A Calibration Approach to Transportability with Observational Data

Kevin P. Josey[1], Fan Yang[2], Debashis Ghosh[2,*], and Sridharan Raghavan[3]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

[2]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO

[3]Department of Veterans Affairs Eastern Colorado Health Care System, Aurora, CO

[*]Corresponding Author: Debashis Ghosh, debashis.ghosh@cuanschutz.edu

August 18, 2020

Abstract

An important consideration in clinical research studies is proper evaluation of internal and external validity. While randomized clinical trials can overcome several threats to internal validity, they may be prone to poor external validity. Conversely, large prospective observational studies sampled from a broadly generalizable population may be externally valid, yet susceptible to threats to internal validity, particularly confounding. Thus, methods that address confounding and enhance transportability of study results across populations are essential for internally and externally valid causal inference, respectively. We develop a weighting method which estimates the effect of an intervention on an outcome in an observational study which can then be transported to a second, possibly unrelated target population. The proposed methodology employs calibration estimators to generate complementary balancing and sampling weights to address confounding and transportability, respectively, enabling valid estimation of the target population average treatment effect. A simulation study is conducted to demonstrate the advantages and similarities of the calibration approach against alternative techniques. We also test the performance of the calibration estimator-based inference in a motivating real data example comparing whether the effect of biguanides versus sulfonylureas - the two most common oral diabetes medication classes for initial treatment - on all-cause mortality described in a historical cohort applies to a contemporary cohort of US Veterans with diabetes.

1 Introduction

Two common and related problems in statistics involve causal inference and generalizing study results to a population of interest. For the former, the principal barrier is confounding associated with the exposure or treatment variable. One solution might be to conduct a randomized trial, but a randomized trial is impractical in many scientific and medical contexts. Therefore, methods for causal inference from observational data are essential. Even when valid causal inference can be drawn from a study, generalization is often limited by the difficulty of randomly sampling from the population of interest. The population that is sampled from in an observational study or randomized trial - i.e. the *study population* - might diverge from the specific population of interest for applying the study results – i.e. the *target population*. For example, the population of patients with a given disease may differ across important characteristics from the population receiving a specific intervention for that disease. In this simplistic scenario, extending valid inferences which evaluate the efficacy of the intervention to the targeted population containing every patient with the disease may be of interest. Throughout this manuscript we focus on a particular yet well-defined setup for generalizing results onto a target population known as transportability (Pearl and Bareinboim, 2014).

Methods for causal inference and transporting effect estimates to a target population have been a recent topic of much interest in the statistical literature. Our goal is to combine methods found in these two respective areas in order to minimize bias due to confounding, which is unavoidable in observational studies, and to account for differences between study and target populations that could influence the causal effect estimate on the population of interest. The propensity score, or the probability of exposure given a set of measured covariates, has emerged as a popular tool in causal inference as the basis for balancing the distribution of confounders between the exposed and unexposed participants in an observational study (Rosenbaum and Rubin, 1983). A closely related approach models the probability of being sampled for a study as opposed to the probability of receiving a treatment in order to transport the results of a randomized trial onto a target population (Westreich et al., 2017). Extensions and comparisons of these methods for transporting results from an observational study are more limited, with most methods opting to focus on transporting results from a randomized controlled trial onto a population characterized by an observational cohort. Moreover, there is limited discussion on extensions for many of these methods to reconcile multiple observational studies which examine the same exposure/outcome relationship - a closely related problem which we will refer to as *data-fusion* (Bareinboim and Pearl, 2016). Further compounding the issue, fitting parametric models of the probability of treatment and inverse odds of sampling with maximum likelihood estimation has several limitations in cases where model misspecification is rife and leads to problems in finite samples (Kang and Schafer, 2007). Therefore, estimation procedures that complement new methodological developments that overcome these limitations would be desirable when transporting causal effect estimates from observational studies.

Calibration estimators (Deville and Särndal, 1992) have had recent success in both transporting causal effects from a randomized controlled trial to a target population and estimating causal effects using obser-

vational data. We propose combining the approach of Chan et al. (2015), which finds *balancing weights* that correct for treatment group heterogeneity, with an exponential tilting function that estimates the *sampling weights*, which removes bias present between the study participants and non-participants sampled from the target population (Signorovitch et al., 2010). A similar exponential tilting function was used to generalize results of a randomized controlled trial onto an observational cohort by Dong et al. (2020). Their solution estimates the target population average treatment effect using a class of estimators that augment the treatment assignment and sampling indicator models with an outcome model (Robins et al., 1994). By contrast, we avoid any attempt to identify the outcome process and instead emphasize efforts to identify the true sampling and treatment processes in the spirit of Rubin (2008). In our experience, the exercise of identifying potential confounders that contribute to sampling and treatment selection bias is more straightforward than identifying the potential variables that affect the outcome in many statistical applications. If there is some conflict in attempting to specify both models correctly, then correct specification of the sampling and treatment assignment models should take priority over correctly specifying the outcome model. Regardless of the design choices, we show that our so-called full calibration approach is doubly-robust, meaning that if either the outcome model is correctly specified or the probability of treatment assignment and the probability of being sampled are correctly specified, then the average treatment effect estimator is consistent. We also show how the full calibration approach can be adapted to solve the data-fusion problem of combining two observational studies that compare the same exposure and response.

In addition to the proposed full calibration approach, we adapt and examine two other methods for transporting causal effect estimates using observational data. One of these methods is the augmented estimator proposed by Dong et al. (2020), which we have already mentioned. The other method uses the targeted maximum likelihood estimation framework and is discussed by Rudolph and van der Laan (2017). Each of these methods are doubly-robust given certain assumptions about the data generating processes, although they require differing degrees of parametric assumptions to achieve this property. The targeted maximum likelihood approach makes the fewest parametric assumptions while the full calibration approach makes the most parametric assumptions. Having fewer parametric assumptions allows for more intricate modeling choices, including machine learning techniques. However, as opposed to the targeted maximum likelihood and augmented approaches, the parametric nature of the full calibration approach allows us to relax the assumption of propensity score exchangeability. Propensity score exchangeability assumes the propensity score is the same across populations. We acknowledge that no single modeling approach will perform best universally in any given observational study but are nevertheless interested in identifying scenarios where one approach may be better suited than the others. With these considerations in mind, a summary of our primary aims are as follows:

- Derive a full calibration approach for transporting observational study results across populations while relaxing the requirement of propensity score exchangeability;
- Extend two other doubly-robust methods for transporting treatment effect estimates to accommodate

observational data while varying the parametric assumptions about the data generating processes;

- Examine extensions of methods for transportability to solve the problem of data-fusion;
- Compare the three doubly-robust methods for transportability identified throughout in a simulation study.
- Apply the full calibration approach to a study comparing mortality rates for diabetic patients prescribed either a sulfonylurea or metformin monotherapy.

The remainder of this article is structured as follows. In Section 1.1 is a motivating dataset that we will analyze in Section 6 to provide some context to the problem of transportability. In Section 2 we introduce the notation, assumptions, and previous methods for transporting causal effects, which are adapted to make use of observational data. In Section 4 we introduce a calibration approach to transporting results from observational studies. In Section 5, we compare the full calibration approach with the methods we will have described in Section 2. Section 6 contains a data analysis of the illustrative example from the dataset introduced in Section 1.1. Finally, we conclude with a discussion in Section 7.

1.1 Motivating Dataset

To help showcase the solutions we propose, we will address two aspects of treatment of type 2 diabetes for which there is limited head-to-head clinical trial data. First, we will compare the effectiveness of monotherapy with metformin and sulfonylureas as first-line treatment for type 2 diabetes mellitus in a cohort of patients receiving care in the US Veterans Affairs Healthcare System (VA). Metformin is a member of the biguanide class of oral diabetes medications and is the most commonly used initial treatment for type 2 diabetes in the United States (Desai et al., 2012; Berkowitz et al., 2014; Hampp et al., 2014). Sulfonylureas are a class of oral diabetes medications that comprise the next most commonly used initial treatments for type 2 diabetes in the United States.

Despite their long-time use for the treatment of type 2 diabetes, head-to-head comparisons have been inconclusive regarding the effects of metformin and sulfonylureas on clinical outcomes such as cardiovascular events and mortality. Results of observational studies have conflicted, and meta-analyses have drawn attention to risk of bias in the observational studies (Azoulay and Suissa, 2017). That said, several of the larger observational studies considered at low-risk of bias reach a similar conclusion: that sulfonylurea use is associated with higher mortality and cardiovascular risk than metformin (Schramm et al., 2011; Roumie et al., 2012; Wheeler et al., 2013). Similarly, randomized trials have been contradictory, though most were not designed to address a direct comparison of the two oral diabetes medication classes (Varvaki Rados et al., 2016). As with the observational studies, however, the randomized trials raise a safety concern regarding the use of sulfonylureas, particularly with regard to cardiovascular disease and cardiovascular mortality, though not always reaching the prespecified threshold for statistical significance (Hong et al., 2013; Varvaki Rados

et al., 2016). The development and approval of a number of new type 2 diabetes medications in the last several years with particular benefit in individuals with or at high-risk for cardiovascular disease (Zinman et al., 2015; Marso et al., 2016a,b; Neal et al., 2017; Holman et al., 2017) has diversified the treatment options available to patients and providers, prompting a reevaluation of the comparative effectiveness of metformin and sulfonylureas. Moreover, the population of diabetes patients has changed temporally with regard to cardiovascular disease prevalence and risk factor control (Ali et al., 2013; Selvin et al., 2014; Gregg et al., 2014; Geiss et al., 2014; Gregg et al., 2018; Cheng et al., 2018; Raghavan et al., 2019), which could affect associations of metformin and sulfonylureas with outcomes in contemporary diabetes patient cohorts.

As a basis for analyses in this paper, we use Wheeler et al. (2013), which found that US military veterans with diabetes receiving sulfonylurea treatment were at higher risk of mortality than patients receiving metformin after adjusting for potential confounders. Even before the publication of these results, the use of sulfonylureas began to decrease dramatically within VA hospitals due to the recognition that metformin had lower marginal risk of adverse outcomes as compared to sulfonylureas in a more general population (Johnson et al., 2002). Concurrent with this reduction in sulfonylurea use were the aforementioned changes in the population of veterans receiving care for diabetes, particularly with regard to cardiovascular disease prevalence and risk factor control (Ali et al., 2013; Selvin et al., 2014; Gregg et al., 2014; Geiss et al., 2014; Gregg et al., 2018; Cheng et al., 2018; Raghavan et al., 2019). Using data available from 2004-2009, which overlaps with the cohort analyzed by (Wheeler et al., 2013), we will transport the risk difference of mortality among newly diagnosed patients receiving initial monotherapy with either a sulfonylurea or metformin to a more contemporary cohort of diabetes patients diagnosed between 2010-2014. We also find the data-fusion estimates of the 2010-2014 risk difference using the combined 2004-2009 and 2010-2014 cohorts. Patients initialized between 2004 and 2009 fall into a date range when sulfonylurea monotherapy was a more widely prescribed method of treatment. The subsequent years saw a dramatic decrease in sulfonylurea use as a monotherapy. The decrease in sulfonylurea is an obvious signal of a propensity score exchangeability violation, thus requiring more flexible methods such as the full calibration approach that can accommodate such circumstances.

Splitting this dataset into separate cohorts has two main advantages for illustrative purposes. First, we can obtain consistent effect estimates on the 2010-2014 sample without using data from 2004-2009. This estimate will provide a benchmark for both the transported estimate from the 2004-2009 cohort and for the data-fusion estimate. The second advantage has to do with the structure of the data. The difference between cohorts will be small given the close proximity of the time intervals, ensuring that the sampling positivity assumption holds (defined in Section 2.2). Both cohorts belong to the same *superpopulation*, i.e. newly diagnosed veterans receiving care at the VA, with the only differences being related to the temporal trends observed in the covariates. Further adding to the strong overlap and completeness of the data, both cohorts contain a large number of patients lending to more accurate estimates.

As a counterpoint to the example transporting results within the VA system across two temporally distinct

cohorts, we will also apply the methods developed in this paper to transport results from a randomized trial comparing two diabetes treatment strategies, the Bypass Angioplasty Revascularization Investigation 2 Diabetes (BARI 2D) trial, to the aforementioned 2010-2014 cohort of veterans with diabetes (The BARI 2D Study Group (2009)). The BARI 2D trial attempted to address a second pressing knowledge gap pertaining to type 2 diabetes treatment: whether diabetes patients with coronary artery disease (CAD), the most common underlying cause of mortality for diabetes patients (Rao Kondapally Seshasai et al., 2011), benefited from an insulin sensitization or insulin provision strategy for treating their diabetes. Thus, the BARI 2D trial randomized diabetes patients with known CAD to receive either an insulin sensitization strategy (largely treatment with metformin) or an insulin provision strategy (treatment with sulfonylureas and/or insulin). Given the importance of optimal glycemic management in the particularly high-risk population of diabetes patients with CAD, evaluating transportability of results from BARI 2D to a real-world population of VA diabetes patients could provide applicable insight into diabetes population health management within the VA health system.

2 Setting and Preliminaries

2.1 Notation and Definitions

The setup for transportability and data-fusion with observational data requires - first and foremost - data from two separate observational studies. Define $S_i \in \{0, 1\}$ as a sampling indicator denoting whether the independent sampling unit $i = 1, 2, \ldots, n$ is a study non-participant or participant. We will refer to units $i \in \{i : S_i = 1\}$ as sample A and units $i \in \{i : S_i = 0\}$ as sample B. We denote $n_1 = \sum_{i=1}^{n} S_i$ and $n_0 = \sum_{i=1}^{n}(1 - S_i)$ with $n = n_1 + n_0$. We suppose that the non-participants in sample B represent a random sample from the target population, the population we would like to infer upon, whereas sample A is a representative sample of the study population.

For each $i = 1, 2, \ldots, n$, let $\mathbf{X}_i \in \mathcal{X}$ denote a vector of measured covariates, $Y_i \in \Re$ denote the outcome, and $Z_i \in \{0, 1\}$ denote the treatment assignment. We employ the potential outcomes framework (Rubin, 1974) to construct the causal estimand of interest and the assumptions for transportability (Lesko et al., 2017). Let $Y_i(0)$ denote the potential outcome when $Z_i = 0$ and $Y_i(1)$ denote the potential outcome when $Z_i = 1$. This means the observed outcome is equivalent to $Y_i \equiv Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. The target population average treatment effect is defined as $\tau_0 \equiv \mathbb{E}[Y_i(1) - Y_i(0)|S_i = 0]$.

Conditioned on $\mathbf{X}_i$, we set $\rho(\mathbf{X}_i) \equiv \Pr\{S_i = 1|\mathbf{X}_i\}$, $\pi_1(\mathbf{X}_i) \equiv \Pr\{Z_i = 1|\mathbf{X}_i, S_i = 1\}$, and $\pi_0(\mathbf{X}_i) \equiv \Pr\{Z_i = 1|\mathbf{X}_i, S_i = 0\}$ for all $i = 1, 2, \ldots, n$. Note that the probability of treatment conditioned on the sample indicator and covariates can be alternatively expressed as

$$\pi(S_i, \mathbf{X}_i) \equiv S_i \pi_1(\mathbf{X}_i) + (1 - S_i)\pi_0(\mathbf{X}_i).$$

Define $\{c_j(\mathbf{X}); j = 1, 2, \ldots, m\}$ as the set of functions that generate linearly independent features to be

balanced between treatment groups and the samples A and B. We will refer to these quantities as *balance functions*. Furthermore, we will assume $c_1(\mathbf{X}_i) = 1$ for all $i = 1, 2, \ldots, n$ throughout. The target sample moments of the balance functions are defined as $\hat{\theta}_j = n_0^{-1} \sum_{i=1}^{n} (1 - S_i) c_j(\mathbf{X}_i)$, which is a consistent estimator for $\theta_j \equiv \mathbb{E}[c_j(\mathbf{X}_i)|S_i = 0]$ for all $j = 1, 2, \ldots, m$.

2.2 Assumptions for Transportability and Data-Fusion

Under the potential outcomes model and given the definitions listed in the previous section, we may begin to develop the setting for which transportability and data-fusion are feasible using observational data (Pearl and Bareinboim, 2014; Bareinboim and Pearl, 2016). We frame the setup to both problems through the following set of assumptions which require several conditions which the data generating mechanisms for $(S_i, \mathbf{X}_i, Y_i, Z_i)$ must satisfy. These assumptions are an extension to those proposed in other articles regarding the transportability of experimental results across populations (Rudolph and van der Laan, 2017; Dong et al., 2020). We combined these assumptions with the assumptions necessary for conducting causal inference in the presence of treatment group heterogeneity (Rubin, 1974).

Assumption 1 (Strongly Ignorable Treatment Assignment)**.** *The potential outcomes among both the study participants and the study non-participants are independent of the treatment assignment given* $\mathbf{X}_i$*:*

$$[Y_i(0), Y_i(1)]^T \perp\!\!\!\perp Z_i | (\mathbf{X}_i, S_i) \text{ for all } i = 1, 2, \ldots, n.$$

Assumption 2 (Mean Exchangeability)**.** *Among all independent sampling units in either sample A or B, the expected value of the potential outcomes conditioned on the covariates are exchangeable:* $\mathbb{E}[Y_i(1)|\mathbf{X}_i, S_i] = \mathbb{E}[Y_i(1)|\mathbf{X}_i]$ *and* $\mathbb{E}[Y_i(0)|\mathbf{X}_i, S_i] = \mathbb{E}[Y_i(0)|\mathbf{X}_i]$ *for all* $i = 1, 2, \ldots, n$*.*

Assumption 3 (Sampling Positivity)**.** *The probability of study participation, conditioned on the baseline covariates necessary to ensure Assumption 2, is bounded away from zero and one:*

$$0 < \Pr\{S_i = 1|\mathbf{X}_i\} < 1 \text{ for all } i = 1, 2, \ldots, n.$$

Assumption 4 (Treatment Positivity)**.** *The probability of treatment conditioned on the baseline covariates, and given the two samples, is bounded away from zero and one:*

$$0 < \Pr\{Z_i = 1|\mathbf{X}_i, S_i\} < 1 \text{ for all } i = 1, 2, \ldots, n.$$

The distinction between transportability and data-fusion essentially amounts to how much data we are provided from sample A and sample B. For problems of transportability, we require the complete individual-level data from sample A, but only the individual-level covariate data from sample B (i.e. $\mathbf{X}_i$ for all $i \in \{i : S_i = 0\}$). There is another setting for transportability wherein we only require the sample moments of the covariates from sample B. However, this scenario is fraught with challenges, particularly for inference involving a population-level estimand (Josey et al., 2020a). We leave any further discussion of this setting for

Section 7. In data-fusion, both samples A and B provide $(\mathbf{X}_i, Y_i, Z_i)$ for all $i = 1, 2, \ldots, n$. It should not be a surprise that the latter setting is more powerful given the additional data. However, in many data analysis applications, Y_i and Z_i are not available from sample B, leaving data-fusion infeasible yet transportability as an appealing alternative.

In addition to Assumptions 2-4, the following set of assumptions is needed to establish the double-robustness property of the proposed full calibration estimator. For more context, we will show that if either Assumption 5 is satisfied or both Assumptions 6 and 7 hold, then the full calibration estimator proposed in Section 4 is consistent.

Assumption 5 (Conditional Linearity for the Potential Outcomes)**.** *The expected value of the potential outcomes, conditioned on* $\mathbf{X_i}$*, is linear across the span of the covariates:* $\mathbb{E}[Y_i(1)|\mathbf{X}_i] = \sum_{j=1}^{m} c_j(\mathbf{X}_i)\beta_j$ *and* $\mathbb{E}[Y_i(0)|\mathbf{X}_i] = \sum_{j=1}^{m} c_j(\mathbf{X}_i)\alpha_j$ *for all* $i = 1, 2, \ldots, n$ *and* $\alpha_j, \beta_j \in \Re$ *for all* $j = 1, 2, \ldots, m$*.*

Assumption 6 (Conditional Linear Log-Odds for Sampling)**.** *The log-odds of being in sample A versus sample B are linear across the span of the covariates:* $logit[\rho(\mathbf{X}_i)] = \sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j$ *for all* $i = 1, 2, \ldots, n$ *and* $\gamma_j \in \Re$ *for all* $j = 1, 2, \ldots, m$*.*

Assumption 7 (Conditional Linear Log-Probability for Treatment)**.** *The log-probability of treatment in sample A and B are linear across the span of the covariates:*

$$log[\pi(S_i, \mathbf{X}_i)] = S_i \sum_{j=1}^{m} c_j(\mathbf{X}_i)\delta_{j1} + (1 - S_i) \sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j1}$$

and

$$log[1 - \pi(S_i, \mathbf{X}_i)] = S_i \sum_{j=1}^{m} c_j(\mathbf{X}_i)\delta_{j0} + (1 - S_i) \sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j0}$$

for all $i = 1, 2 \ldots, n$ *where* $\delta_{j0}, \delta_{j1}, \lambda_{j0}, \lambda_{j1} \in \Re$ *for all* $j = 1, 2, \ldots, m$*.*

The alternative methods presented in Section 3 require similar assumptions, although we will be able to relax the linearity conditions found in Assumptions 5-7. We will see later on that these alternative methods require an additional assumption regarding the scenario when the outcome model is misspecified in order to remain consistent. This additional consideration is referenced in Assumption 8. Propensity score exchangeability is not required for our proposed method in lieu of Assumptions 6 and 7. Verifying this assumption is not so much of an issue for transporting observational study results since Z_i should not exist for any $i \in \{i : S_i = 0\}$, allowing for speculation about $\pi_0(\mathbf{X}_i)$. In this scenario, it is natural to assume that treatment assignment is the same in both samples. In data-fusion, where Z_i is observed for all $i = 1, 2, \ldots, n$ however, this assumption is often violated.

Assumption 8 (Propensity Score Exchangeability)**.** *For all* $\mathbf{X}_i \in \mathcal{X}$*, we have* $\pi_1(\mathbf{X}_i) = \pi_0(\mathbf{X}_i)$*.*

3 Previous Methods for Transportability Extended for Observational Data

Targeted maximum likelihood estimation (TMLE) has emerged as a flexible framework for estimating a variety of causal estimands (van der Laan and Rubin, 2006). Specifically, Rudolph and van der Laan (2017) apply this framework to estimate τ_0 within the transportability setting described in Section 2.2. TMLE is solved in an iterative manner by initially finding $\hat{\mu}_1(\mathbf{X}_i)$ and $\hat{\mu}_0(\mathbf{X}_i)$ using the independent sampling units $i \in \{i : S_i = 1\}$, which estimate $\mu_1(\mathbf{X}_i)$ and $\mu_0(\mathbf{X}_i)$, respectively. If we were to stop here, we could formulate the G-computation approach for estimating τ_0 by solving for

$$\hat{\tau}_{\text{G}} = \frac{1}{n_0} \sum_{\{i:S_i=0\}} \left[\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)\right].$$

If we can show $\hat{\mu}_1(\mathbf{X}_i) \to_p \mu_1(\mathbf{X}_i)$ and $\hat{\mu}_0(\mathbf{X}_i) \to_p \mu_0(\mathbf{X}_i)$, then $\hat{\tau}_G \to_p \tau_0$. Rudolph and van der Laan (2017) extend this intuitive approach in an attempt to account for potential bias induced from misspecifying $\hat{\mu}_1(\mathbf{X}_i)$ and $\hat{\mu}_0(\mathbf{X}_i)$. Their solution updates the estimates of the conditional means of the potential outcome using consistent estimates of $\rho(\mathbf{X}_i)$ and $\pi_1(\mathbf{X}_i)$, which we denote as $\hat{\rho}(\mathbf{X}_i)$ and $\hat{\pi}_1(\mathbf{X}_i)$. The estimators $\hat{\rho}(\mathbf{X}_i)$ and $\hat{\pi}_1(\mathbf{X}_i)$ are combined into a so-called clever covariate (Schuler and Rose, 2017) to find

$$\begin{aligned} \tilde{\mu}_0(\mathbf{X}_i) &= \hat{\mu}_0(\mathbf{X}_i) + \hat{\epsilon}_0 \frac{(1 - Z_i)[1 - \hat{\rho}(\mathbf{X}_i)]}{\hat{\rho}(\mathbf{X}_i)[1 - \hat{\pi}_1(\mathbf{X}_i)]} \\ \tilde{\mu}_1(\mathbf{X}_i) &= \hat{\mu}_1(\mathbf{X}_i) + \hat{\epsilon}_1 \frac{Z_i[1 - \hat{\rho}(\mathbf{X}_i)]}{\hat{\rho}(\mathbf{X}_i)\hat{\pi}_1(\mathbf{X}_i)}. \end{aligned} \tag{1}$$

Estimates of ϵ_0 and ϵ_1 are found by regressing the clever covariate onto the outcome with the initial mean predictions serving as offsets among the units $i \in \{i : S_i = 1\}$. The final TMLE estimate of τ_0 has a similar form to the G-computation setup, which involves solving for

$$\hat{\tau}_{\text{TMLE}} = \frac{1}{n_0} \sum_{\{i:S_i=0\}} \left[\tilde{\mu}_1(\mathbf{X}_i) - \tilde{\mu}_0(\mathbf{X}_i)\right],$$

while ignoring the indicators Z_i that appear in the clever covariates of (1). Given assumptions 1-4, Rudolph and van der Laan (2017) show that the TMLE estimator is doubly-robust if either $\hat{\mu}_1(\mathbf{X}_i) \to_p \mu_1(\mathbf{X}_i)$ and $\hat{\mu}_0(\mathbf{X}_i) \to_p \mu_0(\mathbf{X}_i)$ or $\hat{\rho}(\mathbf{X}_i) \to_p \rho(\mathbf{X}_i)$ and $\hat{\pi}_1(\mathbf{X}_i) \to_p \pi_0(\mathbf{X}_i)$. For the latter scenario, an implicit assumption for the method to achieve double-robustness is Assumption 8.

Another doubly-robust method originally intended for generalizing experimental data is an augmented estimator which combines a treatment, sampling, and outcome model, similar to the clever covariates found in TMLE (Dong et al., 2020). We present the augmented approach with slight alterations to the estimator presented by Dong et al. (2020) so as to be relevant for the transportability setting. The setup to the problem solved by Dong et al. (2020) assumes that each unit in the target sample is prescribed a vector of known sampling weights. This in turn facilitates inference on the combined population containing samples A and B (i.e. the target population). Our setup to the problem, on the other hand, assumes that the target

sample, i.e. sample B, is drawn uniformly from a target population. The distinction can be drawn from the implication that the target population may differ from the superpopulation. The problem they describe is more akin to generalizability (Cole and Stuart, 2010) over a finite population whereas our focus is on transportability within a superpopulation framework.

The augmented approach to transporting observational study results proceeds by deriving estimators for the component models that generate (S_i, Y_i, Z_i) given Assumptions 1-4. While several estimators for $\hat{\pi}_1(\mathbf{X}_i)$, $\hat{\mu}_0(\mathbf{X}_i)$, and $\hat{\mu}_1(\mathbf{X}_i)$ will suffice, the inverse odds of sampling are specifically estimated by solving the Lagrangian dual,

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in \Re^m}{\arg\max} \sum_{\{i:S_i=1\}} \left\{ -\exp\left[-S_i \sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j \right] - \sum_{j=1}^{m} \hat{\theta}_j \gamma_j \right\}, \tag{2}$$

where $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2, \ldots, \gamma_m)^T$. A Lagrangian dual is an unconstrained optimization objective derived by applying the Lagrangian multiplier theorem to a constrained convex optimization problem. In the case of (2), the constrained optimization, or primal problem, seeks to

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} q(\mathbf{X}_i) \log\left[q(\mathbf{X}_i)\right] - q(\mathbf{X}_i) \\ \text{subject to} \quad & \sum_{i=1}^{n} S_i q(\mathbf{X}_i) c_j(\mathbf{X}_i) = \sum_{i=1}^{n} (1 - S_i) c_j(\mathbf{X}_i) \text{ for all } j = 1, 2, \ldots, m. \end{aligned} \tag{3}$$

The dual solution to (2) can be found using Lagrangian multipliers to be

$$\hat{q}(\mathbf{X}_i) = \exp\left[-S_i \sum_{j=1}^{m} c_j(\mathbf{X}_i)\hat{\gamma}_j \right] \tag{4}$$

for all $i = 1, 2, \ldots, n$, which is the primal solution to (3) (i.e. the sampling weight estimates). The estimated sampling weights have the property that $\sum_{i=1}^{n} S_i \hat{q}(\mathbf{X}_i) c_j(\mathbf{X}_i) = \sum_{i=1}^{n} (1 - S_i) c_j(\mathbf{X}_i)$ for all $j = 1, 2, \ldots, m$. In other words, the weighted sample moments of the balance functions in sample A are equal to the unweighted sample moments of the balance functions in sample B. This estimator of the inverse odds of sampling weights was suggested by Signorovitch et al. (2010) as a pre-processing step for comparing outcomes in two clinical trials when sample heterogeneity is present.

Given the estimated sampling weights, the conditional mean estimates of the potential outcomes, and the probability of treatment model, Dong et al. (2020) construct an augmented estimator for the target population average treatment effect which solves for

$$\hat{\tau}_{\text{AUG}} = \frac{1}{n_1} \sum_{\{i:S_i=1\}} \hat{q}(\mathbf{X}_i) \left[\frac{Z_i[Y_i - \hat{\mu}_1(\mathbf{X}_i)]}{\hat{\pi}_1(\mathbf{X}_i)} - \frac{(1 - Z_i)[Y_i - \hat{\mu}_1(\mathbf{X}_i)]}{1 - \hat{\pi}_1(\mathbf{X}_i)} \right] + \frac{1}{n_0} \sum_{\{i:S_i=0\}} \left[\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i) \right]. \tag{5}$$

Under assumptions 1-4, the augmented estimator $\hat{\tau}_{\text{AUG}}$ is shown to be doubly-robust by Dong et al. (2020). We can easily see this result given the following heuristic. In the scenario where $\hat{\mu}_0(\mathbf{X}_i) \to_p \mu_0(\mathbf{X}_i)$ and $\hat{\mu}_1(\mathbf{X}_i) \to_p \mu_1(\mathbf{X}_i)$, and given Assumption 2, the first sum in (5) has an expected value of zero while the second sum is consistent for τ_0 as $n \to \infty$. In the scenario where $\hat{\pi}_1(\mathbf{X}_i) \to \pi_0(\mathbf{X}_i)$ and $\hat{q}(\mathbf{X}_i) \to$

$[1-\rho(\mathbf{X}_i)]\rho(\mathbf{X}_i)^{-1}$, which requires Assumptions 6 and 8 to hold, then the first sum in (5) is consistent for the bias produced by the second sum as n approaches infinity.

Through Assumption 2, and given that Y_i and Z_i are observed for all $i = 1, 2, \ldots, n$, Dong et al. (2020) extend their estimator to use all available data by estimating $\hat{\mu}_0(\mathbf{X}_i)$ and $\hat{\mu}_1(\mathbf{X}_i)$ over all $i = 1, 2, \ldots, n$. The result provides a solution to the data-fusion problem without changing the estimator in (5). A similar approach is considered by Lu et al. (2019) for generalizing experimental results. As we have previously mentioned, a major issue with the augmented and TMLE approaches for integrating two datasets is the requirement of Assumption 8. In scenarios where this assumption is violated, the issue is most apparent within the data-fusion setting whenever the outcome model is misspecified. Without propensity score exchangeability, i.e. $\pi_1(\mathbf{X}_i) \neq \pi_0(\mathbf{X}_i)$, the first summation in (5) will not consistently estimate the bias induced by the second summation, even when the sampling model is correctly specified. There may be several workarounds to this issue using the augmented estimator. However, they may require stronger assumptions to the underlying models for S_i and Z_i, like those of Assumptions 6 and 7. We defer further discussion of this issue to Section 7.

4 A Full Calibration Approach to Transportability

Our solution to the problem of transporting observational study results combines the calibration estimator approach of Chan et al. (2015), which finds balancing weights that correct for treatment group heterogeneity, with a vector of estimated sampling weights, which removes bias induced by the differences of the covariate distribution between samples A and B (Signorovitch et al., 2010). In other words, we estimate both a vector of balancing weights and a vector of sampling weights which, when estimated in tandem, allow for consistent estimation of the target population average treatment effect. A quick breakdown of the procedure is as follows. First, we balance the study and target sample covariate moments by estimating the sampling weights. Second, we estimate balancing weights on sample A given the estimated sampling weights from the previous step. Finally, we estimate the target population average treatment effect using a Horvitz-Thompson type estimator (Horvitz and Thompson, 1952).

Fortunately, the first step of estimating the sampling weights is solved in the exact same manner suggested in Signorovitch et al. (2010) by finding (2) and (4). The second step is to compute balancing weights that mitigate treatment group heterogeneity by solving for

$$\begin{aligned}
\hat{\boldsymbol{\lambda}}_0 &= \underset{\boldsymbol{\lambda}\in\Re^m}{\arg\max} \sum_{\{i:S_i=1\}} \left\{ -\hat{q}(\mathbf{X}_i) \sum_{j=1}^{m} \hat{\theta}_j \lambda_j - \hat{q}(\mathbf{X}_i) \exp\left[-(1-Z_i) \sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_j \right] \right\} \\
\hat{\boldsymbol{\lambda}}_1 &= \underset{\boldsymbol{\lambda}\in\Re^m}{\arg\max} \sum_{\{i:S_i=1\}} \left\{ -\hat{q}(\mathbf{X}_i) \sum_{j=1}^{m} \hat{\theta}_j \lambda_j - \hat{q}(\mathbf{X}_i) \exp\left[-Z_i \sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_j \right] \right\}
\end{aligned} \tag{6}$$

where $\boldsymbol{\lambda}_0 \equiv (\lambda_{10}, \lambda_{20}, \ldots, \lambda_{m0})^T$, and $\boldsymbol{\lambda}_1 \equiv (\lambda_{11}, \lambda_{21}, \ldots, \lambda_{m1})^T$. The resulting balancing weights are esti-

mated with

$$\hat{p}(\mathbf{X}_i) = \hat{q}(\mathbf{X}_i) \exp\left[-(1 - Z_i)\sum_{j=1}^{m} c_j(\mathbf{X}_i)\hat{\lambda}_{j0} - Z_i \sum_{j=1}^{m} c_j(\mathbf{X}_i)\hat{\lambda}_{j1}\right] \tag{7}$$

for all $i \in \{i : S_i = 1\}$. To estimate the target population average treatment effect, we use a Horvitz-Thompson type estimator (Horvitz and Thompson, 1952) which solves for

$$\hat{\tau}_{\text{CAL}} = \sum_{\{i:S_i=1\}} \frac{\hat{p}(\mathbf{X}_i)(2Z_i - 1)Y_i}{\sum_{\{i:S_i=1\}} \hat{p}(\mathbf{X}_i)Z_i}. \tag{8}$$

Recall that for transportability we are presented with $(\mathbf{X}_i, Y_i, Z_i)$ for sample A, but only $\mathbf{X}_i$ in sample B. Hence the index set $\{i : S_i = 1\}$ for the summations in (6)-(8).

Much like with the primal-dual relationship between (2) and (3), the dual problem in (6) corresponds to the following primal problem:

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n} p(\mathbf{X}_i) \log\left[\frac{p(\mathbf{X}_i)}{\hat{q}(\mathbf{X}_i)}\right] - p(\mathbf{X}_i) + \hat{q}(\mathbf{X}_i) \\
\text{subject to} \quad & \sum_{i=1}^{n} S_i(1 - Z_i)p(\mathbf{X}_i)c_j(\mathbf{X}_i) = \sum_{i=1}^{n} S_i\hat{q}(\mathbf{X}_i)c_j(\mathbf{X}_i) \text{ and} \\
& \sum_{i=1}^{n} S_i Z_i p(\mathbf{X}_i)c_j(\mathbf{X}_i) = \sum_{i=1}^{n} S_i\hat{q}(\mathbf{X}_i)c_j(\mathbf{X}_i) \text{ for all } j = 1, 2, \ldots, m.
\end{aligned} \tag{9}$$

The reasoning for using the relative entropy objective function in (9) is its ability to produce balancing weights which have the same functional form as the inverse probability of treatment weights when the probability of treatment is modeled with a log-linear model. If we suppose that Assumption 6 holds, then (2) and (4) is an unbiased estimator for $[1 - \rho(\mathbf{X}_i)]\rho(\mathbf{X}_i)^{-1}$. Coupled with this result and given Assumption 7, (6)-(7) produces unbiased estimates for $\pi_0(\mathbf{X}_i)$ and $1 - \pi_0(\mathbf{X}_i)$. Another calibration weighting estimator can be found by replacing the relative entropy distance in (9) with the shifted relative entropy (Josey et al., 2020b). The resulting balancing weights using this distance are analogous to the calibration version of the inverse probability of treatment weights, which assume the probability of treatment follows a logistic model. However, we found the approach using (6) and (7) to be more stable and computationally faster.

The balancing weights from (7) represent the generalized projection of the objective entropy curve onto the linear hyperspace that satisfies the condition $\sum_{i=1}^{n} S_i Z_i \hat{p}(\mathbf{X}_i)c_j(\mathbf{X}_i) = \sum_{i=1}^{n} S_i\hat{q}(\mathbf{X}_i)c_j(\mathbf{X}_i)$ for all $j = 1, 2, \ldots, m$. Consequently, the balancing weights achieve exact balance between the treatment-specific sample moments of the covariate distribution in sample A and the covariate sample moments of sample B because $\sum_{i=1}^{n} S_i\hat{q}(\mathbf{X}_i)c_j(\mathbf{X}_i) = \sum_{i=1}^{n}(1 - S_i)c_j(\mathbf{X}_i)$. Combining the results of the sampling and balancing weights with their unbiasedness for $\rho(\mathbf{X}_i)$ and $\pi_0(\mathbf{X}_i)$ allows us to apply M-estimation techniques (Stefanski and Boos, 2002) to show that the double-robustness property holds. Additional details for constructing inferences regarding τ_0 with $\hat{\tau}_{\text{CAL}}$ is contained within the Appendix.

Recall that for the data-fusion setting, we are provided $(\mathbf{X}_i, Y_i, Z_i)$ $(i = 1, \ldots, n)$ for both samples A and B. Given the setup we describe in Section 4, we may estimate $\hat{\boldsymbol{\lambda}}_0 \in \Re^m$ and $\hat{\boldsymbol{\lambda}}_1 \in \Re^m$ over all $i = 1, 2, \ldots, n$

in (6) instead of for only $i \in \{i : S_i = 1\}$. This means that (7) exists for all $i = 1, 2, \ldots, n$ and we can estimate (8) over the combined samples A and B by simply changing the index of the summations to account for all $i = 1, 2, \ldots, n$.

5 Simulation Study

5.1 Simulation Setup

To demonstrate the efficacy of the different methods for transportability and data-fusion, we will conduct a simulation study that evaluates the performance of the three doubly-robust methods we have identified in Sections 3 and 4. As was done in multiple other published articles (Lunceford and Davidian, 2004; Kang and Schafer, 2007), we will test the performance of our proposed methodologies across a range of scenarios which focus on model misspecification. When evaluating methods for transporting observational study results to different populations, we must consider possibilities where the sampling, treatment, and/or the outcome models are misspecified. At the same time, we will adhere to Assumptions 1-4. We will compare the full calibration estimator proposed in Section 4 with the TMLE and augmented estimators described in Section 3. We also considered estimators for τ_0 under the data-fusion setting, where both samples A and B have observed treatments and responses - these extensions are for τ_{AUG} and τ_{CAL} which account for the additional data.

For the sake of presentation, we have described the full calibration estimator in sequential terms by estimating the sample weights preceded by estimation of the balance weights. We found that fitting the sampling and balancing weights in parallel yielded better results. To estimate the combined weights in parallel requires constructing and solving a convex optimization problem which minimizes the relative entropy criterion distance, as it appears in (3), subject to the constraints found in (3) and (9), simultaneously (Censor and Zenios, 1998). To solve this optimization problem simply requires substituting the combined constraints and target margins into the Lagrangian that is used to construct the dual problem found in (2).

The scenarios we examine vary the sample size $n \in \{1000, 2000\}$, the generative process that determines the treatment assignment, the outcome process, and the sampling process. For every $i = 1, 2, \ldots, n$, the covariates $\mathbf{X}_i \equiv (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$ are distributed as $X_{i1}, X_{i2}, X_{i3}, X_{i4} \sim \mathcal{N}(0, 1)$. We also construct the transformed variables $U_{i1} = \exp(X_{i1} + X_{i4})$, $U_{i2} = (X_{i1} + X_{i2})^3$, $U_{i3} = (X_{i2} + X_{i3})^2$ and $U_{i4} = \log(|X_{i3} X_{i4}|)$. Each entry in the vector $\mathbf{U}_i \equiv (U_{i1}, U_{i2}, U_{i3}, U_{i4})^T$ is standardized to have a mean of zero and marginal variances of one - same as for $\mathbf{X}_i$.

The sample indicators are generated assuming $S_i \sim B(\rho_i^{(r)})$, $r \in \{a, b\}$ where

$$\text{logit}\left[\rho_i^{(a)}\right] = -0.5X_{i1} - X_{i2} - 0.5X_{i3} + X_{i4} \text{ and}$$
$$\text{logit}\left[\rho_i^{(b)}\right] = -0.5U_{i1} - U_{i2} - 0.5U_{i3} + U_{i4}.$$

The treatment assignments are generated by sampling from $Z_i \sim B\left(\pi_i^{(k)}\right)$, $k \in \{a, b\}$, where

$$\begin{aligned} \pi_i^{(a)} &= S_i \text{expit}\left(\mathbf{X}_i^T \boldsymbol{\delta}\right) + (1 - S_i)\text{expit}\left(\mathbf{X}_i^T \boldsymbol{\lambda}\right) \text{ and} \\ \pi_i^{(b)} &= S_i \text{expit}\left(\mathbf{U}_i^T \boldsymbol{\delta}\right) + (1 - S_i)\text{expit}\left(\mathbf{U}_i^T \boldsymbol{\lambda}\right). \end{aligned} \tag{10}$$

The coefficients in (10) are $\boldsymbol{\delta} \equiv (-1, -0.5, 0, 0.5)^T$ and $\boldsymbol{\lambda} \equiv (0, 1, -0.5, 0.5)^T$. We generate potential outcomes from the bivariate model

$$\begin{bmatrix} Y_i(0) \\ Y_i(1) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_i^{(\ell)} \\ \kappa_i^{(\ell)} \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \right),$$

where $\ell \in \{a, b\}$ indexes

$$\begin{aligned} \mu_i^{(a)} &= 5 - X_{i1} + 3X_{i2} - 3X_{i3} + X_{i4}, \\ \kappa_i^{(a)} &= 5 - 3X_{i1} - X_{i2} + X_{i3} + 3X_{i4}, \\ \mu_i^{(b)} &= 5 - U_{i1} + 3U_{i2} - 3U_{i3} + U_{i4}, \text{ and} \\ \kappa_i^{(b)} &= 5 - 3U_{i1} - U_{i2} + U_{i3} + 3U_{i4}, \end{aligned} \tag{11}$$

The observed outcome corresponds with the potential outcome of the observed treatment assignment. The counterfactual argument is discarded. In the transportability examples, we also discard both Y_i and Z_i for all $\{i : S_i = 0\}$. The variance of the potential outcomes is set to $\sigma^2 = 4$. Each of the methods we consider are provided the design matrix with an intercept term and the four original covariate values - X_{i1}, X_{i2}, X_{i3}, and X_{i4} for all $i = 1, 2, \ldots, n$. The causal effects are estimated using the respective estimators described in Sections 3 and 4 - $\hat{\tau}_{\text{TMLE}}$, $\hat{\tau}_{\text{AUG}}$ or $\hat{\tau}_{\text{CAL}}$. For TMLE and the augmented estimator, the component models of $\hat{\pi}_1(\mathbf{X}_i)$ and $\hat{\rho}(\mathbf{X}_i)$ are fit with logistic regression while $\hat{\mu}_0(\mathbf{X}_i)$ and $\hat{\mu}_1(\mathbf{X}_i)$ are fit with least squares regression. Between the outcome scenarios, the treatment assignment scenarios, the sampling scenarios, and the sample size, there are a total of 16 experimental scenarios.

5.2 Simulation Results

We report empirical averages and Monte Carlo standard errors for each of the scenarios described in Section 5.1 using the estimators described in Sections 3 and 4, plus the two extensions for data-fusion that were discussed briefly. The results of the experiment are summarized in Table 1. For a graphical interpretation, Figure 1 contains a further subset of the results that make up the contents of Table 1 focusing on the set of scenarios where at least one of the models is misspecified and at least one model is correctly specified. We also report the coverage probabilities for the full calibration estimators in both the transportability and data-fusion settings in Table 2. The 95% confidence interval is estimated using a robust variance estimator which we derive in the Appendix.

Observe that when there is the opportunity to combine datasets and solve the data-fusion problem, we get more precise estimates of the target population average treatment effect among the unbiased scenarios. Overall, the results of the full calibration and augmented approaches were less prone to error than the

TMLE results in each of the scenarios tested. The difference in efficiency between the augmented and full calibration approaches appears to be negligible, although the augmented approach had a noticeably smaller Monte Carlo standard error in the transportability setting across every scenario. TMLE, on the other hand, had the largest Monte Carlo standard errors. We found the results of the augmented approach interesting as we had hypothesized from the outset that adding additional parametric assumptions should have resulted in increased efficiency. One advantage to the parametric models we apply is the easy derivation of a variance estimator (Appendix A). The coverage probabilities of the confidence interval estimates for the full calibration approach are shown in Table 2. Note that given the setup to the simulation, the transportability cases have an effective sample size of around $n_1 \approx 500$ for $n = 1000$ and $n_1 \approx 1000$ when $n = 2000$.

Observe that in the scenarios where the outcome model is correctly specified, the estimated population average treatment effects are consistent. For the scenarios where both the outcome model and either the treatment or sampling model is misspecified, we get biased results. When the outcome is misspecified and the treatment and sampling models are correctly-specified, the full calibration approach produces the least amount of bias, both in the transportability and data-fusion cases. This is due to the conditions of the simulation experiment which violate the propensity score exchangeability assumption. This specific scenario clearly illustrates the shortcomings of the augmented and TMLE approaches when this assumption does not hold. With so much uncertainty arising from model misspecification, we advocate for using doubly-robust methods to increase the chances of better ensuring consistent estimation of the causal effects. Doubly-robust methods present two opportunities for the researcher to correctly model the causal effect whereas with G-computation and with inverse probability of treatment/sampling approaches, either the outcome model or the sampling and treatment models must be correctly specified, respectively.

6 Illustrating Examples

6.1 Evaluating Metformin Versus Sulfonylureas as Monotherapy for VA Diabetes Patients

As we discussed in Section 1.1, the aim of the applied problem is to estimate the risk difference in mortality between sulfonylurea and metformin monotherapy for the 2010-2014 cohort of diabetic VA patients. We use improved covariate balance propensity score (iCBPS) weights (Fan et al., 2016; Josey et al., 2020b) to estimate the risk difference using only the 2010-2014 cohort, which balance the covariates found in Table 3. As we mentioned earlier, this estimate will serve as a benchmark for the transport and data-fusion settings. We also estimate the risk difference for the 2004-2009 cohort without the 2010-2014 data, again for comparative purposes. We use (6)-(8) to find estimates of total mortality in the 2004-2009 sample transported to the 2010-2014 cohort. We also examine the setting which combines the 2004-2009 sample with the 2010-2014 sample to estimate the risk difference among patients in the 2010-2014 cohort using the extensions to (6)-(8) to accommodate the data-fusion setting. Since sulfonylurea use in 2004-2009 represents

27.0% of monotherapy recipients but only 11.8% of patients in 2010-2014 implies we will need methods which account for violations of propensity score exchangeability.

Both cohorts excluded patients with pre-existing forms of cancer. We also omitted patients that received a second-line medication (either insulin, a thiazolidinedione, a sulfonylurea for patients receiving metformin, or metformin for patients receiving a sulfonylurea) within 30 days of their first filled prescription of either metformin or a sulfonylurea. Time to mortality, which is used to create the indicators of the three mortality outcomes, is computed as the number of days to death from the date when the first prescription is filled. Our analysis assumes intention to treat with the first prescribed therapy. We do not censor patients at the time when a second-line medication is prescribed, as was done in Wheeler et al. (2013). The remaining baseline, demographic, laboratory measurements, and comorbidities about the two cohorts are summarized in Table 3.

Table 4 contains the various risk-difference estimates that we computed on the illustrative dataset. We will primarily focus on the estimates of five-year mortality since these figures have the greatest magnitude. Furthermore, the trends that we will report about five-year mortality appear to be the same for one- and two-year mortality. Observe that the crude risk difference in five-year mortality is similar within both the 2004-2009 and in the 2010-2014 cohorts. This implies that there is either limited or no changes to the risk difference attributable to temporal trends in treatment efficacy - i.e. the effectiveness of either therapy has not changed relative to the other. A temporal effect modifier is one factor that we would not be able to accommodate without violating Assumption 3. The adjusted marginal estimates of the risk difference using iCBPS reveal the importance of accounting for differences observed between the study and target cohorts. The risk difference in the 2004-2009 cohort is 12.2% (11.6%, 12.7%) and 4.1% (3.4%, 4.9%) in the 2010-2014 cohort. Given the limited change of the crude estimates between cohorts, these discrepancies are likely due to differences in the distribution of effect modifiers between the two temporally-distinct cohorts. When we transport the estimates of the 2004-2009 cohort onto the 2010-2014 cohort, the risk difference is found to be 4.0% (3.4%, 4.6%). That is, the transported effect estimate is more similar to the iCBPS estimate of 2010-2014 than of the 2004-2009 cohort. Carrying over from the 2004-2009 calibrated estimate, the transported estimate from the 2004-2009 cohort onto the 2010-2014 cohort is more efficient than the calibrated estimate computed with only the data from the 2010-2014 cohort. When we combine the two cohorts, we get an unbiased estimate of the target population average treatment effect with much greater efficiency than using only data from 2010-2014 alone. Using this estimator, we found the estimated risk difference for 2010-2014 to be 4.2% (3.7%, 4.7%).

Regardless of the outcome or estimator, our results suggest that sulfonylurea monotherapy remains more harmful overall than metformin monotherapy for newly diagnosed veterans with diabetes. This is true both in 2004-2009 and 2010-2014, indicating that a continued phaseout of sulfonylureas may be advisable except in specific targeted subgroups where it might be more effective. Continued research will be needed to identify these subgroups, should any exist.

6.2 Risk of Total Mortality from Insulin Provision Versus Sensitization in a Veteran Population with Diabetes and CAD

In addition to transporting the risk difference of total mortality among patients receiving sulfonylurea versus metformin across temporally defined populations within the VA, we also transport estimates from the BARI 2D trial population onto the VA 2010-2014 cohort. This example showcases the fact that our method works for transporting estimates from trial data onto observational data in addition to between observational samples as we showed in Section 6. In this example we compare insulin sensitization therapy, which consists of treatment with metformin and/or a thiazolidinedione (another class of oral diabetes medications), with insulin provision therapy, which consists of treatment with a sulfonylurea and/or insulin, on the risk of total mortality three years after randomization.

The BARI 2D study enrolled 2,368 diabetes patients with untreated coronary artery disease (CAD) into four treatment groups along a 2x2 factorial design. In addition to examining the effects of glycemic control strategies, the BARI 2D study also tested the effects of delayed versus contemporaneous treatment of CAD. We will ignore this portion of the study and focus solely on glycemic control. We construct a representative cohort from the VA electronic health record as the target sample. This sample included all diabetes patients diagnosed within the calendar years of 2010-2014 with prior history of CAD that received either insulin provision or insulin sensitization therapy after diagnosis with diabetes mellitus (n = 30,393). We note that there is some misalignment with the two samples in this example since patients within the BARI 2D study had a longer duration of diabetes prior to randomization, during which time most patients had some diabetes treatment prior to randomization. If we were to try and balance this variable between samples, we would likely violate Assumption 3. Nevertheless we should still be able to estimate informative risk differences after accounting for the other risk factors attributable to total mortality considered for diabetes patients and after accounting for the covariates with differences between samples in our model. The factors which we balance between the two samples and the treatment groups are found in Figure 2. Here we display the standardized mean differences of the covariates between the treated and controls as well as between the BARI 2D sample and the sample of 2010-2014 VA diabetes patients with CAD, both before and after adjustment using weights for transportability and data-fusion. These covariates are measured in both the BARI 2D trial sample and within the VA cohort.

Similar to the analysis in Section 6, we find the unadjusted and the iCBPS adjusted risk difference estimates in both the VA and BARI 2D cohorts. We then transport the BARI 2D results to the VA cohort using the calibration methods discussed in Section 6. We supplement this estimate by finding the risk difference under the data-fusion setting which combines the responses, treatment, and covariates of the BARI 2D study with the VA cohort. Both the crude and adjusted results using data only from the BARI 2D study without integrating VA data corroborate what was originally found in the trial analysis - insulin provision has no effect on total mortality compared to insulin sensitization. After three years and adjusting using the iCBPS weights, the BARI 2D study saw no change (-2.1%, 2.2%) in total mortality between the

two treatment groups. However, after weighting the BARI 2D responses to transport the estimated risk difference onto the VA cohort, we observed a 2.4% (-4.1%, 8.8%) increased risk of death among patients receiving insulin provision therapy. For the data-fusion result, we estimate an increase risk in total mortality of 4.2% (2.9%, 5.5%) three years after randomization. This better aligns with the risk difference estimated using only the VA cohort which found an increased risk difference of 4.2% (3.0%, 5.4%).

7 Discussion

By sequentially estimating the vector of sampling weights followed by the vector of balancing weights, we show how constrained convex optimization techniques can be applied in joining the covariate balancing problem described in Chan et al. (2015) with solutions for transporting experimental effect estimates (Signorovitch et al., 2010; Westreich et al., 2017). The resulting estimator eliminates both within-treatment group heterogeneity as well as any heterogeneity that might occur between study participants and non-participants attributed to sampling. This allows us to transport estimates found with observational data across populations. We also examined two alternative approaches for transportability which we adapted to accommodate data for study and target populations derived from observational studies. Along with the full calibration method, the augmented estimator can be extended to solve the problem of data-fusion in an observational context. The TMLE (Rudolph and van der Laan, 2017) and augmented estimators (Dong et al., 2020) are less constrained by parametric assumptions, but they do not account for the possibility of the propensity score differing between the target and study samples. One way to avoid this problem might be to estimate $\pi_0(\mathbf{X}_i)$ by conditioning the estimator for $\pi_1(\mathbf{X}_i)$ on the estimated inverse odds of sampling weights - (4) for example. This approach mirrors the process taken by (6), which finds $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_0$ conditioned on the sampling weight estimates (4).

In the simulation study conducted in Section 5, we found that using some form of calibration, either with the augmented approach or the full calibration approach, yielded the most efficient estimates. In cases where the outcome model is correctly specified, the augmented estimator performed the best of the three methods we tested. However, we also demonstrate that the augmented estimator requires Assumption 7 to remain consistent when the outcome model is misspecified but the treatment assignment and sampling models are correctly specified. As we have mentioned frequently, the propensity score exchangeability assumption is critical within the data-fusion setting. In addition to the simulation study, we show in an illustrative example of US veterans with diabetes how different populations produce different effect estimates for the same outcome. We then demonstrate how eliminating the sampling bias induced by differences in the distribution of the effect modifiers between cohorts produces consistent estimates of the treatment effect on the target population.

One of the major shortcomings of the full calibration method is the set of linearity conditions nested within Assumptions 5-7. These assumptions are necessary to guarantee the double-robustness property

as shown in the Appendix. The TMLE approach does not require any assumption about the functional form of $\rho(\mathbf{X}_i)$ or $\pi_0(\mathbf{X}_i)$ while the augmented approach only requires Assumption 6 to hold to guarantee double-robustness. If the outcome model is not misspecified, then neither the augmented estimator nor TMLE require any assumptions about the form of $\mu_0(\mathbf{X}_i)$ and $\mu_1(\mathbf{X}_i)$. We note that the more stringent linearity assumptions culminate in a tradeoff between more flexible modeling strategies and the requirement of Assumption 8. One solution to relax the linearity conditions in the full calibration approach might be to use sieve regression methods (Geman and Hwang, 1982) that replace the balance functions in (2), (4), (6), and (7) with polynomial expansions and interactions of the covariates. This nonparametric approach was explored in Chan et al. (2015) for estimating balancing weights to mitigate treatment group heterogeneity and briefly for generalizability in Dong et al. (2020).

We stated that the complete individual-level covariate data were required for both samples A and B. It would be advantageous to only require the marginal moment values of the covariate distribution in sample B for transporting observational study results as these entries are often found in the scientific and medical literature in a so-called Table 1. This setting is discussed in more detail by Josey et al. (2020a) under the setting of transporting randomized clinical trial results. In that article we point out that any resulting inference in such a setting would involve the target *sample* average treatment effect, $\tau_0' \equiv n_0^{-1} \sum_{\{i:S_i=0\}} Y_i(1) - Y_i(0)$, instead of the target population average treatment effect τ_0. There is nothing to indicate the same argument would not be true for transporting observational study results.

References

Ali, M. K., Bullard, K. M., Saaddine, J. B., Cowie, C. C., Imperatore, G., and Gregg, E. W. (2013). Achievement of goals in U.S. diabetes care, 1999-2010. *The New England Journal of Medicine*, 368(17):1613–1624.

Azoulay, L. and Suissa, S. (2017). Sulfonylureas and the risks of cardiovascular events and death: a methodological meta-regression analysis of the observational studies. *Diabetes Care*, 40(5):706–714.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Berkowitz, S. A., Krumme, A. A., Avorn, J., Brennan, T., Matlin, O. S., Spettell, C. M., Pezalla, E. J., Brill, G., Shrank, W. H., and Choudhry, N. K. (2014). Initial choice of oral glucose-lowering medication for diabetes mellitus: a patient-centered comparative effectiveness study. *JAMA internal medicine*, 174(12):1955–1962.

Censor, Y. and Zenios, S. A. (1998). *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, New York, first edition.

Chan, K. C. G., Yam, S. C. P., and Zheng, Z. (2015). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700.

Cheng, Y. J., Imperatore, G., Geiss, L. S., Saydah, S. H., Albright, A. L., Ali, M. K., and Gregg, E. W. (2018). Trends and disparities in cardiovascular mortality among U.S. adults with and without self-reported diabetes, 1988-2015. *Diabetes Care*, 41(11):2306–2315.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American Journal of Epidemiology*, 172(1):107–115.

Desai, N. R., Shrank, W. H., Fischer, M. A., Avorn, J., Liberman, J. N., Schneeweiss, S., Pakes, J., Brennan, T. A., and Choudhry, N. K. (2012). Patterns of medication initiation in newly diagnosed diabetes mellitus: quality and cost implications. *The American Journal of Medicine*, 125(3):302.e1–7.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Dong, L., Yang, S., Wang, X., Zeng, D., and Cai, J. (2020). Integrative analysis of randomized clinical trials with real world evidence studies. *arXiv:2003.01242 [stat]*.

Fan, J., Imai, K., Liu, H., Ning, Y., and Yang, X. (2016). Improving covariate balancing propensity score: a doubly robust and efficient approach. *Technical Report*.

Geiss, L. S., Wang, J., Cheng, Y. J., Thompson, T. J., Barker, L., Li, Y., Albright, A. L., and Gregg, E. W. (2014). Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012. *JAMA*, 312(12):1218–1226.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414.

Gregg, E. W., Cheng, Y. J., Srinivasan, M., Lin, J., Geiss, L. S., Albright, A. L., and Imperatore, G. (2018). Trends in cause-specific mortality among adults with and without diagnosed diabetes in the USA: an epidemiological analysis of linked national survey and vital statistics data. *Lancet (London, England)*, 391(10138):2430–2440.

Gregg, E. W., Li, Y., Wang, J., Burrows, N. R., Ali, M. K., Rolka, D., Williams, D. E., and Geiss, L. (2014). Changes in diabetes-related complications in the United States, 1990-2010. *The New England Journal of Medicine*, 370(16):1514–1523.

Hampp, C., Borders-Hemphill, V., Moeny, D. G., and Wysowski, D. K. (2014). Use of antidiabetic drugs in the U.S., 2003-2012. *Diabetes Care*, 37(5):1367–1374.

Holman, R. R., Bethel, M. A., Mentz, R. J., Thompson, V. P., Lokhnygina, Y., Buse, J. B., Chan, J. C., Choi, J., Gustavson, S. M., Iqbal, N., Maggioni, A. P., Marso, S. P., Öhman, P., Pagidipati, N. J., Poulter, N., Ramachandran, A., Zinman, B., and Hernandez, A. F. (2017). Effects of once-weekly exenatide on cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 377(13):1228–1239.

Hong, J., Zhang, Y., Lai, S., Lv, A., Su, Q., Dong, Y., Zhou, Z., Tang, W., Zhao, J., Cui, L., Zou, D., Wang, D., Li, H., Liu, C., Wu, G., Shen, J., Zhu, D., Wang, W., Shen, W., Ning, G., and SPREAD-DIMCAD Investigators (2013). Effects of metformin versus glipizide on cardiovascular outcomes in patients with type 2 diabetes and coronary artery disease. *Diabetes Care*, 36(5):1304–1311.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Johnson, J. A., Majumdar, S. R., Simpson, S. H., and Toth, E. L. (2002). Decreased mortality associated with the use of metformin compared with sulfonylurea monotherapy in type 2 diabetes. *Diabetes Care*, 25(12):2244–2248.

Josey, K. P., Berkowitz, S. A., Ghosh, D., and Raghavan, S. (2020a). Transporting experimental results with entropy balancing. *arXiv:2002.07899 [stat]*.

Josey, K. P., Juarez-Colunga, E., Yang, F., and Ghosh, D. (2020b). A Framework for Covariate Balance using Bregman Distances. *Scandinavian Journal of Statistics*, pages 1–27.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., and Cole, S. R. (2017). Generalizing study results: a potential outcomes perspective. *Epidemiology (Cambridge, Mass.)*, 28(4):553–561.

Lu, Y., Scharfstein, D. O., Brooks, M. M., Quach, K., and Kennedy, E. H. (2019). Causal inference for comprehensive cohort studies. *arXiv:1910.03531 [stat]*.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Marso, S. P., Bain, S. C., Consoli, A., Eliaschewitz, F. G., Jódar, E., Leiter, L. A., Lingvay, I., Rosenstock, J., Seufert, J., Warren, M. L., Woo, V., Hansen, O., Holst, A. G., Pettersson, J., and Vilsbøll, T. (2016a). Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine*, 375(19):1834–1844.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Mann, J. F., Nauck, M. A., Nissen, S. E., Pocock, S., Poulter, N. R., Ravn, L. S., Steinberg, W. M., Stockner, M., Zinman, B., Bergenstal, R. M.,

and Buse, J. B. (2016b). Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 375(4):311–322.

Neal, B., Perkovic, V., Mahaffey, K. W., de Zeeuw, D., Fulcher, G., Erondu, N., Shaw, W., Law, G., Desai, M., and Matthews, D. R. (2017). Canagliflozin and cardiovascular and renal events in type 2 diabetes. *New England Journal of Medicine*, 377(7):644–657.

Pearl, J. and Bareinboim, E. (2014). External validity: from do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595.

Raghavan, S., Vassy, J. L., Ho, Y.-L., Song, R. J., Gagnon, D. R., Cho, K., Wilson, P. W. F., and Phillips, L. S. (2019). Diabetes mellitus-related all-cause and cardiovascular mortality in a national cohort of adults. *Journal of the American Heart Association*, 8(4):e011295.

Rao Kondapally Seshasai, S., Kaptoge, S., Thompson, A., Di Angelantonio, E., Gao, P., Sarwar, N., Whincup, P. H., Mukamal, K. J., Gillum, R. F., Holme, I., Njølstad, I., Fletcher, A., Nilsson, P., Lewington, S., Collins, R., Gudnason, V., Thompson, S. G., Sattar, N., Selvin, E., Hu, F. B., Danesh, J., and Emerging Risk Factors Collaboration (2011). Diabetes mellitus, fasting glucose, and risk of cause-specific death. *The New England Journal of Medicine*, 364(9):829–841.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Roumie, C. L., Hung, A. M., Greevy, R. A., Grijalva, C. G., Liu, X., Murff, H. J., Elasy, T. A., and Griffin, M. R. (2012). Comparative effectiveness of sulfonylurea and metformin monotherapy on cardiovascular events in type 2 diabetes mellitus: a cohort study. *Annals of Internal Medicine*, 157(9):601–610.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

Rudolph, K. E. and van der Laan, M. J. (2017). Robust estimation of encouragement design intervention effects transported across sites. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1509–1525.

Schramm, T. K., Gislason, G. H., Vaag, A., Rasmussen, J. N., Folke, F., Hansen, M. L., Fosbøl, E. L., Køber, L., Norgaard, M. L., Madsen, M., Hansen, P. R., and Torp-Pedersen, C. (2011). Mortality and cardiovascular risk associated with different insulin secretagogues compared with metformin in type 2

diabetes, with or without a previous myocardial infarction: a nationwide study. *European Heart Journal*, 32(15):1900–1908.

Schuler, M. S. and Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73.

Selvin, E., Parrinello, C. M., Sacks, D. B., and Coresh, J. (2014). Trends in prevalence and control of diabetes in the United States, 1988-1994 and 1999-2010. *Annals of Internal Medicine*, 160(8):517–525.

Signorovitch, J. E., Wu, E. Q., Yu, A. P., Gerrits, C. M., Kantor, E., Bao, Y., Gupta, S. R., and Mulani, P. M. (2010). Comparative effectiveness without head-to-head trials. *PharmacoEconomics*, 28(10):935–945.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

The BARI 2D Study Group (2009). A Randomized Trial of Therapies for Type 2 Diabetes and Coronary Artery Disease. *New England Journal of Medicine*, 360(24):2503–2515.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer-Verlag, New York.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Varvaki Rados, D., Catani Pinto, L., Reck Remonti, L., Bauermann Leitão, C., and Gross, J. L. (2016). The association between sulfonylurea use and all-cause and cardiovascular mortality: a meta-analysis with trial sequential analysis of randomized clinical trials. *PLoS medicine*, 13(4):e1001992.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., and Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, 186(8):1010–1014.

Wheeler, S., Moore, K., Forsberg, C. W., Riley, K., Floyd, J. S., Smith, N. L., and Boyko, E. J. (2013). Mortality among veterans with type 2 diabetes initiating metformin, sulfonylurea or rosiglitazone monotherapy. *Diabetologia*, 56(9):1934–1943.

Zinman, B., Wanner, C., Lachin, J. M., Fitchett, D., Bluhmki, E., Hantel, S., Mattheus, M., Devins, T., Johansen, O. E., Woerle, H. J., Broedl, U. C., and Inzucchi, S. E. (2015). Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. *New England Journal of Medicine*, 373(22):2117–2128.

A Inference using Calibration Weights for Transportability

Consider the transportability case. We begin by defining the estimating equations for the parameters introduced in Sections 3 and 4. Let $\mathbf{c}(\mathbf{X}_i) \equiv [c_1(\mathbf{X}_i), c_2(\mathbf{X}_i), \ldots, c_m(\mathbf{X}_i)]^T$. First, we define $\boldsymbol{\omega}(S_i, \mathbf{X}_i; \boldsymbol{\theta}) \equiv$

$(1 - S_i)\left[\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}\right]$, which is solved by $\sum_{i=1}^{n} \boldsymbol{\omega}(S_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}})] = \mathbf{0}_m$ where $\hat{\boldsymbol{\theta}} = \frac{1}{n_0}\sum_{i=1^n}(1 - S_i)\mathbf{c}(\mathbf{X}_i)$. Next, define

$$\boldsymbol{\zeta}(S_i, \mathbf{X}_i; \boldsymbol{\gamma}, \boldsymbol{\theta}) \equiv S_i \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j\right]\left[\mathbf{c}(\mathbf{X}) - \boldsymbol{\theta}\right],$$

which is used to solve for $\sum_{i=1}^{n} \zeta(S_i, \mathbf{X}_i; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}) = \mathbf{0}_m$ as a result of the Lagrangian multiplier theorem discussed in Section 3. The score equations for $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ are identified as

$$\boldsymbol{\xi}_0(S_i, \mathbf{X}_i, Z_i; \boldsymbol{\gamma}, \boldsymbol{\lambda}_0, \boldsymbol{\theta}) = S_i(1 - Z_i)\exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)(\gamma_j + \lambda_{j0})\right]\left[\mathbf{c}(\mathbf{X}) - \boldsymbol{\theta}\right]$$

$$\boldsymbol{\xi}_1(S_i, \mathbf{X}_i, Z_i; \boldsymbol{\gamma}, \boldsymbol{\lambda}_1, \boldsymbol{\theta}) = S_i Z_i \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)(\gamma_j + \lambda_{j1})\right]\left[\mathbf{c}(\mathbf{X}) - \boldsymbol{\theta}\right]$$

which can be used to solve for $\sum_{i=1}^{n} \boldsymbol{\xi}_0(S_i, \mathbf{X}_i, Z_i; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\theta}}) = \mathbf{0}_m$ and $\sum_{i=1}^{n} \boldsymbol{\xi}_1(S_i, \mathbf{X}_i, Z_i; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\theta}}) = \mathbf{0}_m$. Finally, we can write the score equation for τ_0 as

$$\phi(S_i, \mathbf{X}_i, Y_i, Z_i; \boldsymbol{\gamma}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \tau) = S_i Z_i p(\mathbf{X}_i)\left[Y_i(1) - \tau\right] - S_i(1 - Z_i)p(\mathbf{X}_i)Y_i(0) \tag{12}$$

which we may solve as $\sum_{i=1}^{n} \phi(S_i, \mathbf{X}_i, Y_i Z_i; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1, \hat{\tau}_{\text{CAL}}) = 0$. Given this setup, we can see that these equations are conducive of M-estimation theory. For simplicity, we will often drop the parameter values in the notation when using the functional representations of the estimating equations.

To show double-robustness, we first prove that $\hat{\tau}_{\text{CAL}}$ is consistent for τ_0 given Assumption 5. This means we can assume

$$\mu_0^*(\mathbf{X}_i) = \sum_{j=1}^{m} c_j(\mathbf{X}_i)\alpha_j^* \text{ and}$$
$$\mu_1^*(\mathbf{X}_i) = \sum_{j=1}^{m} c_j(\mathbf{X}_i)\beta_j^*.$$

Let $\hat{p}(\mathbf{X}_i)$ be determined by (7) where $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_0$ are solved using the objectives in (6). If we assume $c_1(\mathbf{X}_i) = 1$ for all $i = \{i : S_i = 1\}$, then $n_1 = \sum_{\{i:S_i=1\}} \hat{p}(\mathbf{X}_i)Z_i$. If we substitute $\hat{p}(\mathbf{X}_i)$ for $p(\mathbf{X}_i)$ into (12), regardless of whether it is a correctly specified model for the balancing and sampling weights, to find the expectation

$$\begin{aligned}
\frac{1}{n_1}\mathbb{E}\left[\sum_{i=1}^{n} \phi(S_i, \mathbf{X}_i, Y_i, Z_i)\right] &= \frac{1}{n_1}\mathbb{E}\left\{\sum_{i=1}^{n} \mathbb{E}\left[S_i Z_i \hat{p}(\mathbf{X}_i)Y_i(1) - S_i(1 - Z_i)\hat{p}(\mathbf{X}_i)Y_i(0)|S_i, \mathbf{X}_i, Z_i\right]\right\} - \tau_0 \\
&= \frac{1}{n_1}\mathbb{E}\left[\sum_{i=1}^{n} S_i Z_i \hat{p}(\mathbf{X}_i)\sum_{j=1}^{m} c_j(\mathbf{X}_i)\beta_j^* - \sum_{i=1}^{n} S_i(1 - Z_i)\hat{p}(\mathbf{X}_i)\sum_{j=1}^{m} c_j(\mathbf{X}_i)\alpha_j^*\right] - \tau_0 \\
&= \mathbb{E}\left[\sum_{j=1}^{m} \hat{\theta}_j\beta_j^* - \sum_{j=1}^{m} \hat{\theta}_j\alpha_j^*\right] - \tau_0 \\
&= 0
\end{aligned}$$

Now suppose Assumptions 6 and 7 are given. This means

$$\begin{aligned}
\log[\pi^*(S_i, \mathbf{X}_i)] &= S_i \sum_{j=1}^{m} c_j(\mathbf{X}_i)\delta_{j1}^* + (1 - S_i) \sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j1}^*, \\
\log[1 - \pi^*(S_i, \mathbf{X}_i)] &= S_i \sum_{j=1}^{m} c_j(\mathbf{X}_i)\delta_{j0}^* + (1 - S_i) \sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j0}^*, \text{ and} \\
\text{logit}[\rho^*(\mathbf{X}_i)] &= \sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j^*.
\end{aligned}$$

It is trivial to see that $\mathbb{E}[\boldsymbol{\omega}(S_i, \mathbf{X}_i; \boldsymbol{\theta}^*)] = \mathbf{0}_m$, where $\boldsymbol{\theta}^* = \mathbb{E}[\mathbf{c}(\mathbf{X}_i)|S_i = 0]$. The expectation of the estimating equation for $\boldsymbol{\gamma}$ can be expanded to reveal

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\zeta}(S_i, \mathbf{X}_i; \theta_j^*, \boldsymbol{\gamma}^*, \theta_j^*)\right] &= \mathbb{E}\left(\mathbb{E}\left\{ S_i \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j\right] [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}^*] \middle| \mathbf{X}_i \right\}\right) \\
&= \mathbb{E}\left\{ \frac{\exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j\right]}{1 + \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j^*\right]} [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}^*] \right\},
\end{aligned}$$

which can only evaluate to zero if

$$\frac{\exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j\right]}{1 + \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\gamma_j^*\right]} \propto \Pr\{S_i = 0|\mathbf{X}_i\}.$$

Therefore $\mathbb{E}\left[\boldsymbol{\zeta}(S_i, \mathbf{X}_i; \boldsymbol{\gamma}, \boldsymbol{\theta}^*)\right] \propto \mathbb{E}\left[\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}^* | S_i = 0\right] = \mathbf{0}_m$. This result helps simplify solving the expectations of the estimating equations for $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$, leading us to find

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\xi}_1(S_i, \mathbf{X}_i, Z_i; \theta_j^*, \boldsymbol{\gamma}^*, \boldsymbol{\lambda}_1)\right] &= \mathbb{E}\left[\mathbb{E}\left(S_i Z_i \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)(\gamma_j + \lambda_{j1})\right] [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}^*] \middle| \mathbf{X}_i, Z_i \right)\right] \\
&\propto \mathbb{E}\left[\mathbb{E}\left\{ Z_i \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j1}\right] [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}^*] \middle| S_i = 0, \mathbf{X}_i \right)\right\} \qquad (13) \\
&= \mathbb{E}\left(\frac{\exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j1}\right]}{\exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j1}^*\right]} [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}^*] \middle| S_i = 0 \right).
\end{aligned}$$

As with taking the expectation of $\boldsymbol{\zeta}$, the only way for (13) to evaluate to zero is if $\lambda_{j1} = \lambda_{j1}^*$ for all $j = 1, 2, \ldots, m$. A similar result can be shown for $\boldsymbol{\xi}_0$ where $\lambda_{j0} = \lambda_{j0}^*$ in order for

$$\mathbb{E}\left[\boldsymbol{\xi}_0(S_i, \mathbf{X}_i, Z_i; \theta_j^*, \boldsymbol{\gamma}^*, \boldsymbol{\lambda}_0^*)\right] = 0.$$

Combining these results, we can then conclude

$$\mathbb{E}\left[\phi(S_i, \mathbf{X}_i, Y_i, Z_i; \boldsymbol{\gamma}^*, \boldsymbol{\lambda}_0^*, \boldsymbol{\lambda}_1^*, \tau_0)\right] = 0.$$

If we concatenate the parameter values into $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\lambda}_0^T, \boldsymbol{\lambda}_1^T, \tau)$ and stacking the estimating equations with

$$\boldsymbol{\psi}(S_i, \mathbf{X}_i, Y_i, Z_i; \boldsymbol{\eta}) \equiv [\boldsymbol{\omega}(S_i, \mathbf{X}_i)^T, \boldsymbol{\zeta}(S_i, \mathbf{X}_i)^T, \boldsymbol{\xi}_0(S_i, \mathbf{X}_i, Z_i)^T, \boldsymbol{\xi}_1(S_i, \mathbf{X}_i, Z_i)^T, \phi(S_i, \mathbf{X}_i, Y_i, Z_i)]^T,$$

then given the properties of m-estimators under mild regularity conditions (Tsiatis, 2006), we can generate the influence function that finds

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^* = -\mathbb{E}\left[\frac{\partial \psi(S_i, \mathbf{X}_i, Y_i, Z_i; \boldsymbol{\eta}^*)}{\partial \boldsymbol{\eta}}\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \psi(S_i, \mathbf{X}_i, Y_i, Z_i; \boldsymbol{\eta}^*)\right] + o_p(n^{-1/2}). \tag{14}$$

Equation (14) is known as the influence function for $\boldsymbol{\eta}$ and implies $\mathbb{E}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\right) = \mathbf{0}_{4m+1}$. By applying the weak law of large numbers, we conclude $\hat{\boldsymbol{\eta}} \to_p \boldsymbol{\eta}^*$ implying $\hat{\tau}_{\text{CAL}} \to_p \tau_0$.

Another result of M-estimation theory shows that under the weak law of large numbers,

$$n^{-1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \to_d \mathcal{N}(\mathbf{0}_{4m+1}, \boldsymbol{\Sigma}^*)$$

where

$$\boldsymbol{\Sigma}^* = \mathbb{E}\left[\frac{\partial \psi(S_i, \mathbf{X}_i, Y_i, Z_i; \boldsymbol{\eta}^*)}{\partial \boldsymbol{\eta}}\right]^{-1} \mathbb{E}\left[\psi(S_i, \mathbf{X}_i, Y_i, Z_i; \boldsymbol{\eta}^*)^{\otimes 2}\right] \mathbb{E}\left[\frac{\partial \psi(S_i, \mathbf{X}_i, Y_i, Z_i; \boldsymbol{\eta}^*)}{\partial \boldsymbol{\eta}}\right]^{-T}.$$

Therefore, the robust variance estimator we use is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\left[\sum_{i=1}^{n} \frac{\partial \psi(S_i, \mathbf{X}_i, Y_i, Z_i; \hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}}\right]^{-1} \left[\sum_{i=1}^{n} \psi(S_i, \mathbf{X}_i, Y_i, Z_i; \hat{\boldsymbol{\eta}})^{\otimes 2}\right] \left[\sum_{i=1}^{n} \frac{\partial \psi(S_i, \mathbf{X}_i, Y_i, Z_i; \hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}}\right]^{-T}.$$

B Estimating Equations for Data-Fusion

Data-fusion requires slight modifications to some of the estimating equations in A. The score equations for $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ are instead

$$\begin{aligned}
\boldsymbol{\xi}_0(S_i, \mathbf{X}_i, Z_i; \boldsymbol{\gamma}, \boldsymbol{\lambda}_0, \boldsymbol{\theta}) \equiv\ & S_i(1 - Z_i) \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)(\gamma_j + \lambda_{j0})\right] [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}] \\
& + (1 - S_i)(1 - Z_i) \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j0}\right] [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}] \\
\boldsymbol{\xi}_1(S_i, \mathbf{X}_i, Z_i; \boldsymbol{\gamma}, \boldsymbol{\lambda}_1, \boldsymbol{\theta}) \equiv\ & S_i Z_i q \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)(\gamma_j + \lambda_{j1})\right] [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}] \\
& + (1 - S_i) Z_i \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j1}\right] [\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta}].
\end{aligned}$$

Following a similar theme, we also rewrite the score equation for τ_0 as

$$\begin{aligned}\phi(\mathbf{X}_i, Y_i, Z_i; \boldsymbol{\gamma}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \tau) \equiv S_i \Bigg\{ & Z_i \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)(\gamma_j + \lambda_{j1}) \right] [Y_i(1) - \tau] \\ & - (1 - Z_i) \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)(\gamma_j + \lambda_{j0}) \right] Y_i(0) \Bigg\} \\ & + (1 - S_i) \Bigg\{ Z_i \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j1} \right] [Y_i(1) - \tau] \\ & - (1 - Z_i) \exp\left[-\sum_{j=1}^{m} c_j(\mathbf{X}_i)\lambda_{j0} \right] Y_i(0) \Bigg\}.\end{aligned}$$

The proof is the same as in Appendix A using these updated equations.

Acknowledgements

Fan Yang was supported in part by the National Science Foundation, NSF SES-1659935, Debashis Ghosh was supported in part by the National Science Foundation, NSF DMS-1914937, and Sridharan Raghavan was supported in part by the US Department of Veterans Affairs Award IK2-CX001907-01.

Disclaimer

This manuscript was submitted to the Department of Biostatistics and Informatics in the Colorado School of Public Health, University of Colorado Anschutz Medical Campus, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biostatistics for Kevin Josey.

Supplementary Material

The R package used to fit balancing and sampling weights is in development with a working version available at https://github.com/kevjosey/cbal/. The code used to conduct the simulation study in Section 5 is available at the following URL: https://github.com/kevjosey/fusion-sim/.

BARI 2D study data is publicly available through the US National Institutes of Health, National Heart, Lung, and Blood Institute's Biologic Specimen and Data Repository Information Coordinating Center (https://biolincc.nhlbi.nih.gov/studies/bari2d/).

VA diabetes patient data included in this study are available on reasonable request to SR and upon obtaining required regulatory approvals according to current VA guidelines. Due to the sensitivity of the clinical data collected for this study, data requests must be from qualified researchers with approved human subjects research protocols.

Tables and Figures

n	Outcome Scenario	Treatment Scenario	Sampling Scenario	τ_0	Transportability			Data-Fusion	
					TMLE	AUG	CAL	AUG	CAL
1000	a	a	a	3.28	3.28 (3.75)	3.28 (0.42)	3.29 (0.53)	3.28 (0.34)	3.28 (0.33)
1000	a	a	b	4.06	4.06 (0.41)	4.06 (0.35)	4.06 (0.37)	4.06 (0.31)	4.06 (0.31)
1000	a	b	a	3.26	3.26 (0.58)	3.26 (0.4)	3.27 (0.48)	3.26 (0.33)	3.26 (0.33)
1000	a	b	b	4.08	4.07 (0.35)	4.07 (0.34)	4.07 (0.35)	4.07 (0.31)	4.07 (0.31)
1000	b	a	a	4.08	10.04 (9.74)	2.28 (0.71)	4.13 (0.91)	3.68 (0.42)	4.04 (0.35)
1000	b	a	b	3.36	4.88 (0.58)	4.55 (0.38)	4.70 (0.41)	3.84 (0.27)	3.57 (0.31)
1000	b	b	a	4.09	5.87 (1.31)	3.60 (0.67)	4.95 (0.76)	4.87 (0.38)	4.94 (0.33)
1000	b	b	b	3.36	5.18 (0.36)	5.14 (0.34)	5.20 (0.34)	4.68 (0.27)	4.63 (0.29)
2000	a	a	a	3.27	3.24 (1.39)	3.28 (0.30)	3.27 (0.38)	3.28 (0.25)	3.27 (0.24)
2000	a	a	b	4.06	4.06 (0.29)	4.06 (0.25)	4.06 (0.26)	4.06 (0.22)	4.06 (0.23)
2000	a	b	a	3.28	3.28 (0.41)	3.27 (0.29)	3.28 (0.37)	3.28 (0.24)	3.28 (0.23)
2000	a	b	b	4.07	4.08 (0.25)	4.08 (0.24)	4.08 (0.24)	4.07 (0.21)	4.07 (0.22)
2000	b	a	a	4.10	8.47 (9.23)	2.25 (0.50)	4.12 (0.64)	3.68 (0.32)	4.05 (0.26)
2000	b	a	b	3.38	4.83 (0.40)	4.54 (0.28)	4.68 (0.30)	3.84 (0.21)	3.57 (0.22)
2000	b	b	a	4.09	5.61 (0.88)	3.55 (0.44)	4.91 (0.56)	4.83 (0.27)	4.91 (0.24)
2000	b	b	b	3.38	5.20 (0.26)	5.17 (0.24)	5.22 (0.25)	4.71 (0.20)	4.65 (0.21)

Table 1: Average estimate and Monte Carlo standard error under the different simulation scenarios of the treatment assignment and outcome processes. The target population average treatment effect is evaluated under either the transportability or the data-fusion settings.

n	Outcome Scenario	Treatment Scenario	Sampling Scenario	τ_0	Transportability	Data-Fusion
1000	a	a	a	3.28	0.901	0.944
1000	a	a	b	4.06	0.940	0.953
1000	a	b	a	3.26	0.922	0.941
1000	a	b	b	4.08	0.943	0.948
1000	b	a	a	4.08	0.848	0.964
1000	b	a	b	3.36	0.095	0.955
1000	b	b	a	4.09	0.631	0.290
1000	b	b	b	3.36	0.002	0.028
2000	a	a	a	3.27	0.922	0.954
2000	a	a	b	4.06	0.933	0.931
2000	a	b	a	3.28	0.911	0.939
2000	a	b	b	4.07	0.944	0.946
2000	b	a	a	4.10	0.883	0.952
2000	b	a	b	3.38	0.017	0.928
2000	b	b	a	4.09	0.525	0.099
2000	b	b	b	3.38	0.000	0.000

Table 2: Coverage probabilities of the target population average treatment effect for the full calibration approaches described in Section 4.

	Metformin (2004-2009)	Metformin (2010-2014)	Sulfonylurea (2004-2009)	Sulfonylurea (2010-2014)
Patient Count	84003	100612	29447	11736
Baseline Age	61.90 (11.64)	60.45 (11.50)	67.32 (12.47)	66.95 (12.72)
Male	80964 (96.4)	95906 (95.3)	28912 (98.2)	11472 (97.8)
Race/Ethnicity				
Non-hispanic White	11927 (14.2)	20132 (20.0)	4038 (13.7)	2253 (19.2)
Non-hispanic Black	4730 (5.6)	6810 (6.8)	1517 (5.2)	614 (5.2)
Hispanic	10897 (13.0)	8583 (8.5)	5073 (17.2)	1123 (9.6)
Other	56449 (67.2)	65087 (64.7)	18819 (63.9)	7746 (66.0)
Smoking Status				
Current	21475 (25.6)	30603 (30.4)	6365 (21.6)	2925 (24.9)
Former	42751 (50.9)	44354 (44.1)	16396 (55.7)	5935 (50.6)
Never	19777 (23.5)	25655 (25.5)	6686 (22.7)	2876 (24.5)
BMI	32.96 (6.43)	33.66 (6.52)	31.20 (6.00)	32.08 (6.22)
SBP	133.03 (16.61)	132.35 (15.93)	133.83 (18.37)	131.80 (17.31)
DBP	76.83 (10.82)	78.69 (10.76)	74.90 (11.59)	75.81 (11.45)
HDL	39.19 (10.76)	40.14 (10.97)	39.23 (11.51)	39.60 (11.50)
LDL	103.43 (34.93)	102.68 (35.30)	101.32 (35.46)	96.99 (35.06)
Total Cholesterol	179.60 (44.15)	178.57 (44.51)	177.91 (46.41)	173.49 (46.68)
Triglycerides	205.36 (192.15)	203.67 (198.23)	206.01 (198.88)	208.68 (216.49)
Fasting Plasma Glucose	151.54 (62.26)	149.71 (63.64)	165.87 (80.76)	163.67 (78.23)
HbA1c	7.17 (1.43)	7.28 (1.41)	7.42 (1.63)	7.54 (1.54)
Estimated GFR	78.41 (18.59)	83.88 (20.08)	66.83 (22.97)	66.15 (24.76)
Creatinine	1.03 (0.20)	0.98 (0.20)	1.25 (0.48)	1.29 (0.56)
History of CAD	31211 (37.2)	25771 (25.6)	14071 (47.8)	4535 (38.6)
History of CHF	8768 (10.4)	6086 (6.0)	5916 (20.1)	1752 (14.9)
History of Stroke	10572 (12.6)	8385 (8.3)	5230 (17.8)	1645 (14.0)
History of Kidney Disease	571 (0.7)	237 (0.2)	1214 (4.1)	316 (2.7)
History of Liver Disease	424 (0.5)	349 (0.3)	278 (0.9)	86 (0.7)

Table 3: Summary statistics for covariates measured on newly diagnosed diabetic patients receiving care in the VA healthcare system stratified by years (2004-2009 and 2010-2014) and monotherapy type (Metformin or Sulfonylurea).

Method and Sample	One Year after Rx	Two Years after Rx	Five Years after Rx
Unadjusted 2004-2009	2.6% (2.3%, 2.8%)	5.3% (4.9%,5.7%)	12.6% (11.9%, 13.2%)
Unadjusted 2010-2014	2.4% (2.0%, 2.7%)	5.1% (4.6%, 5.7%)	12.5% (11.7%, 13.4%)
iCBPS 2004-2009	2.3% (2.1%, 2.6%)	5.0% (4.7%, 5.4%)	12.2% (11.6%, 12.7%)
iCBPS 2010-2014	1.0% (0.7%, 1.4%)	2.0% (1.5%, 2.5%)	4.1% (3.4%, 4.9%)
Transported 2004-2009	0.8% (0.5%, 1.1%)	1.9% (1.4%, 2.3%)	4.0% (3.4%, 4.6%)
Data-Fusion	0.9% (0.7%, 1.1%)	2.1% (1.7%, 2.4%)	4.2% (3.7%, 4.7%)

Table 4: Risk differences in total mortality between sulonylurea and metformin monotherapy in a VA cohort starting from the date of first prescription (Rx) using a variety of causal effect estimators.

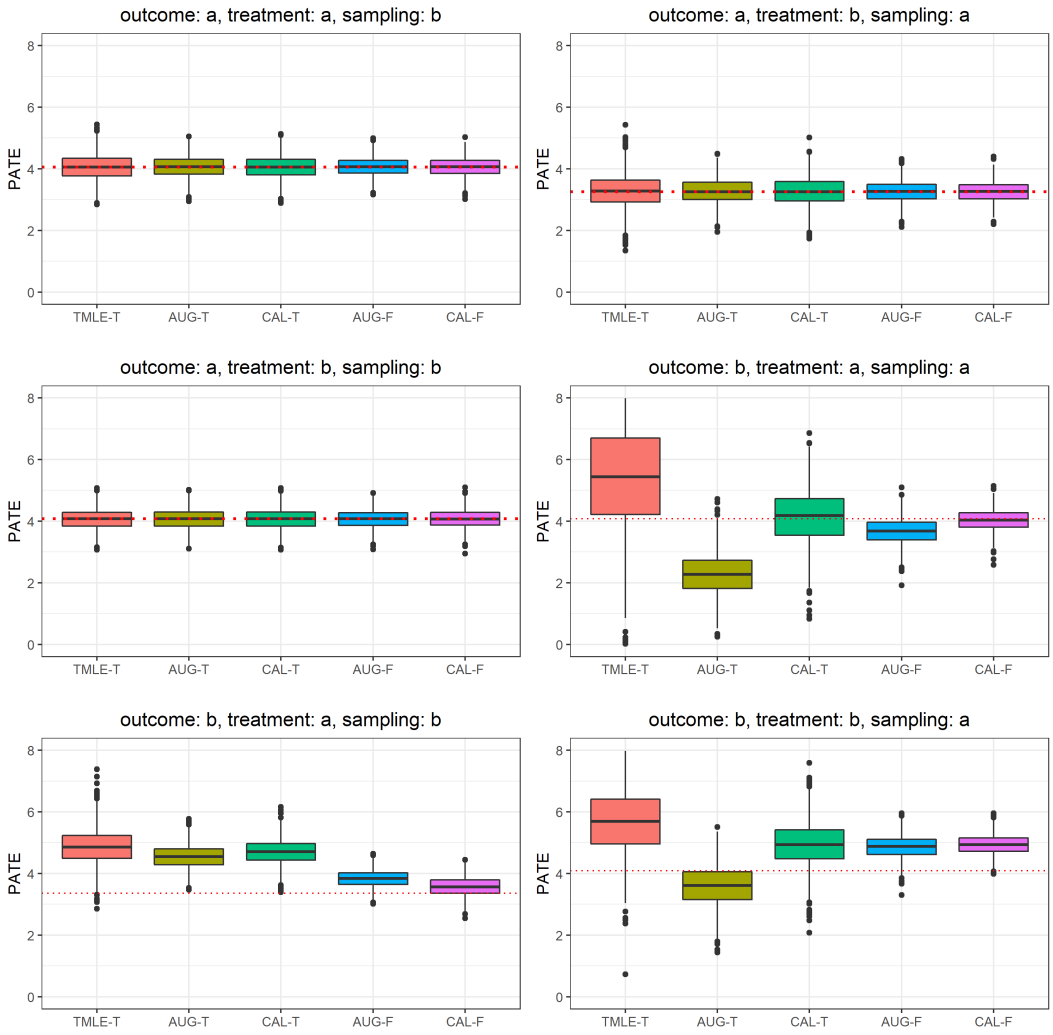

Figure 1: A subset of the constant conditional ATE estimates using four different methods for estimating balancing weights. Each boxplot is composed of 1000 estimates from the replicates that generate the values in Table 1. The suffixes -T denotes the transportability case, while -F denotes the data-fusion case.

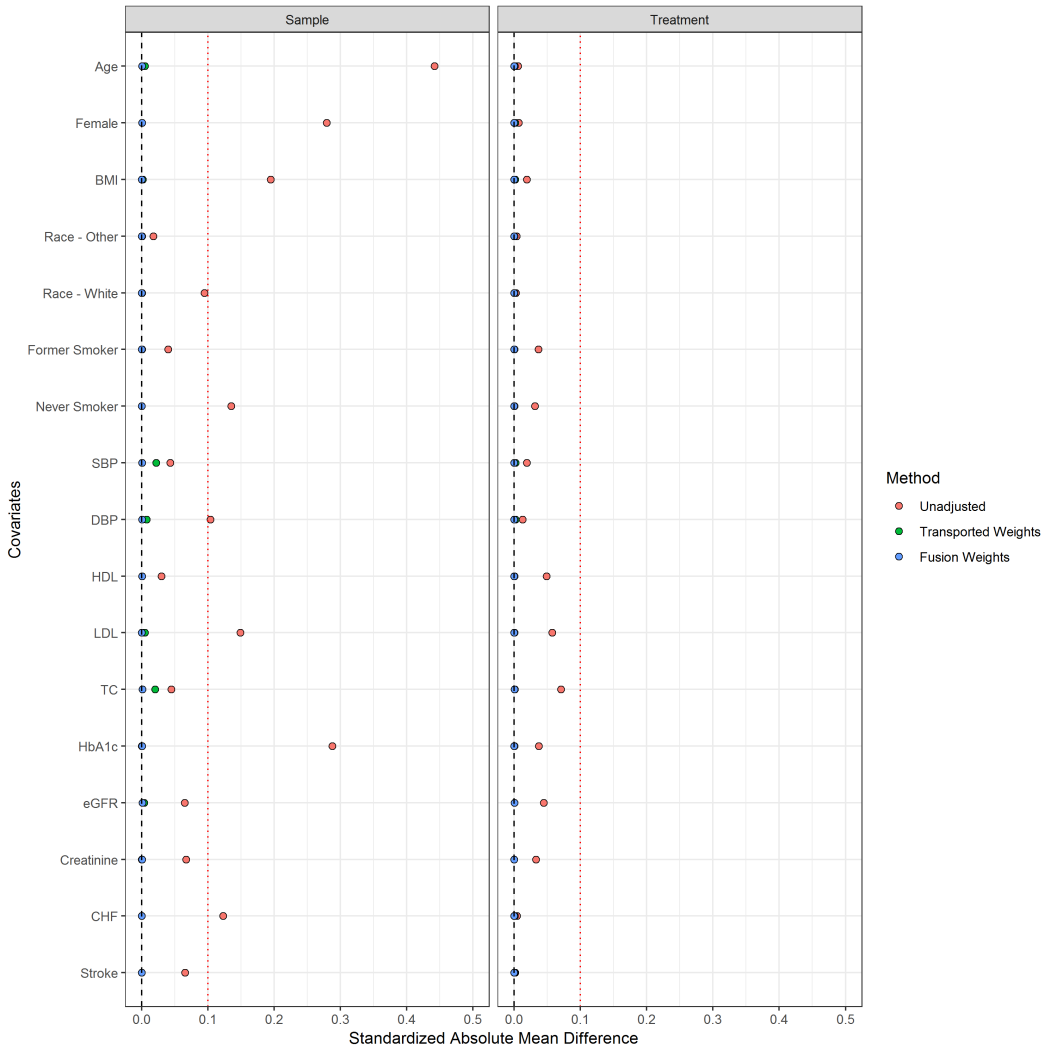

Figure 2: The standardized absolute mean differences between samples (BARI 2D versus 2010-2014 new VA diabetes patients) and treatment groups (insulin sensitization versus insulin provision) using weighting methods discussed in Section 4. The standardized mean differences between treatment groups are estimated over the patients $i \in \{S_i = 1\}$ with the Unadjusted and Transport Weights but for all $i = 1, 2, \ldots, n$ with the Fusion Weights. The entire dataset is used to find the differences between samples for all three methods.