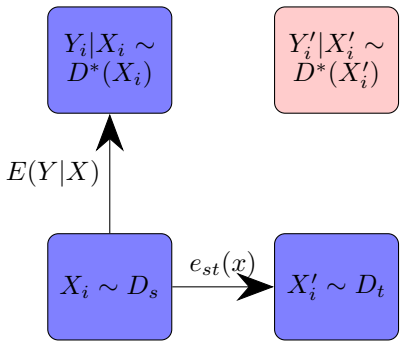


Figure 1: Problem Setting

2.1 Assumptions

The first assumption we make states that the relationship between outcome Y and covariates X is the same over the source and target distributions. This assumption is often referred to in the causal

inference literature as *ignorability*. Formally, we assume,

$$\Pr[Y|X, Z] = \Pr[Y|X].$$

Technically, our results depend only on a weaker form of ignorability, *weak mean-ignorability*, defined in Kallus [2020b] as

$$E[Y|X, Z] = E[Y|X]$$

Informally, this assumption states that the set of covariates must capture all of the systematic differences between the source and target distribution, at least as far as they relate to the outcome. Note that this depends crucially on how the covariates X are defined. For example, in the educational program example, if income is not included as a covariate, this assumption might be violated if (1) the distribution over income categories differs between the schools and (2) income is associated with different performance outcomes. The assumption would be satisfied if either of these was false.

We also make the standard *overlap* assumption that the probability $Pr(Z = s|X)$ is bounded away from 0 and 1 for all $X \in \mathcal{X}$.

Finally, we assume a uniform prior over $Z \in \{s, t\}$. This is a standard convention as the relative prior probabilities of source vs target distribution only affect the constant factor used in propensity score weighting.

3 Imputation Approach

A common approach to domain adaptation is to first attempt to learn a *predictor* which approximates the conditional expectation $E(Y|X)$, and then apply this function to unlabelled samples from the target distribution. This approach is illustrated in 2a. The corresponding estimator is defined as follows.

Definition 1 (Imputation-Based Estimator)**.** *For a predictor p, we define the imputation-based estimator for μ_t^* as*

$$\mu_t(p) = E_{D_t}[p(X)]$$

When the predictor is equal to the true conditional expectation, this estimator is accurate: $\mu_t(p) = \mu_t^*$.

4 Propensity Score Weighting

An alternative approach is *propensity score weighting*. A distribution shift can be characterized by its *propensity score odds*, which is defined as the ratio of the probabilities of being in the target and source distributions, respectively, for a given covariate profile $X = x$. In the causal inference setting, this is the odds of being untreated.

Definition 2 (True Propensity Odds)**.**

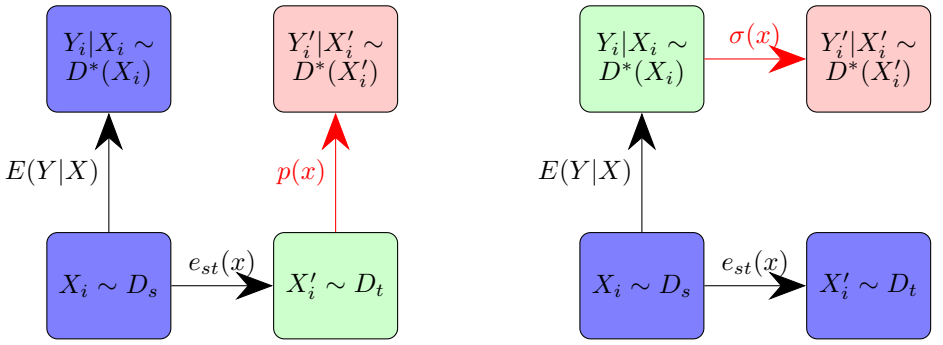
$$e_{st}(x) = \frac{\Pr[Z = t|X = x]}{\Pr[Z = s|X = x]} = \frac{1 - \Pr[Z = s|X = x]}{\Pr[Z = s|X = x]}$$

Reweighting samples from the source distribution according to an estimate of the propensity odds yields an unbiased estimate of the expectation of Y in the target population. Formally, the propensity score weighted estimator is defined as follows.

Definition 3 (Propensity Score-Based Estimator)**.** *For a candidate propensity odds $\sigma(X)$, we define the propensity score-based estimator of μ_t^* as*

$$\mu_t^{PS}(\sigma) = E_{D_s}[\sigma(X)Y]$$

When the candidate propensity odds is equal to the true propensity odds, the estimator is accurate: $\mu_t^{PS}(\sigma) = \mu_t^*$.



(a) Imputation-Based Approach (b) Weighting-Based Approach

Figure 2: Two broad approaches for domain adaptation. Blue blocks denote data necessary for the first stage of each method, green blocks denote data necessary for the second stage, and the pink block denotes the unobserved data to be estimated.

4.1 Propensity Score vs Propensity Odds

The probability of being in the source population (or of being treated), $P[Z = s|X = x]$, is often referred to as the *propensity score*. Weighting the outcomes of the untreated population by the inverse propensity score yields an unbiased estimate of the expected potential outcome $Y(0)$ under the full population, which can then be used to estimate the *average treatment effect*. In contrast, we focus on estimating the expected counterfactual $Y(0)$ just for the treated population, which can be estimated by weighting the outcomes of the untreated population with the propensity odds.

5 Practical Considerations

Note that both approaches proceed in two stages. In the first, we estimate, respectively, the outcome function or the propensity odds. In the second, we apply this learned function to, respectively, unlabeled data from the target distribution, or labelled data from the source distribution. Intuitively, the "heavy lifting" occurs in the first stage, as the second simply involves evaluating the learned function and taking a sample average.

Note that estimating the outcome function for the imputation-based approach only requires samples from the source distribution. This means a practitioner can estimate this function prior to observing examples from the target distribution. Furthermore, they can use the same estimated function to estimate the expected outcome over multiple different target distributions. This motivates the characterization of imputation-based approaches as *target-independent* [Kim et al., 2022].

Conversely, estimating the propensity odds only requires unlabelled samples from the source and target distributions. This means a practitioner can estimate the propensity odds prior to collecting outcome data. Furthermore, they can use the same estimated propensity odds to estimate multiple different outcomes. In contrast to the target target-independence of imputation-based approaches, propensity score-based estimation can be characterized as *multi-task oriented* [Wang et al., 2022].

6 Error Characterizations

We note that the expectation in both estimators must be approximated by an empirical mean. However, we assume for this analysis that labeled examples from the source population and unlabeled examples from the target population are inexpensive enough that the error induced by this approximation to the expectation is negligible. Following Kim et al. [2022], we define the error of a statistic $\tilde{\mu}$ as the absolute distance from the true expectation over the target distribution. In the propensity-score based setting (see, e.g. Kallus [2020a]), this is defined as the conditional bias, where we condition on the data used to estimate the propensity score.

If the propensity odds is estimated perfectly, then the propensity score-based estimate will have no error and if the conditional expectation $E(Y|X)$ is estimated perfectly by the predictor $p(X)$, then the imputation-based estimate will have no error. However, the error induced by incorrectly estimating either the propensity odds $e_{st}(x)$ or the conditional expectation $f(x) = E(Y|X = x)$ for their respective methods is not simply the l_1-error of the respective estimators, $\sigma(x)$ and $p(x)$. In particular, the error of the imputation-based estimate for an estimated conditional expectation $p(X)$, true conditional expectation $f(x) = E(Y|X)$, and true propensity odds $e_{st}(x)$ is

$$|\mu_t^I(p) - \mu_t^*| = |E_{D_s}[e_{st}(X)\,(p(X) - f(X))]| \tag{1}$$

Similarly, the error of the propensity score-based estimate for an estimated propensity odds $\sigma(x)$, true propensity odds $e_{st}(x)$, and true conditional expectation $f(x) = E(Y|X = x)$ is

$$|\mu_t^{PS}(\sigma) - \mu_t^*| = |E_{D_s}[f(X)\,(\sigma(X) - e_{st}(X))]| \tag{2}$$

While these results have been shown in other contexts, both proofs are included the appendix for completeness.

7 General Methods for Bounding Error

These characterizations imply that, if we know (or hypothesize) that the true propensity odds $e_{st}(x)$ is in some class of functions Σ, we can bound the error of the imputation-based estimate by learning an estimate $p(x)$ of $f(x)$ such that

$$\max_{\sigma \in \Sigma} E_{D_s}[\sigma(X)\,(p(X) - f(X))] \tag{3}$$

is bounded. Conversely, if we know (or hypothesize) that the true conditional expectation $f(x)$ is in some class of functions $\mathcal{P}$, we can bound the error of the propensity score-based estimate by learning an estimate $\sigma(x)$ of $e_{st}(x)$ such that

$$\max_{p \in \mathcal{P}} E_{D_s}[p(X)\,(\sigma(X) - e_{st}(X))] \tag{4}$$

is bounded.

7.1 Connection to Multi-Accuracy

Both of these general methods can be stated with the notion of *multi-accuracy*, a concept first defined by Hébert-Johnson et al. [2018] as a measure of the fairness of a prediction algorithm. The definition has since been generalized and applied to various settings [Kim et al., 2022, Gopalan et al., 2021a,b].

Definition 4 (Multi-Accuracy)**.** *A hypothesis/predictor h is α-multi-accurate for outcome Y with respect to a class of functions $\mathcal{C}$ if for all $c \in \mathcal{C}$,*

$$\max_{c \in \mathcal{C}} |E_{X \sim \mu}[c(X)\,(h(X) - Y)]| \leq \alpha$$

Hébert-Johnson et al. [2018] also proposed a boosting-based algorithm to learn a multi-accurate predictor. Since then, others have proposed alternative algorithms and further analyzed the sample complexity of this learning task [Hu et al., 2022].

Motivated by this connection, Kim et al. [2022] propose estimating the outcome function for imputation-based estimation and then post-processing this estimated function to ensure it satisfies multi-accuracy for some class of propensity odds ratios. In the propensity score-based setting, Gopalan et al. [2021b] suggest requiring multi-calibration, a slightly stronger requirement than multi-accuracy, of learned propensity odds, but they do not motivate their suggestion via this error characterization.

By the error characterizations above,

- α-multi-accuracy of a predictor $\tilde{p}$ with respect to a class of candidate propensity odds Σ ensures the corresponding imputation-based estimate has error at most α if the shift is correctly specified (i.e. $\sigma^* \in \Sigma$) and
- α-multi-accuracy of a propensity odds $\tilde{\sigma}$ with respect to a class of candidate predictors $\mathcal{P}$ ensures the corresponding propensity score-based estimate has error at most α if the conditional expectation is correctly specified (i.e. $E(Y|X) \in \mathcal{P}$).

However, this does not provide any guarantees for when Σ or $\mathcal{P}$, do not contain, respectively, the true propensity odds or the true outcome function. It turns out we can still bound the error when the relevant estimator is multi-accurate.

Informally, Kim et al. [2022] show the following:

Multi-accuracy of a predictor $\tilde{p}$ with respect to a class of candidate propensity odds Σ ensures the corresponding imputation-based estimate has (1) unconditional low error if the shift is correctly specified (i.e. $\sigma^* \in \Sigma$), and (2) not much more error than the propensity score-based estimator corresponding to the best $\sigma \in \Sigma$ otherwise.

Analogously, I show:

Multi-accuracy of a propensity odds $\tilde{\sigma}$ with respect to a class of candidate predictors $\mathcal{P}$ ensures the corresponding propensity score-based estimate has (1) unconditional low error if the conditional expectation is correctly specified (i.e. $E(Y|X) \in \mathcal{P}$), and (2) not much more error than the imputation-based estimator corresponding to the best $p \in \mathcal{P}$ otherwise.

To state our main result, we first define the *outcome-misspecification error*, which measures how close a predictor $p \in \mathcal{P}$ is to the true outcome function.

Definition 5 (Outcome-Misspecification Error)**.**

$$\Delta_s(p) = E_{D_s}[|(Y - p(X))|]$$

Theorem 1 formalizes our main result. We state it using the same notation and structure as Kim et al. [2022] to emphasize the symmetry between our result and theirs.

Theorem 1 (Universal Adaptability)**.** *For source and target distributions D_s and D_t over $\mathcal{X}$ with true propensity odds $e_{st} : \mathcal{X} \to [0, 1]$, suppose $\mathcal{P} \subseteq \{\mathcal{X} \to \mathcal{Y}\}$ is a collection of candidate predictor functions. Suppose $\tilde{\sigma} : \mathcal{X} \to \mathbb{R}^+$ is a $(\mathcal{P}, \alpha)$-multi-accurate propensity odds over the source D_s; then, for any conditional distribution $D^*(X)$, the propensity score-weighted estimator $\mu_t^{PS}(\tilde{\sigma}) = \mathbb{E}_{D_s}[\tilde{\sigma}(X) \cdot Y]$ is $(\Delta_{p^*}(p) + \alpha)$-close to the imputation-based estimator corresponding to any $p \in \mathcal{P}$. Furthermore, the estimation error of the propensity score-weighted estimator is bounded by the sum of the imputation-based estimation error, the outcome-misspecification error, and α, under the best-fit $p \in \mathcal{P}$. That is,*

$$er_t(\mu_t^{PS}(\tilde{\sigma})) \leq \min_{p \in \mathcal{P}} \{er_t(\mu_t(p)) + \Delta_{p_s}(p) + \alpha\}$$

This result demonstrates that, if one can determine a class of functions $\mathcal{P}$ with respect to which a propensity score odds estimate bounds the quantity 4, the corresponding propensity score-weighted estimate will be provably close to the imputation-based estimate corresponding to the optimal $p \in \mathcal{P}$.

Conversely, the result shown in Kim et al. [2022] demonstrates that, if one can determine a class of functions Σ with respect to which a predictor bounds the quantity 3, the corresponding imputation-based estimate will be provably close to the propensity score-weighted estimate corresponding to the optimal $\sigma \in \mathcal{C}$.

8 Connection to Balancing Weights

Some literature considers a generalization of the propensity score-based approach as described here, where instead of attempting to estimate the propensity odds directly, they learn a function of the data that, when used to weight new samples from the source distribution as in the propensity score-weighting approach 1 yields a corresponding estimate with minimal error. Here we refer to these

learned functions as *weighting functions* and the related class of methods as *weighting-based methods*, to emphasize that they do not directly aim to approximate the propensity score odds. A common formulation is to consider minimizing the *conditional mean squared error (CMSE)* of the corresponding estimate where the data used to estimate the weighting function are viewed as fixed and the randomness is over the data used to estimate the expectation. When the CMSE is decomposed into a bias and variance term, the bias term is equivalent to the error 4 we consider here. The variance term depends on the variance of the weights and the variance of the conditional distribution $Y|X$.

Many weighting-based approaches characterize the minimization of the CMSE as a constrained optimization problem, where the constraints correspond to bounding the error 4 for some class of functions $\mathcal{P}$ and the goal is to minimize the variance of the weights while satisfying this constraint. The error characterization 3 is also known as an *Integral Probability Metric* between distributions D_s and D_t for function class $\mathcal{P}$ [Müller, 1997]. The functions $p \in \mathcal{P}$ in 4 are sometimes called moment functions and requiring that 4 is 0 for a fixed function p is referred to as a *balancing constraint* [Zhao and Percival, 2017]. For different choices of the class of outcome functions $\mathcal{P}$, such as all bounded functions, Lipschitz functions, or RKHS functions, many weighting methods can be viewed as solutions to this optimization problem [Ben-Michael et al., 2021]. Ben-Michael et al. [2021] provide a unified summary of such methods in the causal inference setting, Wang et al. [2022] propose a similar style of balancing weights for the *non-probability sampling* setting, and Kallus [2020a] show how some matching methods can also be viewed as weights minimizing the conditional mean squared error.

While the learned weighting functions are not explicitly trained to estimate the propensity score odds, note that,if $\mathcal{P}$ is the set of all possible functions from $\mathcal{X}$ to $\mathcal{Y}$ (or equivalently satisfies all possible balancing constraints), then the only function that ensures the error 4 is 0 is the true propensity score odds. However, when $\mathcal{P}$ is some smaller set of functions, the true propensity score odds is not the unique function that ensures the error 4 is 0, and these approaches selecting the function that minimizes variance. Consequently, these weighting functions can be viewed as regularized propensity score estimators. In fact, Wang and Zubizarreta [2020] show that approximate covariate balancing weights are equivalent to shrinkage estimation of the propensity score.

9 Implied Outcome Functions in Weighting-Based Methods

In some cases, weighting-based methods can be framed as imputation-based methods. For example, if we assume the outcome function is linear with coefficients bounded in l_2 norm, learning a weighting scheme by minimizing the CMSE and implementing the weighting-based approach 3 is equivalent to fitting a ridge regression model for $E[Y|X]$ from source samples and implementing the imputation-based approach 1 [Zhao and Percival, 2017]. In particular, even though the weighting function learns from target samples, the weighted estimate 3 is equivalent to one obtained via the imputation-based method 1 where the learned outcome function depends only on (and is linear in) labelled source data. Bruns-Smith and Feller [2022] show that this result extends to any class of outcome functions satisfying some minor conditions.

10 Implied Weights in Imputation-Based Methods

Conversely, some imputation-based methods can be framed as weighting-based methods. Chattopadhyay and Zubizarreta [2021] show that if a linear model is fit to labelled source distribution data, the imputation-based estimate 1 can be expressed as a weighting-based estimate 3 with *implied weights* that are functions only of the unlabeled observations. They show that these weights converge point-wise to the true inverse probability weights if the true propensity score is logistic in the covariates.

11 Data Application

11.1 Data

We illustrate these two general approaches via an epidemiological task, where the goal is to estimate 15-year mortality in a target population which differs in substantial ways from the source population. This task and the corresponding data is described in detail in Kim et al. [2022]. Briefly, the datasets for both distributions are publicly available samples from two US household surveys, each linked to death certificate records from the National Death Index [NCHS, 1995a]. The source distribution is based on data collected through the third US National Health and Nutrition Examination Survey (NHANES, n=20,050) and the target distribution is based on data collected through the US National Health Interview Survey (NHIS, n=19,738) [NCHS, 1995b, HHS, 1996]. Table 2 in the Appendix summarises the demographic composition and mortality rates for each population.

11.2 Methods

As a baseline, we first estimate the mortality in the target population simply as the mortality in the source population. Next, we consider three imputation-based methods. First, we fit a random forest (RF) model to the labelled source distribution. Next, we fit a random forest to a subset of the source distribution data and post-process the outcome model with the remaining source distribution data to ensure either multi-accuracy (MA) or multi-calibration (MC) with respect to a class of propensity odds functions defined by ridge regression models. [1]

We implement a hybrid approach which first learns a propensity score model and then is trained to predict the outcome over a source population weighted by this learned propensity score model.

We consider four propensity score weighting approaches. We fit a propensity score model via logistic regression to the full population and use the weighting-based estimator 3. We also fit separate propensity score models on subsets of the data defined by demographic groups and use these subgroup-specific weights to obtain subgroup-specific weighting-based estimates. Next, we fit another logistic regression model to a subset of the unlabelled data and post-process the model with the remaining unlabeled data to ensure either multi-accuracy (MA) or multi-calibration (MC) of the propensity score with respect to a class of outcome functions defined by regression trees. We use the implementation of MCBoost by Kim et al. [2022]. Unfortunately, this implementation requires the outcome to be bounded by 0 and 1 for each observation, precluding us from ensuring multi-accuracy of the propensity score odds. In future work, it would be interesting to ensure multi-accuracy of the propensity score odds, rather than the propensity score, as this would enable direct application of our theoretical results.

11.3 Results

The naive approach yields severely biased estimates, indicating that adjustment for distribution shift is necessary. All adjustment methods perform similarly well, and there is no evidence that multi-calibration provides substantial benefit as compared to multi-accuracy. There also does not appear to be clear evidence that either multi-accuracy or multi-calibration post-processing significantly improves estimation error, but this could be due to the fact that the initial predictor and propensity score model already satisfied the multi-accuracy criteria.

12 Conclusion

We have demonstrated theoretically and empirically various connections between weighting-based and imputation-based estimators. In many cases, the approaches are equivalent, yet a full characterization of such cases is unknown.

[1] While our theoretical results rely only on MA, we show empirical results for MA for comparability to results in Kim et al. [2022].

		Imputation-Based			Hybrid	Propensity Score Weighting			
	Naive	RF Naive	RF MA	RF MC	RF	Naive Overall	Naive Subgroup	Logistic MA	Logistic MC
Overall	10.1 (57.5)	1.2 (6.8)	0.9 (4.9)	0.9 (4.9)	**0.3 (1.5)**	-0.6 (3.5)		1.1 (6.4)	1.1 (6.1)
Male	11.8 (62.9)	-0.3 (1.7)	0.4 (2.4)	0.4 (2.4)	-1.5 (7.9)	**-0.1 (0.7)**	-1.3 (6.9)	0.6 (3.2)	0.3 (1.8)
Female	8.6 (52.4)	2.6 (15.5)	1.2 (7.5)	1.2 (7.5)	1.9 (11.3)	**-1.0 (5.8)**	**0.5 (3.2)**	1.7 (10.4)	1.8 (11.0)
Age 18y to 24y	1.6 (70.5)	6.0 (267.5)	1.7 (76.0)	1.7 (76.0)	4.7 (209.4)	**0.0 (1.1)**	0.2 (9.5)	-0.1 (5.8)	-0.2 (11.2)
Age 25y to 44y	1.8 (47.6)	0.9 (23.4)	0.7 (18.8)	0.7 (18.8)	**0.3 (7.7)**	**-0.3 (6.9)**	**-0.2 (5.4)**	**-0.2 (5.6)**	**-0.3 (7.0)**
Age 45y to 64y	5.1 (28.6)	1.1 (6.1)	0.1 (0.7)	0.1 (0.7)	0.2 (1.1)	-0.9 (4.9)	-0.1 (0.4)	0.1 (0.7)	**-0.0 (0.2)**
Age 65y to 69y	3.1 (6.8)	-4.0 (8.8)	**-1.3 (2.9)**	**-1.3 (2.9)**	-5.5 (12.2)	-4.3 (9.5)	**-2.3 (5.0)**	**-2.4 (5.3)**	**-2.4 (5.2)**
Age 70y to 74y	4.2 (7.0)	-2.9 (4.8)	1.8 (3.0)	1.8 (3.0)	-4.9 (8.2)	-1.4 (2.3)	**0.0 (0.0)**	-0.2 (0.3)	0.6 (1.0)
Age 75+ y	4.2 (4.9)	**0.7 (0.8)**	4.0 (4.7)	4.0 (4.7)	**-0.5 (0.6)**	3.7 (4.3)	3.2 (3.7)	3.8 (4.4)	3.5 (4.0)
White	18.6 (99.2)	1.1 (5.9)	1.1 (5.7)	1.1 (5.7)	**0.1 (0.5)**	**-0.1 (0.7)**	-0.7 (4.0)	1.1 (5.7)	0.8 (4.3)
Black	4.1 (21.9)	-0.6 (3.1)	**-0.2 (0.9)**	**-0.2 (0.9)**	-1.3 (6.8)	-5.2 (27.5)	-1.6 (8.6)	-0.4 (2.1)	0.5 (2.4)
Hispanic	8.2 (80.5)	3.0 (29.7)	1.7 (16.8)	1.7 (16.8)	2.7 (27.0)	**0.8 (7.9)**	**1.2 (11.3)**	3.5 (34.1)	3.8 (37.4)
Other	6.7 (74.4)	3.5 (39.3)	-2.0 (22.2)	-2.0 (22.2)	2.2 (24.6)	**-1.3 (14.0)**	**-1.3 (14.9)**	**-0.9 (10.0)**	**-1.4 (15.3)**

Table 1: Comparison of domain imputation methods to estimate mortality rate in target population. Estimation error overall and within demographic groups is shown with absolute percent error in parenthesis. Results within 2x the optimal method are in bold.

References

E. Ben-Michael, A. Feller, D. A. Hirshberg, and J. R. Zubizarreta. The balancing act in causal inference, 2021. URL https://arxiv.org/abs/2110.14831.

D. Bruns-Smith and A. Feller. Outcome assumptions and duality theory for balancing weights, 2022. URL https://arxiv.org/abs/2203.09557.

A. Chattopadhyay and J. R. Zubizarreta. On the implied weights of linear regression for causal inference. *arXiv preprint arXiv:2104.06581*, 2021.

P. Gopalan, A. T. Kalai, O. Reingold, V. Sharan, and U. Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021a.

P. Gopalan, O. Reingold, V. Sharan, and U. Wieder. Multicalibrated partitions for importance weights, 2021b. URL https://arxiv.org/abs/2103.05853.

U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

HHS. Department of health and human services, third national health and nutrition examination survey, 1988–1994, nhanes iii household adult file. 1996.

L. Hu, C. Peale, and O. Reingold. Metric entropy duality and the sample complexity of outcome indistinguishability, 2022. URL https://arxiv.org/abs/2203.04536.

N. Kallus. Generalized Optimal Matching Methods for Causal Inference. *Journal of Machine Learning Research*, 21(62):1–54, 2020a. URL http://jmlr.org/papers/v21/19-120.html.

N. Kallus. Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.*, 21: 62–1, 2020b.

M. P. Kim, C. Kern, S. Goldwasser, F. Kreuter, and O. Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022. doi: 10.1073/pnas.2108097119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2108097119.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

NCHS. National center for health statistics, office of analysis and epidemiology,public-use linked mortality file, 2015, 1995a. URL https://www.cdc.gov/nchs/data-linkage/mortality-public.htm.

NCHS. National center for health statistics, public use data tape documentation, part i,national health interview survey, 1994, 1995b.

Y. Wang and J. R. Zubizarreta. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 2020.

Z. Wang, X. Mao, and J. K. Kim. Functional calibration under non-probability survey sampling. *arXiv preprint arXiv:2204.09193*, 2022.

Q. Zhao and D. Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

13 Appendix

13.1 Survey Data Distribution Shift

	Composition		Mortality	
	NHANES	NHIS	NHANES	NHIS
Overall			27.7	17.6
Male	46.9	47.7	29.9	18.8
Female	53.1	52.3	24.6	16.5
Age 18y to 24y	15.8	13.4	3.3	2.2
Age 25y to 44y	35.4	43.6	5.7	3.9
Age 45y to 64y	22.6	26.6	22.7	17.7
Age 65y to 69y	6.3	5.1	48.6	45.5
Age 70y to 74y	6.4	4.6	64.2	60.0
Age 75+ y	13.5	6.8	0.1	86.2
White	42.3	75.8	36.7	18.7
Black	27.4	11.2	22.5	18.9
Hispanic	28.9	9.0	17.8	10.2
Other	1.5	4.0	15.4	9.0

Table 2: Overall and within each demographic group, percent of population and percent mortality rate

13.2 Lemmas and Proofs

Lemma 1. *For any random variable W,*

$$E_{D_t}[W] = E_{D_s}[e_{st}(X)W]$$

Proof of Lemma 1. Expanding the expectation,

$$E_{D_t}[W] = \sum_{x \in \mathcal{X}} \Pr[X = x | Z = t] \cdot W$$

By Bayes rule:

$$= \sum_{x \in \mathcal{X}} \frac{\Pr[Z = t | X = x] \Pr[X = x]}{\Pr[Z = t]} \cdot W$$

By the uniform priors assumption that $\Pr[Z = t] = \Pr[Z = s]$,

$$= \sum_{x \in \mathcal{X}} \frac{\Pr[Z = t | X = x] \Pr[X = x]}{\Pr[Z = s]} \cdot W$$

Multiplying by $\frac{\Pr[Z=s|X=x]}{\Pr[Z=s|X=x]}$,

$$= \sum_{x \in \mathcal{X}} \frac{\Pr[Z = t | X = x]}{\Pr[Z = s | X = x]} \frac{\Pr[Z = s | X = x] \Pr[X = x]}{\Pr[Z = s]} \cdot W$$

By Bayes rule and definition of $e_{st}(x)$,

$$= \sum_{x \in \mathcal{X}} e_{st}(x) \cdot \Pr[X = x | Z = s] \cdot W$$

Collapsing the expectation,

$$= E_{D_s}[e_{st}(X) \cdot W]$$

□

Proof of Eq 1. By definition,

$$|\mu_t^I(p) - \mu_t^*| = |E_{D_t}[p(X)] - E_{D_t}[Y]|$$

By lemma 1,

$$= |E_{D_s}[e_{st}(X)p(X)] - E_{D_s}[e_{st}(X)Y]|$$
$$= |E_{D_s}[e_{st}(X)(p(X) - Y)]|$$

By iterated expectations,

$$= |E_{D_s}[E_{D_s}[e_{st}(X)(p(X) - Y)|X]]|$$
$$= |E_{D_s}[e_{st}(X)\,(p(X) - f(X))]|$$

□

Proof of Eq 2. By definition,

$$|\mu_t^{PS}(\sigma) - \mu_t^*| = |E_{D_s}[\sigma(X)Y] - E_{D_t}[Y]|$$

By lemma 1,

$$= |E_{D_s}[\sigma(X)Y] - E_{D_s}[e_{st}(X)Y]|$$
$$= |E_{D_s}[Y(\sigma(X) - e_{st}(X))]|$$

By iterated expectations,

$$= |E_{D_s}[E_{D_s}[Y(\sigma(X) - e_{st}(X))|X]]|$$
$$= |E_{D_s}[f(X)(\sigma(X) - e_{st}(X))]|$$

□

Proof of Theorem 1. Fix an arbitrary $p \in \mathcal{P}$. We first show that the propensity-score based estimator is $(\Delta_{p^*}(p) + \alpha)$-close to the imputation-based estimator corresponding to p. Formally, we show,

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

By definition,

$$|\mu_t^{PS}(\tilde{\sigma}) - \mu_t(p)| = |E_{D_s}[\tilde{\sigma}(X)Y] - E_{D_t}[p(X)]|$$

By Lemma 1 (below),

$$= |E_{D_s}[\tilde{\sigma}(X) \cdot Y] - E_{D_s}[e_{st}(X) \cdot p(X)]|$$

Subtracting and adding $E_{D_s}[\tilde{\sigma}(X)p(X)]$,

$$= |E_{D_s}[\tilde{\sigma}(X) \cdot (Y - p(X))] + E_{D_s}[\tilde{\sigma}(X) \cdot p(X)] - E_{D_s}[e_{st}(X) \cdot p(X)]|$$

$$= |E_{D_s}[\tilde{\sigma}(X) \cdot (Y - p(X))] + E_{D_s}[p(X)\,(\tilde{\sigma}(X) - e_{st}(X))]|$$

By the triangle inequality,

$$= |E_{D_s}[\tilde{\sigma}(X) \cdot (Y - p(X))]| + |E_{D_s}[p(X)\,(\tilde{\sigma}(X) - e_{st}(X))]|$$

By the Cauchy-Schwarz inequality,

$$\leq E_{D_s}[|Y - p(X)|] + |E_{D_s}[p(X)\,(\tilde{\sigma}(X) - e_{st}(X))]|$$

By multi-accuracy, the second term is bounded by α and the first is defined as $\Delta p_s(p)$, so

$$\leq \Delta_{p_s}(p) + \alpha$$

Now that we have shown $\mu_t^{PS}(\tilde{\sigma})$ is close to $\mu_t(p)$ for all $p \in \mathcal{P}$, we can bound the excess error of $\mu_t^{PS}(\tilde{\sigma})$ to prove the theorem. Fix an arbitrary $p \in \mathcal{P}$.

$$er_t(\mu_t^{PS}(\tilde{\sigma})) = |\mu_t^{PS}(\tilde{\sigma}) - \mu_t^*|$$

$$\leq |\mu_t^{PS}(\tilde{\sigma}) - \mu_t(p)| + |\mu_t(p) - \mu_t^*|$$
$$\leq er_t(\mu_t(p)) + \Delta_{p_s}(p) + \alpha$$

Because this holds for all $p \in \mathcal{P}$, this proves the theorem. □

13.3 Overlap with Ongoing Research

I believe the result in Theorem 1 is new, but a similar result was shown in Kim et al. [2022] (where the roles of $\mathcal{P}$ and Σ are flipped). Besides that, the theoretical results I reference are not new, but I try to connect them in a unifying framework. As discussed, the data application is an extension of that in Kim et al. [2022].