A calibration approach to transportability and data-fusion with observational data

Kevin P. Josey | Fan Yang | Debashis Ghosh | Sridharan Raghavan

1Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
2Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA
3Rocky Mountain Regional VA Medical Center, Aurora, Colorado, USA
4Division of Hospital Medicine, University of Colorado School of Medicine, Aurora, Colorado, USA

Correspondence
Sridharan Raghavan, Rocky Mountain Regional VA Medical Center, 1700 N Wheeling St, Aurora, CO 80045, USA.
Email: sridharan.raghavan@va.gov

Funding information
Division of Mathematical Sciences, Grant/Award Number: DMS-1914937;
Division of Social and Economic Sciences, Grant/Award Number: SES-1659935;
National Institute of Environomental Health Sciences, Grant/Award Number: ST32ES007142; U.S. Department of Veterans Affairs, Grant/Award Number: IK2-CX001907-01

1 INTRODUCTION

Two important considerations in clinical research studies are proper evaluations 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. These issues persist for another problem closely related to transportability known as data-fusion. We develop a calibration method to generate balancing weights that address confounding and sampling bias, thereby enabling valid estimation of the target population average treatment effect. We compare the calibration approach to two additional doubly robust methods that estimate the effect of an intervention on an outcome within a second, possibly unrelated target population. The proposed methodologies can be extended to resolve data-fusion problems that seek to evaluate the effects of an intervention using data from two related studies sampled from different populations. A simulation study is conducted to demonstrate the advantages and similarities of the different techniques. We also test the performance of the calibration approach in a motivating real data example comparing whether the effect of biguanides vs 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.

KEYWORDS
causal inference, covariate balance, data-fusion, transportability, type 2 diabetes
from a study, the effect estimates may not be accurate for the population of interest. This discordance occurs when the population that is sampled for an observational study or randomized trial—that is, the study population—diverges from the specific population of interest for applying the study results—that is, 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 target population containing every patient with the disease is of practical importance. For much of this article, we focus on a more general yet well-defined setup for generalizing results onto a target population known as transportability.1 Whereas generalizability requires the study population to be nested entirely within the target population, transportability allows the study and target populations to be disjoint. More information about the distinction between generalizability and transportability can be found elsewhere.1,2 We then transition our attention to the problem of data-fusion,3 a closely related problem to transportability, wherein inferences occur on the target population and data for the outcome and treatment assignment are available from two respective samples of the target and study populations. Relative to transportability, where treatment and outcome data are only available for units in the study sample, the data-fusion estimator of a target population causal effect should be more efficient given the greater availability of data.

Methods for causal inference and transporting effect estimates to a target population are two recent topics of substantial 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 in 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. Conditioning a causal effect estimator on the propensity score in essence balances the distribution of confounders between the exposed and unexposed participants in an observational study, thereby eliminating confounding bias.4 Through a similar mechanism, by conditioning a causal effect estimator on the sampling score, or the probability of being sampled for a study given the measured covariates, we can transport study results onto a target population in the presence of effect modification.1 Extensions and comparisons of methods for transporting observational study results are limited, with most methods opting to focus on transporting results from a randomized controlled trial.2,5-7 Moreover, there is little discussion on extensions for many of these methods to reconcile differences between two observational studies which examine the same exposure/outcome relationship, as in the data-fusion problem.3 Further compounding the issues already facing transportability and data-fusion, fitting parametric models of the propensity and sampling scores with maximum likelihood estimation has several limitations, particularly in finite sample settings. One of these issues is model misspecification.8 Therefore, the proposed implementations for extending inferences, either through transportability or data-fusion, should complement methodological developments found in the causal inference literature that overcome some of these limitations, such as double-robustness to help ameliorate model misspecification.

Calibration estimators9 have had recent success in both transporting causal effects from a randomized controlled trial to a target population and estimating causal effects using observational data. We propose combining the approach of Chan et al10 and Josey et al,11 which finds balancing weights that correct for confounding bias, with an exponential tilting estimator that estimates sampling weights to remove any bias attributable to differences in the covariate distribution between the study participants and nonparticipants sampled from the target population.12 The exponential tilting approach to estimate sampling weights was used by Lee et al6 to generalize randomized controlled trial results onto a target population characterized by an observational cohort. Their solution estimates the target population average treatment effect using a class of estimators that augment models for the propensity and sampling scores with a model of the outcome process.13 However, they stop short of identifying calibration weights that will also balance the covariate distribution between the treatment groups. To supplement their results, we propose a full calibration approach for transporting causal effects that produces weights that balance the covariate distribution between the treatment groups and between the study and target samples. We then show how the full calibration approach can be adapted to solve the data-fusion problem, enabling us to combine two observational studies that compare the same exposure and outcome. In the data-fusion setting, we will need to carefully consider additional assumptions appended to those already required for transportability. We demonstrate that the full calibration approach is doubly robust, meaning that if either the potential outcome model is correctly specified or both the propensity and sampling score models are correctly specified, then the full calibration estimator of the target population average treatment effect is consistent.

To help showcase the proposed calibration solutions, we will evaluate effect estimates for diabetes treatment derived from older observational data on a contemporary real-world patient sample. This scenario is well-suited for
transportability and data-fusion as diabetes patient characteristics, which potentially modify the effects of the initial diabetes treatments, have changed substantially over the last 10 to 20 years.\textsuperscript{14-20} The analysis here focuses on the comparative effectiveness of metformin and sulfonylureas—the two most commonly used drugs/classes for initial diabetes treatment in the US\textsuperscript{21-23}—as first-line diabetes treatments. Several large observational studies and clinical trials comparing metformin and sulfonylureas have found that sulfonylurea use may be associated with poorer cardiovascular and mortality outcomes than metformin.\textsuperscript{24-28} Consequently, sulfonylurea use has declined over the last decade. However, they remain inexpensive and effective diabetes medications, making them an attractive treatment option in the right patients—for example, those at low risk of treatment related adverse events. With changing patient characteristics and more effective patient selection for sulfonylurea treatment, it is possible that the comparative effectiveness of sulfonylureas relative to metformin in real-world use has improved over time. We apply the transportability and data-fusion methods described in this article to the problem of updating effect estimates of metformin vs sulfonylureas as a first-line diabetes treatment on all-cause mortality in the United States Veterans Affairs (VA) population. Using data available from 2004 to 2009, we will transport the risk difference of mortality from initial monotherapy with either a sulfonylurea or metformin regimen to a more contemporary cohort of diabetes patients diagnosed between 2010 and 2014. The estimated effects in the 2010 to 2014 sample will serve as a benchmark for the 2004 to 2009 cohort estimates transported to the 2010 to 2014 population. We also use the data-fusion approach to combine the 2004 to 2009 and 2010 to 2014 cohorts to estimate the treatment effect in the 2010 to 2014 VA diabetic population. Understanding how temporal changes in a health system’s patient population can modify the effect estimates from treatment choices for diabetes could have substantial population health impacts.

The remainder of this article is structured as follows. In Section 2, we introduce the notation and several assumptions that will be referenced throughout. In Section 3, we describe various methods for transporting causal effects in observational settings, including our proposed full calibration approach. In Section 4, we propose an extension to the full calibration method in Section 3 so that it is applicable in a data-fusion setting that requires combining observational studies. In Section 5, we compare the methods described in Sections 3 and 4 through simulations. Section 6 contains a data analysis using the full calibration approaches to transportability and data-fusion. Finally, we conclude with a discussion in Section 7.

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 nonparticipant or participant. We sometimes refer to units $\{i : S_i = 1\}$ as the study sample and units $\{i : S_i = 0\}$ as the target sample. 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 nonparticipants in the target sample represent a random sample from the target population—the population we would like to infer upon. The study sample is representative of the study population which is fundamentally different than the target population.

For each $i = 1, 2, \ldots, n$, let $X_i \in \mathcal{X}$ denote a vector of measured covariates, $Y_i \in \mathcal{Y}$ denote the outcome and $Z_i \in \{0, 1\}$ denote the treatment assignment. We apply the potential outcomes framework\textsuperscript{29} to identify the causal estimand of interest and the necessary assumptions for transportability.\textsuperscript{30} Let $Y_i(0)$ denote the potential outcome when $Z_i = 0$ and $Y_i(1)$ denote the potential outcome when $Z_i = 1$. The conditional expectations of these potential outcomes are denoted $\mu_0(\cdot) \equiv \mathbb{E}[Y_i(0) | \cdot]$ and $\mu_1(\cdot) \equiv \mathbb{E}[Y_i(1) | \cdot]$. The target population average treatment effect is defined as $\tau_0 \equiv \mathbb{E}[Y_i(1) - Y_i(0) | S_i = 0]$. $\tau_0$ is the estimand of interest for both transportability and data-fusion.

Conditioned on $X_i$, we set $\rho(X_i) \equiv \Pr(S_i = 1 | X_i)$, $\pi_1(X_i) \equiv \Pr(Z_i = 1 | S_i = 1, X_i)$, and $\pi_0(X_i) \equiv \Pr(Z_i = 1 | S_i = 0, X_i)$ 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

$$\pi_i(X_i) \equiv \Pr(Z_i = 1 | S_i = 1, X_i) \equiv s\pi_1(X_i) + (1 - s)\pi_0(X_i).$$

Define $\{c_j(X) : j = 1, 2, \ldots, m\}$ as the set of functions that generate linearly independent features to be balanced between treatment groups and the two samples. We will refer to these quantities as \textit{balance functions}. Furthermore, we will assume
\(c_1(X_i) = 1\) for all \(i = 1, 2, \ldots, n\). The target sample moments of the balance functions are defined as \(\hat{\theta}_0 = n^{-1} \sum_{i=1}^n (1 - S_i) c_j(X_i)\), which is a consistent and unbiased estimator for \(\theta_0 \equiv E[c_j(X_i)|S_i = 0]\) for all \(j = 1, 2, \ldots, m\). Finally, we define the logit link function as \(g(w) = \log[w/(1 - w)]\) and its inverse \(g^{-1}(w) = \exp(w)/[1 + \exp(w)]\).

### 2.2 Assumptions for transportability and data-fusion

#### 2.2.1 Nonparametric identifiability assumptions

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 methods are applicable using observational data.\(^3,31\) We frame the setup to both problems through the following set of assumptions. These assumptions are an extension to those proposed in other articles regarding the transportability of experimental results across populations.\(^5,6\) We combined these assumptions with the assumptions necessary for conducting causal inference in the presence of confounding.\(^29\) To begin, we invoke the stable unit treatment value assumption which requires consistency, \(Y_i(z) = Y_i\) when \(Z_i = z\), and no interference between units.\(^29\) This means the observed outcome is equivalent to \(Y_i \equiv Z_i Y_i(1) + (1 - Z_i) Y_i(0)\).

**Assumption 1** (Strongly ignorable treatment assignment). In the case of transportability, the potential outcomes among the study participants are independent of the treatment assignment given \(X_i\), and therefore: \(E[Y_i(z)|S_i = 1, X_i] = E[Y_i(z)|S_i = 1, X_i]\) for both \(z \in \{0, 1\}\). For data-fusion, this condition is required for both study participants and nonparticipants: \(E[Y_i(z)|S_i, X_i, Z_i] = E[Y_i(z)|S_i, X_i]\) for both \(z \in \{0, 1\}\).

**Assumption 2** (Treatment effect exchangeability). Among all independent sampling units in either the study sample or the target sample, the expected value of the individual treatment effects conditioned on the covariates are exchangeable between samples: \(E[Y_i(1) - Y_i(0)|S_i, X_i] = E[Y_i(1) - Y_i(0)|X_i]\) for all \(i = 1, 2, \ldots, n\).

**Assumption 3** (Sample positivity). The probability of nonstudy participation, conditioned on the baseline covariates necessary to ensure treatment effect exchangeability, is bounded below by zero: \(Pr\{S_i = 0|X_i\} > 0\) for all \(i = 1, 2, \ldots, n\).

**Assumption 4** (Treatment positivity). For transportability, the probability of treatment conditioned on the baseline covariates in the study sample is bounded between zero and one: \(0 < Pr\{Z_i = 1|S_i = 1, X_i\} < 1\) for all \(i = 1, 2, \ldots, n\). For data-fusion, this condition is required for the both the study and target samples: \(0 < Pr\{Z_i = 1|S_i, X_i\} < 1\).

#### 2.2.2 Parametric assumptions for calibration estimators

In addition to Assumptions 1 to 4, the following set of assumptions are necessary to establish the doubly robust properties of the calibration weighted estimates for transportability and data-fusion. For more context, we will show that if either Assumption 5 is satisfied or both Assumptions 6 and 7 hold, then the so-called full calibration method that we present in Section 3.3 is consistent. If 6 and 7 hold, then 5 is not required to achieve statistical consistency. Likewise, if Assumption 5 holds then 6 and 7 are not required to ensure statistical consistency.

**Assumption 5** (Conditional linearity of the potential outcomes). The expected value of the potential outcomes, conditioned on \(X_i\), is linear across the span of the balancing functions while adhering to treatment effect exchangeability (Assumption 2):

\[
\mu_0(S_i = s, X_i) = \sum_{j=1}^m c_j(X_i) \beta_{s,j} \quad \text{and} \quad \mu_1(S_i = s, X_i) = \mu_0(S_i = s, X_i) + \sum_{j=1}^m c_j(X_i) \alpha_j
\]

for all \(i = 1, 2, \ldots, n\) with \(\alpha_j, \beta_{s,j} \in \mathcal{R}\) for all \(j = 1, 2, \ldots, m\) and \(s \in \{0, 1\}\).

**Assumption 6** (Conditional linear log-odds for sampling). The log-odds of being in the study sample vs the target sample are linear across the span of the covariates: \(g(\rho(X_i)) = \sum_{j=1}^m c_j(X_i) \gamma_j\) for all \(i = 1, 2, \ldots, n\) and \(\gamma_j \in \mathcal{R}\) for all \(j = 1, 2, \ldots, m\).
Assumption 7  (Conditional linear log-odds for treatment). The probability of treatment in the study sample and the target sample are linear across the span of the covariates:

\[
g[\pi_i(X_i)] = s \sum_{j=1}^{m} c_j(X_i) \lambda_{ij} + (1 - s) \sum_{j=1}^{m} c_j(X_i) \lambda_{ij}
\]

for all \( i = 1, 2 \ldots, n \) and \( j = 1, 2, \ldots, m \) with \( \lambda_{ij}, \lambda_{ij} \in \mathbb{R} \) and \( S_i = s \).

The linearity conditions found in Assumptions 5 to 7 are relatively strong. Whereas the calibration estimators require the above representations of the potential outcome model, or the sampling and propensity scores, the alternative methods presented in Sections 3.1 and 3.2 have less stringent requirements to achieve statistical consistency. We will show that the targeted maximum likelihood approach makes the fewest parametric assumptions of all the methods that we will examine whereas the full calibration approach in Section 3.3 makes the most parametric assumptions for identification. Having fewer parametric assumptions allows for more intricate modeling choices, including the ability to employ machine learning techniques. However, these methods lose some interpretability provided by the calibration weights, which will exactly balance whichever distributional features are specified in the primal problem between the study and target samples and between the treatment groups—a quality often desired in causal analyses. While the TMLE and augmented approaches we present later on do not require parametric assumptions for identification, the methods used to estimate the nuisance parameters can be parametric. If the nuisance parameters are estimated with parametric models, then the TMLE and augmented approaches are also making the same strong parametric assumptions as the calibration approach. We demonstrate this fact in our simulation study in order to evaluate the doubly robust properties of the various estimators outlined for \( r_0 \).

2.2.3  Data-fusion specific assumptions

The distinction between transportability and data-fusion essentially amounts to how much data we are provided from the study and target samples. For problems of transportability, we require the complete individual-level data from study sample, but only the individual-level covariate data from the target sample (i.e., \( X_i \) for all \( i : S_i = 0 \)). In data-fusion, both the study and target samples provide data on \( X_i, Y_i, \) and \( Z_i \) 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 the target sample leaving data-fusion impracticable. In those cases, transportability is an appealing alternative.

Assumption 8  (Potential outcome exchangeability). Among all independent sampling units in either the study or target sample, the expected value of the potential outcomes conditioned on the covariates are exchangeable between samples:

\[
\mathbb{E}[Y_i(z)|S_i, X_i] = \mathbb{E}[Y_i(z)|X_i] \text{ for all } z \in \{0, 1\}.
\]

Assumption 9  (Propensity score exchangeability). For all \( X_i \in \mathcal{X} \), we have \( \pi_i(X_i) = \pi_0(X_i) \).

The exchangeability assumptions of Assumptions 8 and 9 are key when considering the model design for estimators of \( r_0 \) in the data-fusion setting. Potential outcome exchangeability is a stronger assumption than the assumptions of treatment effect exchangeability (Assumption 2). The former requires the expected values of the potential outcomes be exchangeable, conditional on the effect modifiers, confounders, and prognostic variables. We refer to prognostic variables as covariate measurements that are predictive of the outcome, but are otherwise uncorrelated with the treatment to be exclusive of confounders. The treatment effect exchangeability assumption, on the other hand, requires exchangeability conditional on both the effect modifiers to eliminate sampling bias and the confounders to eliminate confounding bias. Potential outcome exchangeability is a key limitation to more general approaches that impute potential outcomes for specific populations without necessarily evaluating a treatment effect. Since comparable methods like the augmented approach and the target maximum likelihood approach that we will introduce in the next section require estimates of \( \mu_0(S_i, X_i) \) and \( \mu_1(S_i, X_i) \), potential outcome exchangeability becomes important if not necessary whenever both the study and target samples are used to estimate \( r_0 \). Propensity score exchangeability assumes the propensity score is the same across both populations. This assumption is not an issue for transporting observational study results since these problems implicitly assume that the treatment assignment mechanism is the same for both samples. This is not an unreasonable assumption given that \( Z_i \) is not observed in the target sample in the transportability case. In the data-fusion setting, where the distribution of the treatment assignment can vary between samples, propensity score exchangeability becomes an
important consideration. If both Assumptions 8 and 9 are violated, then $S_i$ behaves like a surrogate confounder$^{35}$ and omitting it from either the potential outcome models or the propensity score model may lead to biased effect estimates.

# METHODS FOR TRANSPORTING OBSERVATIONAL RESULTS

## Targeted maximum likelihood

Targeted maximum likelihood estimation (TMLE) has emerged as a flexible framework for estimating a variety of causal estimands.$^{36}$ Specifically, Rudolph and van der Laan$^5$ apply this framework to estimate $\tau_0$ within the transportability setting described in Section 2.2. TMLE is adapted from the G-computation family of estimators which finds $\hat{\mu}_0(S_i = 1, X_i) \equiv \hat{E}(Y_i | S_i = 1, X_i, Z_i = 0)$ and $\hat{\mu}_1(S_i = 1, X_i) \equiv \hat{E}(Y_i | S_i = 1, X_i, Z_i = 1)$ using the independent sampling units $\{i : S_i = 1\}$ to solve for

$$\hat{\tau}_G = \frac{1}{n_0} \sum_{i: S_i = 0} \left[ \hat{\mu}_1(S_i = 1, X_i) - \hat{\mu}_0(S_i = 1, X_i) \right]. \quad (1)$$

For this estimator, if we can show $\hat{\mu}_1(S_i = 1, X_i) \rightarrow \mu_1(S_i = 1, X_i)$ and $\hat{\mu}_0(S_i = 1, X_i) \rightarrow \mu_0(S_i = 1, X_i)$, then given Assumptions 1 and 2 we have $\hat{\tau}_G \rightarrow \tau_0$.

Rudolph and van der Laan$^5$ extend this intuitive approach to account for potential bias induced from misspecifying $\hat{\mu}_0(S_i = 1, X_i)$ and $\hat{\mu}_1(S_i = 1, X_i)$. First, the proposed TMLE update will find estimates of the expected potential outcomes on the logit scale.$^{37}$ This requires standardizing $Y_i$ as $Y_i^* \equiv [Y_i - Y^-] \left[ Y^+ - Y^- \right]^{-1}$ where $Y^- \equiv \min_{i: S_i = 1}(Y_i)$ and $Y^+ \equiv \max_{i: S_i = 1}(Y_i)$ from the available outcome data. Additionally, we will also need to find $\hat{\mu}_0^*(S_i = 1, X_i) \equiv \hat{E}[Y_i^* | S_i = 1, X_i]$ and $\hat{\mu}_1^*(S_i = 1, X_i) \equiv \hat{E}[Y_i^* | S_i = 1, X_i, Z_i = 1]$. The TMLE solution updates the estimates of the conditional means for the transformed potential outcomes using consistent estimates for $\rho(X_i)$ and $\pi_1(X_i)$, which we denote as $\hat{\rho}(X_i) = \hat{P}_{\tau} \{ S_i = 1 | X_i \}$ and $\hat{\pi}_1(X_i) \equiv \hat{P}_{\tau} \{ Z_i = 1 | S_i = 1, X_i \}$. These estimators are combined into a so-called clever covariate$^{38}$ to find

$$\hat{\eta}_0^*(S_i = 1, X_i) \equiv \eta^0 = \left\{ g \left[ \hat{\mu}_0^*(S_i = 1, X_i) \right] + \hat{\epsilon}_0 \frac{[1 - \hat{\rho}(X_i)]}{\hat{\rho}(X_i)[1 - \hat{\pi}_1(X_i)]} \right\}^{\frac{1}{g^1}} \quad \text{and} \quad \hat{\eta}_1^*(S_i = 1, X_i) \equiv \eta^1 = \left\{ g \left[ \hat{\mu}_1^*(S_i = 1, X_i) \right] + \hat{\epsilon}_1 \frac{[1 - \hat{\rho}(X_i)]}{\hat{\rho}(X_i)\hat{\pi}_1(X_i)} \right\}^{\frac{1}{g^1}}.$$ 

Estimates of the fluctuation coefficients $\hat{\epsilon} \equiv (\hat{\epsilon}_0, \hat{\epsilon}_1)^T$ are found by projecting the transformed outcomes $X_i^*$ onto the clever covariates

$$\hat{h}_0(X_i) = \frac{(1 - Z_i)[1 - \hat{\rho}(X_i)]}{\hat{\rho}(X_i)[1 - \hat{\pi}_1(X_i)]} \quad \text{and} \quad \hat{h}_1(X_i) = \frac{Z_i[1 - \hat{\rho}(X_i)]}{\hat{\rho}(X_i)\hat{\pi}_1(X_i)},$$

with $\hat{h}(X_i) \equiv \left[ \hat{h}_0(X_i), \hat{h}_1(X_i) \right]^T$. In this regression, the initial estimates of the potential outcome means $\hat{\mu}_0^*(S_i = 1, X_i)$ and $\hat{\mu}_1^*(S_i = 1, X_i)$ serve as offsets. Analytically, estimates of the fluctuation coefficients $\hat{\epsilon}$ are found such that

$$\sum_{i: S_i = 1} \hat{h}(X_i) \left( Y_i^* - g^{-1} \left\{ Z_i g \left[ \hat{\mu}_1^*(S_i = 1, X_i) \right] + (1 - Z_i) g \left[ \hat{\mu}_0^*(S_i = 1, X_i) \right] + \hat{h}^T(X_i) \hat{\epsilon} \right\} \right) = 0. \quad (2)$$

Since this regression requires $Y_i^*$ and $Z_i$ to be observed, estimates are calculated using the units $\{i : S_i = 1\}$. The TMLE estimate of $\tau_0$ has a similar form to the G-computation setup, solving for

$$\hat{\tau}_{TMLE} = \frac{1}{n_0} \sum_{i: S_i = 0} \left[ \hat{\eta}_1(S_i = 1, X_i) - \hat{\eta}_0(S_i = 1, X_i) \right], \quad (3)$$

where $\hat{\eta}_0(S_i = 1, X_i) = \left[ Y^+ - Y^- \right] \hat{\eta}_0^*(S_i = 1, X_i) + Y^-$ and $\hat{\eta}_1(S_i = 1, X_i) = \left[ Y^+ - Y^- \right] \hat{\eta}_1^*(S_i = 1, X_i) + Y^-$. Given Assumptions 1 to 4, Rudolph and van der Laan$^5$ show that the TMLE estimator for $\tau_0$ is consistent if either the potential outcome
models are consistent, or if \( \hat{p}(\mathbf{X}_i) \rightarrow_p p(\mathbf{X}_i) \) and \( \hat{\pi}_1(\mathbf{X}_i) \rightarrow_p \pi_1(\mathbf{X}_i) \). Gruber and van der Laan\(^6\) argue that using the logit scale within TMLE constrains the fluctuation coefficients such that \( Y^- \leq \hat{\eta}_0(S_i = 1, \mathbf{X}_i) \leq Y^+ \) and \( Y^- \leq \hat{\eta}_1(S_i = 1, \mathbf{X}_i) \leq Y^+ \), thus avoiding extrapolations of the potential outcome mean estimates while retaining the doubly robust properties inherent with TMLE.

### 3.2 Potential outcome model augmentation

Another doubly robust method for transporting causal effects is to augment the potential outcome model with weights derived from estimates of the propensity and sampling scores, similar to the implementation of the clever covariates used in TMLE.\(^6\) This method also builds upon the intuition surrounding the G-computation estimator found in (1). Below we present a brief summary of the augmented approach relevant for the transportability setting as opposed to the generalizability setting.\(^2\) For a more complete description about how the method is applied in the generalizability setting, we suggest referring to the original text of Lee et al.\(^6\)

The augmented approach proceeds by fitting component models of the data generating mechanisms for \( S_i, Y_i, \) and \( Z_i \) given Assumptions 1 to 4. While several estimators for the other nuisance parameters will suffice, the inverse odds of sampling are specifically estimated to satisfy the primal constrained optimization problem to

\[
\begin{align*}
\text{minimize} & \quad \sum_{i:S_i=1} \left\{ q(S_i, \mathbf{X}_i) \log[q(S_i, \mathbf{X}_i)] - q(S_i, \mathbf{X}_i) \right\}, \\
\text{subject to} & \quad \sum_{i=1}^n S_i q(S_i, \mathbf{X}_i) c_j(\mathbf{X}_i) = n_1 \theta_{ij},
\end{align*}
\]  

(4)

for all \( j = 1, 2, \ldots, m \). Equation (4) can equivalently be solved with a Lagrangian dual problem which finds

\[
\hat{\gamma} = \arg \max_{\gamma \in \mathbb{R}^m} \sum_{i:S_i=1} \left\{ -\exp \left[ -S_i \sum_{j=1}^m c_j(\mathbf{X}_i) \gamma_j \right] - \sum_{j=1}^m \theta_{0j} \gamma_j \right\},
\]

(5)

where \( \hat{\gamma} \equiv (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_m)^T \). A Lagrangian dual is an unconstrained optimization problem derived by applying the Lagrangian multiplier theorem to a primal constrained convex optimization problem, like the one in (4). With the dual solution to (5), the resulting balancing weights that satisfy (4) can be found with

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

(6)

With the estimated sampling weights, the conditional mean estimates of the potential outcomes, and an appropriate propensity score model, Lee et al.\(^6\) construct an augmented estimator for \( \tau_0 \) which solves for

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

(7)

The estimated inverse odds of sampling weights have the property that \( \sum_{i=1}^n S_i \hat{q}(S_i, \mathbf{X}_i) c_j(\mathbf{X}_i) = \sum_{i=1}^n (1 - S_i) c_j(\mathbf{X}_i) \) for all \( j = 1, 2, \ldots, m \). In other words, the weighted sample moments of the balance functions in the study sample are equal to the unweighted sample moments of the balance functions in the target sample. Under Assumptions 1 to 4, the augmented estimator \( \hat{\tau}_{\text{AUG}} \) can be shown to be doubly robust.\(^6\) We can easily see this result given the following heuristic. In the scenario where \( \hat{\mu}_0(S_i = 1, \mathbf{X}_i) \rightarrow_p \mu_0(S_i = 1, \mathbf{X}_i) \), \( \hat{\mu}_1(S_i = 1, \mathbf{X}_i) \rightarrow_p \mu_1(S_i = 1, \mathbf{X}_i) \), the first summation in (7) has an expected value of zero while the second summation is consistent for \( \tau_0 \). When Assumption 6 holds, if \( \hat{\pi}_1(\mathbf{X}_i) \rightarrow_p \pi_1(\mathbf{X}_i) \), then the first sum in (7) will consistently negate the bias produced by the second summation.
3.3 A full calibration approach to transportability

Our proposed solution to the problem of transporting observational study results combines the calibration weighting approach for correcting confounding bias encountered within causal effect estimation with the calibrated sampling weights used in Section 3.2, which removes bias induced by the differences of the covariate distribution between the two samples. In other words, we estimate a vector of weights that concurrently balance both the samples and treatment groups.

The combined balancing and sampling weights are estimated contemporaneously in a manner similar to (5) and (6). The full calibration weights that balance both the treatment and sampling groups are estimated with the primal problem to

\[
\begin{align*}
\text{minimize} & \quad \sum_{i:S_i=1} \{ p(S_i,X_i,Z_i) \log [p(S_i,X_i,Z_i) - p(S_i,X_i,Z_i)] \} . \\
\text{subject to} & \quad \sum_{i=1}^{n} S_i(2Z_i - 1)p(S_i,X_i,Z_i)c_j(X_i) = 0 \text{ (propensity score) and} \\
& \quad \sum_{i=1}^{n} S_ip(S_i,X_i,Z_i)c_j(X_i) = n_1\hat{\theta}_0j \text{ (sampling score)}, \\
\end{align*}
\]

for all \( j = 1, 2, \ldots, m \). Much like with the primal-dual relationship between (4) and (5), the primal problem in (8) can be equivalently solved by finding

\[
\lambda^T, \hat{p}^T = \arg \max_{(\lambda^T, \hat{p}^T) \in \mathbb{R}^{2m}} \sum_{i:S_i=1} \left\{ -\exp \left[ -(2Z_i - 1) \sum_{j=1}^{m} c_j(X_i)\lambda_j - \sum_{j=1}^{m} c_j(X_i)\gamma_j - \sum_{j=1}^{m} \hat{\theta}_0j\gamma_j \right] \right\}
\]

where \( \lambda \equiv (\lambda_1, \lambda_2, \ldots, \lambda_m)^T \) which are then used to generate the balancing weights that satisfy (8) with

\[
\hat{p}(S_i,X_i,Z_i) = \exp \left[ -(2Z_i - 1) \sum_{j=1}^{m} c_j(X_i)\lambda_j - \sum_{j=1}^{m} c_j(X_i)\gamma_j \right].
\]

The desired estimand \( \tau_0 \) is estimated using the Hajek-type estimator

\[
\tau_{CAL} = \frac{1}{n_1} \sum_{i:S_i=1} \hat{p}(S_i,X_i,Z_i)(2Z_i - 1)Y_i.
\]

The balancing weights from (10) achieve exact balance between the treatment-specific sample moments of the balance functions in the study sample and balances the balance function sample moments between the study and target samples. By rearranging the constraints of (8) we get

\[
\sum_{i=1}^{n} S_iZ_i\hat{p}(S_i,X_i,Z_i)c_j(X_i) = \sum_{i=1}^{n} S_i(1 - Z_i)\hat{p}(S_i,X_i,Z_i)c_j(X_i) = \frac{n_1\hat{\theta}_0j}{2}.
\]

Consequently, the estimator in (9) to (11) is equivalent to the estimator proposed in Josey et al, which was intended for transporting randomized experiments. Combining the results of the sampling and balancing weights allows us to apply M-estimation techniques to show that the double-robustness property holds under Assumption 5 or under the combination of Assumptions 6 and 7. Additional details regarding modes of inference about \( \tau_0 \) with \( \tau_{CAL} \) can be found in the online supplement to Josey et al.

4 A full calibration approach for data-fusion

There has been limited discussion in the literature on how the methods for transportability presented in Section 3 might be adapted for data-fusion. Recall that for the data-fusion setting, we are provided \( X_i, Y_i, \) and \( Z_i \) for all \( i = 1, \ldots, n \). The
suggestion by Lee et al.⁶ is to use the outcome and treatment data from the target sample to find estimates $\hat{\mu}_0(S_i = 0, X_i) \equiv \hat{E}[Y_i | S_i = 0, X_i; Z_i = 0]$ and $\hat{\mu}_1(S_i = 0, X_i) \equiv \hat{E}[Y_i | S_i = 0, X_i; Z_i = 1]$, but otherwise keep the procedure in Section 3.2 the same. Invoking potential outcome exchangeability, we can find estimates of $\tau_0$ by solving for

$$
\hat{\tau}_{AUG} = \frac{1}{n_1} \sum_{i: S_i = 1} \hat{q}(S_i, X_i) \left\{ \frac{Z_i [Y_i - \hat{\mu}_1(S_i = 0, X_i)]}{\hat{r}_1(X_i)} - \frac{(1 - Z_i) [Y_i - \hat{\mu}_0(S_i = 0, X_i)]}{1 - \hat{r}_1(X_i)} \right\} + \frac{1}{n_0} \sum_{i: S_i = 0} \left[ \hat{\mu}_1(S_i = 0, X_i) - \hat{\mu}_0(S_i = 0, X_i) \right].
$$

(12)

This simple modification to (7) implies that only units $\{i : S_i = 1\}$ are used to de-bias estimates of the potential outcome means, assuming the potential outcome models are biased. The reasoning for not using all $i = 1, 2, \ldots, n$ for the first line of (12) is due to potential conflicts with propensity score exchangeability. Assuming potential outcome exchangeability holds but propensity score exchangeability is violated, then in order to use all available data to de-bias the second line of (12) requires estimates of both $\pi_0(X_i)$ in addition to $\pi_1(X_i)$. This is because $\hat{r}_1(X_i)$ would produce biased predictions from an incorrectly specified model of $\pi_0(X_i)$. For now, the estimator in (12) does not require this assumption at the cost of losing a considerable degree of efficiency that a data-fusion estimator should otherwise gain. Even if specific models for $\pi_1(X_i)$ and $\pi_0(X_i)$ are fit in cases where propensity score exchangeability is violated, (12) is still subject to potential outcome exchangeability, unlike its transportability analogue. Let’s suppose that the outcome model is correctly specified for the target sample. If potential outcome exchangeability is violated, then the first line of (12) would likely introduce unchecked bias to $\hat{\tau}_{AUG}$, even when the propensity score model of the study sample is correctly specified. This assumption could be counteracted in a similar way to counteracting violations to propensity score exchangeability—by fitting separate outcome models for the two samples and substituting them into (12) accordingly. Because of the extended exchangeability assumptions (Assumptions 8 and 9), the design decisions for data-fusion are more complex than the designs for transportability estimators.

We are intentional in constructing the full calibration data-fusion estimator such that it will not require either Assumptions 8 or 9 while still using all available data. To do this, ultimately, the full calibration approach to data-fusion will need to satisfy two additional constraints to balance treatment groups across both samples relative to the transportability approach in Section 3.3. This objective is not achieved by the primal problem in (8). The new goal is to estimate a set of dual vectors which can generate weights to balance the treatment groups while retaining the doubly robust property. Once we estimate the new calibration weights, we can simply change the index of the summation in (11) to estimate $\tau_0$ over all $i = 1, 2, \ldots, n$.

For data-fusion, the primal problem of (8) is expanded upon to become

$$
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{n} \{ p(S_i, X_i, Z_i) \log [p(S_i, X_i, Z_i)] - p(S_i, X_i, Z_i) \}, \\
\text{subject to} & \quad \sum_{i=1}^{n} S_i (2Z_i - 1)p(S_i, X_i, Z_i)c_j(X_i) = 0 \text{ (study sample propensity score)}, \\
& \quad \sum_{i=1}^{n} S_i p(S_i, X_i, Z_i)c_j(X_i) = n_1 \hat{\theta}_{ij} \text{ (sampling score)}, \\
& \quad \sum_{i=1}^{n} (1 - S_i)(2Z_i - 1)p(S_i, X_i, Z_i)c_j(X_i) = 0 \text{ (target sample propensity score)} \text{ and} \\
& \quad \sum_{i=1}^{n} (1 - S_i)p(S_i, X_i, Z_i)c_j(X_i) = n_0 \hat{\theta}_{ij} \text{ (weight stabilization)},
\end{align*}
$$

(13)

for all $j = 1, 2, \ldots, m$. Balancing the treatment specific moments in the study sample alone is not sufficient to also balance the treatment groups in the target sample that is required for data-fusion. For this reason, we added two extra constraints to the primal problem of (8) resulting in (13). Only the first three constraints would be required if we knew how to correctly specify the sampling and propensity scores, following Assumptions 6 and 7. The fourth constraint corrects the sampling variation from the target sample and is needed to ensure double-robustness when
Assumption 5 holds. This constraint stabilizes the weights in the same way that the improved covariate balance propensity score formulation, thereby making the original proposal for the covariate balance propensity score estimator doubly robust. These constraints may be relaxed and altered further when the potential outcome exchangeability or the propensity score exchangeability assumptions are fulfilled. Careful attention should also be given to the size of the influence attributable to the respective samples after weighting. Note that the sum of the weights on the right-hand side of the second and fourth equality constraints can be modified if more weight should be assigned to one sample vs the other.

Despite the changes to (8) expressed in (13), the primal-dual relationship is unaffected by the added constraints, barring any collinear balance functions. A new dual objective can be derived in the same manner as in Section 3.3 to find

\[
\left( \tilde{\lambda}_s^T , \tilde{\gamma}_s^T \right) = \arg \max_{(\lambda^T , \gamma^T ) \in \mathbb{R}^{2m_1} : \{ i : S_i = s \}} \left\{ -\exp \left[ -(2Z_i - 1) \sum_{j=1}^m c_j(X_i) \lambda_j - \sum_{j=1}^m c_j(X_i) \gamma_j \right] - \sum_{j=1}^m \theta_{0j} \gamma_j \right\} \tag{14}
\]

where \( \tilde{\lambda}_s \equiv (\tilde{\lambda}_{s1}, \tilde{\lambda}_{s2}, \ldots, \tilde{\lambda}_{sm})^T \) and \( \tilde{\gamma}_s \equiv (\tilde{\gamma}_{s1}, \tilde{\gamma}_{s2}, \ldots, \tilde{\gamma}_{sm})^T \) for \( s \in \{0, 1\} \). The balancing weights combine the four estimated dual vectors to produce a solution to (13) with

\[
\tilde{p}(S_i, X_i, Z_i) = \exp \left\{ -S_i \left[ (2Z_i - 1) \sum_{j=1}^m c_j(X_i) \tilde{\lambda}_{ij} + \sum_{j=1}^m c_j(X_i) \tilde{\gamma}_{ij} \right] - (1 - S_i) \left[ (2Z_i - 1) \sum_{j=1}^m c_j(X_i) \tilde{\lambda}_{ij} + \sum_{j=1}^m c_j(X_i) \tilde{\gamma}_{ij} \right] \right\}. \tag{15}
\]

The updated Hajek-estimator finds estimates of \( \tau_0 \) by solving for

\[
\tilde{\tau}_{\text{CAL}} = \frac{1}{n} \sum_{i=1}^n \tilde{p}(S_i, X_i, Z_i)(2Z_i - 1)Y_i. \tag{16}
\]

The summation in (16) uses all available data, \( i = 1, 2, \ldots, n \), meaning this estimator should be more efficient than the transportability proposal in (11). Neither the potential outcome exchangeability assumption nor the propensity score exchangeability assumption are necessarily required for the above approach, but they can be helpful to simplify some of the constraints in (13). This might be important in small sample settings where feasible solutions to (13) are difficult or impossible to achieve. For more information on how inference is conducted using this estimator, please refer to the online supplemental material.

## 5 | SIMULATION STUDY

### 5.1 | Simulation setup

To demonstrate the efficacy of the different methods for transporting observational results, and to demonstrate the advantages of using data-fusion when it is feasible, we will conduct a simulation study that evaluates the performance of the three doubly robust methods for transporting observational results that we have identified in Section 3 and the two data-fusion methods in Section 4. We will examine the performance of the proposed methodologies across a range of scenarios that test these methods when violations to the assumptions in Section 2.2 occur. To establish the doubly robust quality of these estimators, we must consider possibilities where either the sampling and propensity score models are misspecified or the potential outcome models are misspecified. We also consider the sensitivity of these methods to practical violations to overlap. Finally, we examine the implications of Assumptions 8 and 9, which are particularly important for data-fusion. When potential outcome exchangeability holds but propensity score exchangeability does not, then \( S_i \) can be viewed as an instrumental variable as the sample indicator would affect the treatment assignment but would not have a direct impact on the outcome process that is not through the treatment assignment. Conversely, when propensity score exchangeability holds and potential outcome exchangeability is violated, then \( S_i \) is regarded as a nonconfounding prognostic variable in the sense that it predicts the outcome but does not confound the relationship between the treatment assignment and the outcome.
The scenarios we examine vary the sample size \( n \in \{500, 1000, 2000\} \), the generative process that determines the treatment assignment through \( \pi_i(X_i) \), the potential outcome processes through \( \mu_0(S_i, X_i) \) and \( \mu_1(S_i, X_i) \), and the sampling mechanism \( \rho(X_i) \). These scenarios are detailed in Table 1. The covariate data \( X_i \equiv (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T \) are distributed as \( X_{i1}, X_{i2}, X_{i3}, X_{i4} \sim \mathcal{N}(0, 1) \) for every \( i = 1, 2, ... , n \). We also generate the transformed variables \( U_{i1} = \exp \left[ (X_{i1} + X_{i4})/2 \right] \), \( U_{i2} = X_{i2}/\left[ 1 + \exp(X_{i1}) \right] \), \( U_{i3} = \log(|X_{i2}X_{i3}|) \), and \( U_{i4} = (X_{i3} + X_{i4})^2 \) which are used in instances of model misspecification. Each entry in the vector \( U_i \equiv (U_{i1}, U_{i2}, U_{i3}, U_{i4})^T \) is standardized to have a mean of zero and marginal variances of one—the same as for \( X_i \). The sample indicators are generated assuming \( S_i \sim \text{Bern} [\rho(X_i)] \). Treatment assignments are generated by sampling from \( Z_i \sim \text{Bern} [\pi_S(X_i)] \). Finally, we generate the potential outcomes as \( Y_i(0) \sim \mathcal{N}[\mu_0(X_i, S_i), 4] \) and \( Y_i(1) \sim \mathcal{N}[\mu_1(X_i, S_i), 4] \), which we use to generate the observed outcome \( Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0) \).

In the transportability examples, we will discard both \( Y_i \) and \( Z_i \) for all \( i : S_i = 0 \), and use only \( Y_i \) and \( Z_i \) from units \( i : S_i = 1 \). The data-fusion methods will make use of \( Y_i \) and \( Z_i \) for all \( i = 1, 2, ... , n \). Regardless of whether the method estimates \( \tau_0 \) under the transportability case or the data-fusion case, each estimator is provided a design matrix containing an intercept term and the four original covariate measurements—\( X_{i1}, X_{i2}, X_{i3}, \) and \( X_{i4} \) for all \( i = 1, 2, ... , n \). The target population average treatment effect is estimated across each iteration of a given scenario using the respective estimators described in Section 3: \( \hat{\tau}_{\text{TMLE}} \), \( \hat{\tau}_{\text{AUG}} \) and \( \hat{\tau}_{\text{CAL}} \). We also find estimates of \( \tau_0 \) using the data-fusion estimators \( \hat{\tau}_{\text{AUG}} \) and \( \hat{\tau}_{\text{CAL}} \) from Section 4.

## 5.2 Simulation results

We report the average bias and root mean square error (RMSE) for each of the scenarios described in Table 1 across 1000 iterations using the estimators described in Sections 3 and 4. The results of the experiment are summarized in Table 2 for \( n = 500 \) and \( n = 2000 \). Figure 1 provides a graphical representation of the results for \( n = 1000 \). We also report the coverage probabilities in Table 3. The augmented and TMLE approaches use a plug-in estimator of the efficient influence function to find the 95% confidence intervals.\(^5\) For the full calibration approaches, the 95% confidence interval is estimated using a robust variance estimator.\(^{39}\)

In the case of transportability, it is clear that the three methods from Section 3 perform equitably in scenarios where model misspecification occurs, and when there is some superseding exchangeability violation (Assumptions 8 and 9). The former scenarios demonstrate that each method is doubly robust whereas the latter scenarios make it abundantly clear that transportability is not affected by the potential outcome exchangeability assumption nor the propensity score exchangeability assumption. The full calibration approach achieved a lower RMSE whenever the outcome model was misspecified (Scenarios E and H). When the potential outcome models are correctly specified, the trade-offs between the different methods for transporting observational results are negligible. Regarding the scenarios with practical violations to overlap, it is apparent that the full calibration approach was less vulnerable in scenarios with treatment overlap violation (Scenario C) relative to the scenarios in which sample overlap is violated (Scenario B). The augmented approach achieved the least amount of bias in scenarios where there is limited sample overlap in terms of the average bias and the RMSE. When \( n = 500 \), the augmented approach performed better than the full calibration approach in scenarios where treatment group overlap was violated (Scenario C), but when \( n = 2000 \) the degree of bias between the two methods is nearly identical. In this latter case, the full calibration approach had a smaller RMSE. Even though the full calibration approach performed worse than the augmented approach in scenarios involving a sample overlap violation, sample heterogeneity is still resolved using calibration thus marking the importance of this type of approach. This claim is further evidenced by the relatively poor results generated by the TMLE estimator in these settings. Furthermore, in Table 3, it appears that the full calibration approach retains accurate coverage probabilities of \( \tau_0 \), whereas finding the 95% confidence interval with a plug-in estimator of the influence function sometimes overestimates the variance of the target population average treatment effect estimates.

When we combine datasets for data-fusion, we might expect to get more precise estimates of \( \tau_0 \) than with an analogous transportability estimator. However, this is not the case for the data-fusion variant to the augmented approach. In many scenarios, the augmented estimator for data-fusion had a higher RMSE than the augmented estimator for transporting observational results. Recall that the proposed augmented solution has the same estimator in the data-fusion case as in the transportability case. The only difference between the two is the choice of initial outcome model. The only time that the RMSE is smaller was when some overlap violation occurs. The only occurrence where the full
| Exchangeability                     | Misspecification                      | Notes                        | Scenario | Data generating models                                                                                                                                 |
|------------------------------------|---------------------------------------|------------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| No exchangeability violation       | No model misspecification             | Baseline example             | A        | \[g(\rho(X_i)] = 0.5 - 0.5X_{i1} + 0.5X_{i2} - 0.5X_{i3} + 0.5X_{i4} \]  \[g(\pi_S(X_i)] = 0.5X_{i1} - 0.5X_{i2} + 0.5X_{i3} - 0.5X_{i4} \] \[\mu_0(S_i, X_i) = 2 - 3X_{i1} - X_{i2} + X_{i3} + 3X_{i4} \] \[\mu_1(S_i, X_i) = \mu_0(S_i, X_i) - 2 - X_{i1} + 3X_{i2} - 3X_{i3} + X_{i4} \] |
| Sample overlap violation           |                                       |                               | B        | \[g(\rho(X_i)] = 0.5 - 0.5X_{i1} + 0.5X_{i2} - 0.5X_{i3} + 0.5X_{i4} \]  \[g(\pi_S(X_i)] = 0.5X_{i1} - 0.5X_{i2} + 0.5X_{i3} - 0.5X_{i4} \] \[\mu_0(S_i, X_i) = 2 - 3U_{i1} - U_{i2} + U_{i3} + 3U_{i4} \] \[\mu_1(S_i, X_i) = \mu_0(S_i, X_i) - 2 - U_{i1} + 3U_{i2} - 3U_{i3} + U_{i4} \] |
| Outcome model misspecification     |                                       |                               |          | \[g(\rho(X_i)] = 0.5 - 0.5X_{i1} + 0.5X_{i2} - 0.5X_{i3} + 0.5X_{i4} \]  \[g(\pi_S(X_i)] = 0.5X_{i1} - 0.5X_{i2} + 0.5X_{i3} - 0.5X_{i4} \] \[\mu_0(S_i, X_i) = 2 - 3U_{i1} - U_{i2} + U_{i3} + 3U_{i4} \] \[\mu_1(S_i, X_i) = \mu_0(S_i, X_i) - 2 - U_{i1} + 3U_{i2} - 3U_{i3} + U_{i4} \] |
| Treatment overlap violation        |                                       |                               | C        | \[g(\rho(X_i)] = 0.5 - 0.5X_{i1} + 0.5X_{i2} - 0.5X_{i3} + 0.5X_{i4} \]  \[g(\pi_S(X_i)] = 0.5X_{i1} - 0.5X_{i2} + 0.5X_{i3} - 0.5X_{i4} \] \[\mu_0(S_i, X_i) = 2 - 3U_{i1} - U_{i2} + U_{i3} + 3U_{i4} \] \[\mu_1(S_i, X_i) = \mu_0(S_i, X_i) - 2 - U_{i1} + 3U_{i2} - 3U_{i3} + U_{i4} \] |
| Potential outcome exchangeability violation | Sampling and propensity score misspecification | Prognostic sampling variable | D        | \[g(\rho(X_i)] = 0.5 - 0.5U_{i1} + 0.5U_{i2} - 0.5U_{i3} + 0.5U_{i4} \]  \[g(\pi_S(X_i)] = 0.5U_{i1} - 0.5U_{i2} + 0.5U_{i3} - 0.5U_{i4} \] \[\mu_0(S_i, X_i) = S_i (2 - 3X_{i1} - X_{i2} + X_{i3} + 3X_{i4}) \] \[+ (1 - S_i) (2X_{i1} - 2X_{i2} - 2X_{i3} + 3X_{i4}) \] \[\mu_1(S_i, X_i) = \mu_0(S_i, X_i) - 2 - X_{i1} + 3X_{i2} - 3X_{i3} + X_{i4} \] |
| Propensity score exchangeability violation | Outcome model misspecification | Instrumental sampling variable | E        | \[g(\rho(X_i)] = 0.5 - 0.5X_{i1} + 0.5X_{i2} - 0.5X_{i3} + 0.5X_{i4} \]  \[g(\pi_S(X_i)] = S_i (0.5X_{i1} - 0.5X_{i2} + 2X_{i3} - 2X_{i4}) - 0.5(1 - S_i) \] \[\mu_0(S_i, X_i) = 2 - 3U_{i1} - U_{i2} + U_{i3} + 3U_{i4} \] \[\mu_1(S_i, X_i) = \mu_0(S_i, X_i) - 2 - U_{i1} + 3U_{i2} - 3U_{i3} + U_{i4} \] |
| Exchangeability | Misspecification | Notes | Scenario | Data generating models |
|-----------------|-----------------|-------|----------|-----------------------|
| Propensity score and potential outcome exchangeability violation | No model misspecification | Confounding sampling variable | F | $g(\rho(X_i)) = 0.5 - 0.5X_{i1} + 0.5X_{i2} - 0.5X_{i3} + 0.5X_{i4}$ $\mu_{0i}(S_i, X_i) = S_i (2 - 3X_{i1} - X_{i2} + X_{i3} + 3X_{i4})$ $\mu_{1i}(S_i, X_i) = \mu_{0i}(S_i, X_i) - 2 - X_{i1} + 3X_{i2} - 3X_{i3} + X_{i4}$ |
| Sampling and propensity score misspecification | | | G | $g(\rho(X_i)) = 0.5 - 0.5U_{i1} + 0.5U_{i2} - 0.5U_{i3} + 0.5U_{i4}$ $\mu_{0i}(S_i, X_i) = S_i (2 - 3U_{i1} - U_{i2} + U_{i3} + 3U_{i4})$ $\mu_{1i}(S_i, X_i) = \mu_{0i}(S_i, X_i) - 2 - U_{i1} + 3U_{i2} - 3U_{i3} + U_{i4}$ |
| Outcome model misspecification | | | H | $g(\rho(X_i)) = 0.5 - 0.5X_{i1} + 0.5X_{i2} - 0.5X_{i3} + 0.5X_{i4}$ $\mu_{0i}(S_i, X_i) = S_i (2 - 3U_{i1} - U_{i2} + U_{i3} + 3U_{i4})$ $\mu_{1i}(S_i, X_i) = \mu_{0i}(S_i, X_i) - 2 - U_{i1} + 3U_{i2} - 3U_{i3} + U_{i4}$ |
TABLE 2 Average bias (RMSE) of the various transportability and data-fusion estimators under several different simulation scenarios of the treatment assignment, sampling, and outcome processes described in Table 1

| n   | Scenario | \( \tau_0 \) | Transportability | Data-fusion |
|-----|----------|-------------|------------------|-------------|
|     |          |             | TMLE AUG CAL     | AUG CAL     |
|     |          |             |                  |             |
| 500 | A        | -4.00       | -0.01 (0.45) 0.00 (0.38) | -0.01 (0.36) 0.00 (0.38) |
|     | B        | -3.51       | 2.73 (6.10) 0.38 (2.69) | 1.95 (4.69) |
|     | C        | -2.71       | 1.33 (5.50) 0.27 (1.61) | 0.57 (2.54) |
|     | D        | -2.73       | -0.05 (0.37) -0.01 (0.35) | -0.01 (0.35) |
|     | E        | -2.71       | -0.03 (0.95) 0.00 (0.83) | 0.04 (0.72) |
|     | F        | -4.00       | 0.02 (0.46) 0.01 (0.37) | 0.02 (0.36) |
|     | G        | -2.73       | -0.03 (0.38) 0.01 (0.36) | 0.01 (0.36) |
|     | H        | -2.71       | -0.01 (1.02) 0.02 (0.83) | 0.06 (0.78) |
| 2000| A        | -4.00       | -0.01 (0.22) 0.00 (0.19) | 0.00 (0.18) |
|     | B        | -3.51       | 0.85 (4.78) 0.09 (1.63) | 0.25 (2.27) |
|     | C        | -2.71       | 0.31 (4.59) 0.11 (1.09) | 0.13 (0.92) |
|     | D        | -2.73       | -0.04 (0.19) 0.00 (0.17) | 0.00 (0.17) |
|     | E        | -2.71       | -0.04 (0.48) -0.01 (0.46) | 0.03 (0.38) |
|     | F        | -4.00       | -0.01 (0.21) 0.00 (0.19) | 0.00 (0.18) |
|     | G        | -2.73       | -0.04 (0.20) 0.00 (0.19) | 0.00 (0.19) |
|     | H        | -2.71       | -0.01 (0.47) 0.02 (0.43) | 0.05 (0.38) |

6 | EVALUATING METFORMIN VS SULFONYLUREAS AS MONOTHERAPY FOR VA DIABETES PATIENTS

The primary aim of this applied analysis is to estimate the risk difference in mortality between sulfonylurea and metformin monotherapy for the 2010 to 2014 cohort of diabetic VA patients. We first use improved covariate balance propensity score (CBPS) weights\(^{11,40}\) to estimate the risk difference using only the 2010 to 2014 cohort, which balance the covariates found in Table 4. This estimate will serve as a benchmark for the transport and data-fusion settings. We also estimate the risk difference for the 2004 to 2009 cohort without the 2010 to 2014 data, again for comparative purposes. We use (9) to (11) to find estimates of total mortality in the 2004 to 2009 sample transported to the 2010 to 2014 cohort. Since sulfonylurea use in 2004 to 2009 represents 27.0% of monotherapy recipients but only 11.8% of patients in 2010 to 2014 implies the need for data-fusion methods which account for propensity score exchangeability violations, like that of the estimator in (14) to (16).

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.\textsuperscript{25} The remaining baseline demographic and clinical characteristics about the two cohorts are summarized in Table 4.

Analyses based on covariate balance methods typically require checks to evaluate whether balance is actually achieved, both between the two samples and between the two treatment groups under investigation. A common diagnostic statistic for measuring balance is the standardized absolute mean difference evaluated over the various covariate measurements after weighting.\textsuperscript{32} We report this statistic for the unweighted differences in Table 4. With calibration, this step is redundant because the calibration weights should attain a standardized mean difference identical to zero (disregarding computational tolerance) in every feasible occasion (i.e., when a solution to the desired
primal problem exists). However, it can still be helpful from a pedagogical standpoint to visualize this result. In the online supplement, we analyze a similar dataset comparing insulin sensitization therapy with insulin provisioning therapy for Type 2 diabetes patients—transporting and combining the results of the BARI 2D study with the 2010 to 2014 VA cohort. There we show a plot that illustrates the exact balancing nature of the calibration method, both between treatments and samples evaluated with the weighted standardized absolute mean differences. This supplemental analysis also provides an example for how we can combine experimental and observational datasets for data-fusion.

Another diagnostic is to compare the estimated propensity and sampling scores to ensure that there is sufficient overlap between the two treatment groups and between the two samples. We use a nonparametric kernel density estimator of the probability density function for the fitted propensity and sampling scores to visualize the degree that the treatment and sample stratified distributions overlap. Both the propensity score and the sampling score were fit using covariate balance propensity scores. This diagnostic reveals the validity of the overlap assumptions (Assumptions 3 and 4). In Figure 2, we can see that there is indeed broad overlap between the sulfonylurea and metformin group, both in the 2004 to 2009 cohort and in the 2010 to 2014 cohort, and a considerable amount of overlap exists between samples based on the predicted sampling scores. Note that it is typically not possible to visualize the degree of overlap between treatments in the target sample under the transportability scenario, unless the treatment assignment is observed. Since our example works for both the transportability and data-fusion scenarios, we are able to estimate the propensity score in the target sample, which is displayed in Figure 2.

Table 5 contains the various risk-difference estimates that we computed on the illustrative dataset. We will primarily focus on the estimates of 5-year mortality since these figures have the greatest magnitude. Furthermore, the trends that we will report about 5-year mortality appear to be the same for 1- and 2-year mortality. Observe that the crude risk difference in 5-year mortality is similar within both the 2004 to 2009 and in the 2010 to 2014 cohorts. This implies that there is either limited or no changes to the risk difference attributable to temporal trends in treatment effectiveness. Note that a temporal effect modifier is one factor that we would not be able to accommodate without violating sample positivity as it is collinear with the sample indicator. The adjusted marginal estimates of the risk difference using CBPS reveal the importance of accounting for differences observed between the study and target cohorts. The risk difference in the 2004 to 2009 cohort is 12.2%, 95% CI = (11.6%, 12.7%), and 4.1%, 95% CI = (3.4%, 4.9%), in the 2010 to

**Table 3** Coverage probabilities of $\tau_0$ for the methods in Sections 3 and 4

| $n$   | Scenario | $\tau_0$ | Transportability | Data-fusion |
|-------|----------|----------|------------------|-------------|
|       |          |          | TMLE | AUG | CAL | AUG | CAL |
| 500   | A        | -4.00    | 0.946 | 0.946 | 0.929 | 0.954 | 0.940 |
| 500   | B        | -3.51    | 0.371 | 0.843 | 0.625 | 0.913 | 0.625 |
| 500   | C        | -2.71    | 0.134 | 0.898 | 0.690 | 0.903 | 0.694 |
| 500   | D        | -2.73    | 0.968 | 0.956 | 0.957 | 0.995 | 0.956 |
| 500   | E        | -2.71    | 0.899 | 0.952 | 0.927 | 0.976 | 0.951 |
| 500   | F        | -4.00    | 0.942 | 0.949 | 0.939 | 0.991 | 0.947 |
| 500   | G        | -2.73    | 0.970 | 0.958 | 0.960 | 0.993 | 0.957 |
| 500   | H        | -2.71    | 0.867 | 0.941 | 0.899 | 0.991 | 0.943 |
| 2000  | A        | -4.00    | 0.960 | 0.954 | 0.951 | 0.953 | 0.951 |
| 2000  | B        | -3.51    | 0.325 | 0.962 | 0.686 | 0.967 | 0.706 |
| 2000  | C        | -2.71    | 0.113 | 0.917 | 0.736 | 0.927 | 0.774 |
| 2000  | D        | -2.73    | 0.964 | 0.957 | 0.957 | 0.988 | 0.956 |
| 2000  | E        | -2.71    | 0.910 | 0.953 | 0.930 | 0.957 | 0.948 |
| 2000  | F        | -4.00    | 0.962 | 0.946 | 0.941 | 0.987 | 0.949 |
| 2000  | G        | -2.73    | 0.948 | 0.937 | 0.934 | 0.977 | 0.932 |
| 2000  | H        | -2.71    | 0.903 | 0.950 | 0.934 | 0.979 | 0.957 |
### TABLE 4
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)

|                      | Metformin (2004-2009) | Metformin (2010-2014) | Sulfonylurea (2004-2009) | Sulfonylurea (2010-2014) | Standardized absolute mean difference |
|----------------------|-----------------------|-----------------------|--------------------------|--------------------------|---------------------------------------|
| **Patient count**    | 84,003                | 100,612               | 29,447                   | 11,736                   |                                       |
| **Baseline age**     | 61.90(11.64)          | 60.45(11.50)          | 67.32(12.47)             | 66.95(12.72)             | 0.354                                 |
| **Male**             | 80,964(96.4)          | 95,906(95.3)          | 28,912(98.2)             | 11,472(97.8)             | 0.095                                 |
| **Race/ethnicity**   |                       |                       |                          |                          | 0.185                                 |
| Non-Hispanic White   | 11,927(14.2)          | 20,132(20.0)          | 4,038(13.7)              | 2,253(19.2)              |                                       |
| Non-Hispanic Black   | 4,730(5.6)            | 6,810(6.8)            | 1,517(5.2)               | 614(5.2)                 |                                       |
| Hispanic             | 10,897(13.0)          | 8,583(8.5)            | 5,073(17.2)              | 1,123(9.6)               |                                       |
| Other                | 56,449(67.2)          | 65,087(64.7)          | 18,819(63.9)             | 7,746(66.0)              |                                       |
| **Smoking status**   |                       |                       |                          |                          | 0.127                                 |
| Current              | 21,475(25.6)          | 30,603(30.4)          | 6,365(21.6)              | 2,925(24.9)              |                                       |
| Former               | 42,751(50.9)          | 44,354(44.1)          | 16,396(55.7)             | 5,935(50.6)              |                                       |
| Never                | 19,777(23.5)          | 25,655(25.5)          | 6,686(22.7)              | 2,876(24.5)              |                                       |
| **BMI**              | 32.96(6.43)           | 33.66(6.52)           | 31.20(6.00)              | 32.08(6.22)              | 0.220                                 |
| **SBP**              | 133.03(16.61)         | 132.35(15.93)         | 133.83(18.37)            | 131.80(17.31)            | 0.066                                 |
| **DBP**              | 76.83(10.82)          | 78.69(10.76)          | 74.90(11.59)             | 75.81(11.45)             | 0.185                                 |
| **HDL**              | 39.19(10.76)          | 40.14(10.97)          | 39.23(11.51)             | 39.60(11.50)             | 0.048                                 |
| **LDL**              | 103.43(34.93)         | 102.68(35.30)         | 101.32(35.46)            | 96.99(35.06)             | 0.098                                 |
| **Total cholesterol**| 179.60(44.15)         | 178.57(44.51)         | 177.91(46.41)            | 173.49(46.68)            | 0.069                                 |
| **Triglycerides**    | 205.36(192.15)        | 203.67(198.23)        | 206.01(198.88)           | 208.68(216.49)           | 0.013                                 |
| **Fasting plasma glucose** | 151.54(62.26)       | 149.71(63.64)         | 165.87(80.76)            | 163.67(78.23)            | 0.141                                 |
| **HbA1c**            | 7.17(1.43)            | 7.28(1.41)            | 7.42(1.65)               | 7.54(1.54)               | 0.139                                 |
| **Estimated GFR**    | 78.41(18.59)          | 83.88(20.08)          | 66.83(22.97)             | 66.15(24.76)             | 0.500                                 |
| **Creatinine**       | 1.03 (0.20)           | 0.98 (0.20)           | 1.25 (0.48)              | 1.29 (0.56)              | 0.493                                 |
| **History of CAD**   | 31,211 (37.2)         | 25,771 (25.6)         | 14,071 (47.8)            | 4,535 (38.6)             | 0.240                                 |
| **History of CHF**   | 8,768 (10.4)          | 6,086 (6.0)           | 5,916 (20.1)             | 1,752 (14.9)             | 0.237                                 |
| **History of Stroke**| 10,572 (12.6)         | 8,385 (8.3)           | 5,230 (17.8)             | 1,645 (14.0)             | 0.149                                 |
| **History of kidney disease** | 571 (0.7)            | 237 (0.2)             | 1,214 (4.1)              | 316 (2.7)                | 0.167                                 |
| **History of liver disease** | 424 (0.5)            | 349 (0.3)             | 278 (0.9)                | 86 (0.7)                 | 0.043                                 |

Note: The standardized absolute mean differences is an average measurement between the four temporal and treatment defined groups.

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 to 2009 cohort onto the 2010 to 2014 cohort, the risk difference is found to be 4.0%, 95% CI = (3.4%, 4.6%). That is, the transported effect estimate is more similar to the CBPS estimate of 2010 to 2014 than of the CBPS estimate evaluated on the 2004 to 2009 cohort. We also see that the transported estimate from the 2004 to 2009 cohort onto the 2010 to 2014 cohort is more efficient than the calibrated estimate computed with only the data from the 2010 to 2014 cohort.

We also examined the scenario which combines the 2004 to 2009 sample with the 2010 to 2014 sample to estimate the risk difference among patients in the 2010 to 2014 cohort using (14) to (16), which accommodates the data-fusion setting. When we combine the two cohorts, we get an unbiased estimate of $\tau_0$ with much greater efficiency than using only data from 2010 to 2014 alone. The efficiency gain is due to the increased effective sample size after weighting, which
FIGURE 2  Kernel density estimates of the CBPS fitted propensity and sampling scores. Recall that the propensity score measures the probability of being assigned to a sulfonylurea and the sampling score measures the probability of belonging to the 2004 to 2009 cohort vs the 2010 to 2014 cohort.

TABLE 5  Risk differences in total mortality between sulfonylurea and metformin monotherapy in a VA cohort starting from the date of first prescription (Rx) using a variety of causal effect estimators

| Method and sample | One year after Rx | Two years after Rx | Five years after Rx |
|-------------------|-------------------|--------------------|---------------------|
| Unadjusted 2004 to 2009 | 2.6% (2.3%, 2.8%) | 5.3% (4.9%, 5.7%) | 12.6% (11.9%, 13.2%) |
| Unadjusted 2010 to 2014 | 2.4% (2.0%, 2.7%) | 5.1% (4.6%, 5.7%) | 12.5% (11.7%, 13.4%) |
| CBPS 2004 to 2009 | 2.3% (2.1%, 2.6%) | 5.0% (4.7%, 5.4%) | 12.2% (11.6%, 12.7%) |
| CBPS 2010 to 2014 | 1.0% (0.7%, 1.4%) | 2.0% (1.5%, 2.5%) | 4.1% (3.4%, 4.9%) |
| Transported 2004 to 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%) |

increased modestly from 30,883 in the transportability case to 37,682 in the data-fusion case. Using the data-fusion calibration estimator, we found the estimated risk difference for 2010 to 2014 to be 4.2%, 95% CI = (3.7%, 4.7%). Regardless of the estimator, our results suggest that metformin monotherapy remains associated with lower mortality than sulfonylurea treatment for veterans with newly diagnosed diabetes. However, the size of the effect has declined over the two time periods, potentially implying improved patient selection for sulfonylurea treatment.

7 | DISCUSSION

By estimating a vector of weights that simultaneously reduces sample and treatment group heterogeneity, we show how constrained convex optimization techniques can be applied to synergize solutions for causal inference using covariate balancing weights with solutions for transporting experimental effect estimates using sampling weights. The resulting estimator eliminates both confounding and sampling bias. This allows us to transport estimates found with observational data across populations. We also examined two alternative approaches for transporting experimental results which we adapted to accommodate data from observational studies. The TMLE and augmented estimators are less constrained by
parametric assumptions, but their operational qualities are less interpretable if the desired objective is to balance covariate distributions between samples and treatment groups. We then demonstrate how the proposed full calibration approach can be extended to solve the problem of data-fusion in an observational context.

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 potential outcome models are correctly specified, the difference in performance between the three transportability estimators was negligible. For data-fusion, we saw in the simulation that the full calibration approach outperformed the alternative augmented approach. It is argued that the increased efficiency might be the result of the implied parametric nature of the full calibration approaches. However, in this simulation, the nuisance parameters used by the TMLE and augmented estimators were fit using parametric models in an effort to better contrast the results. In addition to the simulation study, we show in an illustrative example evaluating mortality rates in 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 along with eliminating confounder heterogeneity between treatment groups 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 to 7. These assumptions are necessary to guarantee the double-robustness property for this estimator. The TMLE approach does not require any assumption about the functional form of the nuisance parameters - the only requirement is that the nuisance parameter models be statistically consistent. The augmented approach as presented in Section 3.2 only requires Assumption 6, along with a statistically consistent propensity score, or consistent potential outcome models to guarantee double-robustness. In data-fusion, we show that Assumption 8 is required for the augmented approach in instances where the sampling and propensity scores are not correctly specified. We also note that the more stringent linearity assumptions culminate in a trade-off between more flexible modeling strategies and an increase to efficiency in finite sample settings, as demonstrated in the simulation study.

For much of this article, we have left the specification of the balancing weights $c_j(X_i)$ to be ill-defined other than to set $c_j(X_i) = 1$ for all $i = 1, 2, \ldots, n$. This latter property is necessary to stabilize the weights. Indeed, for many problems the choice of $c_j(X_i)$ can simply be the original covariate measurements without any transformation, as we demonstrated in the simulations and illustrative examples. A more flexible yet complex option is to use sieve-type regression methods that replace the balance functions in the dual problems of (5), (9), and (14) with an expanding basis of functions and interactions among the covariate measurements. A sieve regression alternative would relax the linearity conditions in Assumptions 5 to 7. This nonparametric approach was explored in Chan et al. for estimating balancing weights to mitigate confounding bias and in Lee et al. for generalizing treatment effects.

In the case where we need to balance an exceptionally large number of balance functions, there is little to no assurance that a feasible solution to the primal problems in (4), (8), and (13) exists. To counteract this issue, we can introduce a penalization term on the dual variables in (5), (9), and (14). Doing so induces inequality constraints onto (4), (8), and (13), respectively, as opposed to the equality constraints that we currently solve. Introducing a penalization component, like in ridge regression or LASSO, results in a bias-variance trade-off in a finite sample setting. In finite sample settings without exact balance, estimates of $\tau_0$ might be biased when using weights derived with penalization. However, a slightly biased estimate is certainly better than no estimate at all. Using such an approach was not necessary for either the simulation scenarios that we considered or for the illustrative examples, as these two applications are not high-dimensional cases. Using inequality constraints can also be used to reduce the mean squared error of an estimate for $\tau_0$ if that is an objective of the analysis. More information on using inequality constraints with calibration can be found in Wang and Zubizarreta for finding balancing weights and in Lu et al. for transporting effect estimates.

Early on, we stated that the complete individual-level covariate data were required for both samples. It would be advantageous to only require the marginal moment values of the covariate distribution (or balance functions) in the target sample 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. for 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 estimand of $\tau_0$. The findings of Josey et al. can easily be translated to the full calibration approach we propose for transporting observational results in Section 3.3 given that the two estimators of the target population average treatment effect are nearly identical. The augmented and TMLE approaches cannot be resolved in this context without additional assumptions given the need to sum the estimates of the potential outcome means over the target sample, which in general requires individual-level covariate data from the target sample. Given this finding, we believe calibration weights might be integrated or combined...
with the alternative estimators presented in this article to further enhance these solutions when they are hindered by various modeling challenges, such as the setting where individual-participant study data are available but only marginal moment values of the target sample.

ACKNOWLEDGEMENTS
Kevin P. Josey was supported in part by the National Institute of Environmental Health Sciences, NIEHS 5T32ES007142, Fan Yang was supported in part by the National Science Foundation, NSF SES-1659935, Debasis 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.

DATA AVAILABILITY STATEMENT
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 https://github.com/kevjosey/fusion-sim/. BARI 2D study data are 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.

ORCID
Kevin P. Josey https://orcid.org/0000-0003-2490-6272
Debashis Ghosh https://orcid.org/0000-0001-6618-1316
Sridharan Raghavan http://orcid.org/0000-0003-0643-4873

REFERENCES
1. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol. 2017;186(8):1010-1014.
2. Josey KP, Berkowitz SA, Ghosh D, Raghavan S. Transporting experimental results with entropy balancing. Stat Med. 2021;40(19):4310-4326.
3. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Natl Acad Sci. 2016;113(27):7345-7352.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
5. Rudolph KE, van der Laan MJ. Robust estimation of encouragement design intervention effects transported across sites. J Royal Stat Soc Ser B (Stat Methodol). 2017;79(5):1509-1525.
6. Lee D, Yang S, Dong L, Wang X, Zeng D, Cai J. Improving trial generalizability using observational studies. Biometrics. 2021.
7. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. Stat Med. 2020;39(14):1999-2014.
8. Kang J DY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007;22(4):523-539.
9. Deville J-C, Särndal C-E. Calibration estimators in survey sampling. J Am Stat Assoc. 1992;87(418):376-382.
10. Chan KC, Yam SC, Zhang Z. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. J Royal Stat Soc Ser B Stat Methodol. 2016;78(3):673.
11. Josey KP, Juarez-Colunga E, Yang F, Ghosh D. A framework for covariate balance using Bregman distances. Scand Jf Stat. 2021;48(3):790-816.
12. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials. PharmacoEconomics. 2010;28(10):935-945.
13. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89:846-866.
14. Ali MK, Bullard KMK, Saadinne JB, Cowie CC, Imperatore G, Gregg EW. Achievement of goals in U.S. diabetes care, 1999-2010. N Engl J Med. 2013;368(17):1613-1624.
15. Selvin E, Parrinello CM, Sacks DB, Coresh J. Trends in prevalence and control of diabetes in the United States, 1988-1994 and 1999-2010. Ann Internal Med. 2014;160(8):517-525.
16. Gregg EW, Li Y, Wang J, et al. Changes in diabetes-related complications in the United States, 1990-2010. N Engl J Med. 2014;370(16):1514-1523.
17. Geiss LS, Wang J, Cheng YJ, et al. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012. JAMA. 2014;312(12):1218-1226.
18. Gregg EW, Cheng YJ, Srinivasan M, et al. 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*. 2018;391(10138):2430-2440.

19. Cheng YJ, Imperatore G, Geiss LS, et al. Trends and disparities in cardiovascular mortality among U.S. adults with and without self-reported diabetes, 1988-2015. *Diabetes*. 2018;67(11):2306-2315.

20. Raghavan S, Vassy JL, Ho Y-L, et al. Diabetes mellitus-related all-cause and cardiovascular mortality in a national cohort of adults. *J Am Heart Assoc*. 2019;8(4):e011295.

21. Berkowitz SA, Krumme AA, Avorn J, et al. Initial choice of oral glucose-lowering medication for diabetes mellitus: a patient-centered comparative effectiveness study. *JAMA Internal Med*. 2014;174(12):1955-1962.

22. Hampp C, Borders-Hemphill V, Moeny DG, Wysowski DK. Use of antidiabetic drugs in the U.S., 2003-2012. *Diabetes Care*. 2014;37(5):1367-1374.

23. Desai NR, Shrank WH, Fischer MA, et al. Patterns of medication initiation in newly diagnosed diabetes mellitus: quality and cost implications. *Am J Med*. 2012;125(3):302.e1-302.e7.

24. Schramm TK, Gislason GH, Vaag A, et al. 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. *Eur Heart J*. 2011;32(15):1900-1908.

25. Roumie CL, Hung AM, Greevy RA, et al. Comparative effectiveness of sulfonylurea and metformin monotherapy on cardiovascular events in type 2 diabetes mellitus: a cohort study. *Ann Intern Med*. 2012;157(9):601-610.

26. Wheeler S, Moore K, Forsberg CW, et al. Mortality among veterans with type 2 diabetes initiating metformin, sulfonylurea or rosiglitazone monotherapy. *Diabetologia*. 2013;56(9):1934-1943.

27. Hong J, Zhang Y, Lai S, et al. Effects of metformin versus glipizide on cardiovascular outcomes in patients with type 2 diabetes and coronary artery disease. *Diabetes Care*. 2013;36(5):1304-1311.

28. Varvaki RD, Catani PL, Reck RL, Bauermann LC, Gross JL. The association between sulfonylurea use and all-cause and cardiovascular mortality: a meta-analysis with trial sequential analysis of randomized clinical trials. *PLoS Med*. 2016;13(4):e1001992.

29. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701.

30. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiology*. 2017;28(4):553-561.

31. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. *Stat Sci*. 2014;29(4):579-595.

32. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083-3107.

33. Keele L, Ben-Michael E, Feller A, Kelz R, Miratrix L. Hospital quality risk standardization via approximate balancing weights. arXiv preprint arXiv:2007.09056, 2020

34. Hirshberg DA, Maleki A, Zubizarreta JR. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*. 2020;107(1):93-105.

35. Lu B, Ben-Michael E, Miratrix L. Is it who you are or where you are? Accounting for compositional differences in cross-site treatment variation. arXiv preprint arXiv:2103.14765, 2021.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Josey KP, Yang F, Ghosh D, Raghavan S. A calibration approach to transportability and data-fusion with observational data. *Statistics in Medicine*. 2022;41(23):4511-4531. doi: 10.1002/sim.9523