

RESEARCH ARTICLE

Doubly robust augmented weighting estimators for the analysis of externally controlled single-arm trials and unanchored indirect treatment comparisons

Harlan Campbell^{1,2} and Antonio Remiro-Azócar³ 

¹Department of Statistics, The University of British Columbia, Canada

²Precision AQ, Vancouver, Canada

³Methods and Outreach, Novo Nordisk, Spain

Corresponding author: Harlan Campbell; Email: harlancampbell@gmail.com

Received: 16 October 2025 UTC; **Revised:** 11 February 2026 UTC; **Accepted:** 14 April 2026 UTC

Keywords: covariate adjustment; data fusion; evidence synthesis; external control; indirect treatment comparison; single-arm trial

Abstract

Externally controlled single-arm trials are critical to assess treatment efficacy across therapeutic indications for which randomized controlled trials are not feasible. A closely-related research design, the unanchored indirect treatment comparison, is often required for disconnected treatment networks in health technology assessment. We present a unified causal inference framework for both research designs. We develop an estimator that augments a popular weighting approach based on entropy balancing—matching-adjusted indirect comparison (MAIC)—by fitting a model for the conditional outcome expectation. The predictions of the outcome model are combined with the entropy balancing MAIC weights. While the standard MAIC estimator is singly robust where the outcome model is non-linear, our augmented MAIC approach is doubly robust (DR), providing increased robustness against model misspecification. This is demonstrated in a simulation study with binary outcomes and a logistic outcome model, where the augmented estimator demonstrates its DR property, while exhibiting higher precision than all non-augmented weighting estimators and near-identical precision to G-computation. We describe the extension of our estimator to the setting with unavailable individual participant data for the external control, illustrating it through an applied example. Our findings reinforce the understanding that entropy balancing-based approaches have desirable properties compared to standard “modeling” approaches to weighting, but should be augmented to improve protection against bias and guarantee double robustness.

Highlights

What is already known?

- Externally controlled single-arm trials (SATs) are important for assessing treatment efficacy when randomized controlled trials are not feasible.
- Unanchored indirect treatment comparisons (ITCs), often using the entropy balancing or matching-adjusted indirect comparison (MAIC) approach, can provide treatment effect estimates for disconnected networks in health technology assessment.

  This article was awarded Open Data and Open Materials badges for transparent practices. See the Data availability statement for details.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

What is new?

- A unified causal inference framework for both externally controlled SATs and unanchored ITCs is presented.
- A doubly robust (DR) estimator that augments the MAIC approach by combining outcome model predictions with entropy balancing weights is proposed.
- The proposed estimator provides increased protection against model misspecification, as demonstrated in a simulation study with binary outcomes.
- The proposed estimator is extended to the setting with unavailable individual participant data for the external control and illustrated through an applied example.

Potential impact for RSM readers

- DR augmented weighting estimators, particularly using entropy balancing weights, offer increased robustness to model misspecification than conventional non-augmented approaches for confounding adjustment.
- Entropy balancing-based methods have desirable properties compared to standard “modeling” approaches to weighting but should be augmented to improve protection against bias and guarantee double robustness.

1. Background

In pharmaceutical research, randomized controlled trials (RCTs) are the gold standard for evaluating treatment efficacy and safety due to their high internal validity. Random allocation minimizes confounding by balancing, in expectation, prognostic factors across treatment arms. Nevertheless, RCTs are not always feasible, for instance:

- where recruitment is impractical due to small populations, for example, rare diseases or biomarker-specific precision oncology¹;
- for life-threatening conditions with high unmet need and inadequate standard of care, for example, “last-line of therapy” indications in late-stage hematological and solid tumor oncology²
- where placebo controls are unethical, for example, withholding a therapy with proven efficacy in adults from a control group of children in pediatric trials.³

Regulatory agencies such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) emphasize that RCTs provide the highest evidentiary standard.^{4,5} However, regulators recognize that alternative designs may be required in special circumstances. One such design is the externally controlled single-arm trial (SAT), where the control group is fully derived from external data, such as prior clinical trials or secondary real-world data (RWD) sources.^{2,6–8} Marketing authorization applications featuring externally controlled SATs continue to rise, especially under accelerated approval pathways.^{8,9} In 2023, the FDA issued draft guidance for externally controlled trials,¹⁰ and in 2024, the EMA finalized a reflection paper on SATs.¹¹ Approvals have been granted based on such designs, particularly for conditions with predictable natural history, precisely measurable endpoints, and anticipated large effect sizes.^{1,2,8,9,12,13}

For health technology assessment (HTA), head-to-head RCT evidence also remains the gold standard.^{14–16} Nevertheless, as regulators increasingly approve pharmaceuticals based on externally controlled SATs, payer reliance on such designs has grown.^{9,17} HTA bodies are developing recommendations for externally controlled SATs,^{18,19} with acceptability influenced by unmet need and disease rarity.^{9,17,20} HTA also requires comparing new technologies against all existing alternatives.²¹ The scope of assessments often depends on the policy question rather than available data²² and a single RCT cannot typically include all comparators desired for HTA, given the multiplicity of stakeholders and variations in clinical practice across jurisdictions.²³ In the absence of direct RCT comparisons versus all candidate comparators, indirect treatment comparisons (ITCs) are required.²⁴

HTA decision-makers prefer *anchored* ITCs of randomized trials,^{25–27} which respect randomization by using a common control arm to contrast relative treatment effects.²⁸ However, compatible control arms with which to “anchor” the analysis are not always available, especially in rapidly evolving

areas with multiple novel treatments and no single accepted standard of care.^{29,30} In these scenarios, *unanchored* ITCs based on disconnected networks may be required,^{29,30} and recent reviews determined that unanchored ITCs are in fact more common than anchored ITCs.^{31,32} Unanchored ITCs contrast mean treatment-specific absolute outcomes across studies, relying on more restrictive assumptions than anchored ITCs.^{25,28} In essence, they are externally controlled SATs where the external control is often a competitor's historical trial with limited data access: individual participant data (IPD) are available for the SAT, but only published aggregate-level data (AD) for the external control.^{25,28}

The absence of randomization compromises the validity of externally controlled SATs. Various statistical methods have been proposed to adjust for imbalances in baseline covariates³³ and these can potentially mitigate confounding bias and account for the additional variability induced by covariate differences. The most widely-used methods are propensity score-based weighting approaches, typically using logistic regression estimated via maximum-likelihood.^{34–37} For unanchored ITCs, matching-adjusted indirect comparison (MAIC) based on entropy balancing³⁸ is more popular.^{39,40,41} MAIC views covariate balance as a convex optimization problem, estimating weights that directly enforce balance without explicitly modeling the conditional probability of SAT participation. MAIC is attractive for ITCs due to its applicability in IPD-AD situations and is thought to be more stable, precise, and robust to model misspecification than the standard propensity score-based weighting approaches, even in “IPD-IPD” scenarios.^{41–44}

So-called “G-computation” or “model-based standardization” methods have also been developed for the IPD-IPD^{35,45–47} and IPD-AD settings.⁴⁸ For the latter, they are also referred to as “simulated treatment comparison” (STC).⁴⁹ These methods estimate a model for the conditional outcome expectation and average predictions over the target covariate distribution. G-computation exhibits increased precision relative to weighting, particularly when overlap is poor,⁴⁸ but relies on model-based extrapolation and can be prone to bias under model misspecification.⁵⁰

Weighting and G-computation are generally “singly robust”: weighting, in most cases, depends on correct propensity score model specification; G-computation on correct outcome model specification. Decision-makers have expressed a preference for “doubly robust” (DR) estimation approaches that can consistently estimate the treatment effect as long as either the propensity score model or the outcome model is correct, but not necessarily both.^{25,27,28,51} These methods should reduce the risk of bias by offering two opportunities for correct model specification. Despite this, DR methods have rarely been applied to externally controlled SATs⁵² and, to our knowledge, never to unanchored ITCs, despite recommendations from HTA agencies.^{25,28}

One barrier may be a misunderstanding that MAIC is always DR. MAIC enables consistent estimation when an implicit propensity score model is misspecified, but only if the true outcome model is linear with respect to balanced covariate functions—termed “linearly DR.”^{41,43,44} In practice, outcomes rarely vary linearly with covariates. DR methods for ITCs that are not necessarily restricted to linear outcome models are yet to be developed, with Josey et al. recently identifying this as a research priority.⁴¹

This article clarifies existing approaches for DR estimation in externally controlled SATs and proposes a DR augmented MAIC estimator for unanchored ITCs. Section 2 introduces the target estimand. Section 3 outlines available estimators and our proposed approach. Section 4 presents simulation study results comparing our proposed DR augmented estimators against existing singly robust and other augmented estimators. Lastly, we illustrate the application of the methods in an example analysis in Section 5, and conclude in Section 6.

2. Estimands

We begin by defining the *estimands* that can be targeted by externally controlled SATs. An estimand is a precise definition of the treatment effect, which should align with the clinical question of interest, the research design, and the analytical approach. The International Council of Harmonisation E9 (R1) Addendum, adopted by the FDA and EMA, specifies five estimand attributes: population,

treatment(s), endpoint, summary effect measure, and strategies for intercurrent events.⁵³ We focus on the “population” and “summary effect measure” when defining the estimands, which are:

- the average treatment effect (ATE) among the combined SAT and external control;
- the average treatment effect in the treated (ATT), among those participating in the SAT;
- the average treatment effect in the control (ATC), among the external control group.

The difference between these summary effect measures is driven by them targeting different (sub) populations or applying to different “analysis sets.” Having assumed that the SAT and external control are random samples of their underlying target populations, we make no further distinction between sample-level and population-level estimands.

Using potential outcomes notation, let Y^t represent the outcome under intervention $T = t$, with $t \in \{0, 1\}$, where $T = 1$ denotes the SAT intervention (data source $S = 1$) and $T = 0$ the external control (data source $S = 0$). Two potential outcomes, (Y^1, Y^0) , exist for every subject; one is observed, the other counterfactual. The ATE is

$$\text{ATE} = g\left(E\left(Y^1\right)\right) - g\left(E\left(Y^0\right)\right),$$

where the link function $g(\cdot)$ transforms potential outcome means into the plus/minus infinity range, and expectations are over the distribution of potential outcomes in the combined SAT and external control population. For binary outcomes, suitable links include identity, log, or logit to produce a risk difference, log relative risk, or log-odds ratio, respectively, as the summary effect measure. The ATT is

$$\text{ATT} = g\left(E\left(Y^1 \mid S = 1\right)\right) - g\left(E\left(Y^0 \mid S = 1\right)\right),$$

with expectations taken over the SAT (sub) population. The ATC is

$$\text{ATC} = g\left(E\left(Y^1 \mid S = 0\right)\right) - g\left(E\left(Y^0 \mid S = 0\right)\right),$$

with expectations over the external control (sub) population.

Within RCTs, the ATE, ATT, and ATC are identical in expectation. However, they generally differ in externally controlled SATs and will almost invariably do so where there is treatment effect heterogeneity by the covariates, that is, effect measure modification. We view the ATE target population, defined by pooling the SAT and the external control, as somewhat ambiguous in this context. As such, the target estimand in an externally controlled SAT is often either the ATT or the ATC.

The ATT is typically the primary estimand for regulatory drug approval, consistent with emulating a randomized comparison in the pivotal trial population, with the external control mimicking the internal comparator arm of a registrational clinical trial. The ATT is also compatible with the mean absolute outcome that is targeted by the SAT, $E(Y^1 \mid S = 1)$, preserving the original SAT results. Nevertheless, SAT populations are often highly selected and may lack “real-world” representativeness, making the ATT potentially less appealing for HTA, where generalizability to routine clinical practice is a priority.

The ATC can be more desirable for external validity, as natural history and RWD-based external controls have relatively broad inclusion criteria and heterogeneous target populations. However, external controls based on historical trials will not reflect the current standard of care and RWD-derived controls are often country-specific, not necessarily transferable to the relevant jurisdiction for decision-making. Sample size considerations also influence the estimand choice. Both SATs and external controls often have low sample sizes and covariate adjustment may reduce effective sample sizes (ESSs) further. Consider weighting, where the estimand impacts the definition of the weights. Targeting the ATT implies preserving the original SAT, re-weighting, and reducing the ESS of the external control. Conversely, targeting the ATC implies the reverse.

Finally, where IPD are available for the SAT but only AD for the external control, as in unanchored ITCs, the ATC is often targeted by necessity.^{25,28,33} Throughout this manuscript, we assume unlimited subject-level data access but target the ATC. Our methodological approaches are also applicable where

subject-level data are unavailable for the external control, and to target the ATT instead of the ATC, with caveats discussed in Section 3.8 and the [Supplementary Material](#), respectively.

3. Methodology

3.1. Data and assumptions

As per Section 2, let $T = t$ denote a time-fixed binary treatment, with $t \in \{0, 1\}$, such that $T = 1$ represents the active intervention and $T = 0$ the control. Let $S = s$ denote the data source, with $s \in \{0, 1\}$, such that $S = 1$ represents the SAT and $S = 0$ the external data source. In addition, let $\mathbf{X}$ denote vector-valued pre-treatment baseline covariates, for example, clinical or demographic characteristics, measured across the SAT and the external data source. Let Y denote the clinical outcome of interest. We assume that only distributional differences in $\mathbf{X}$ are preventing exchangeability between the SAT and external subjects, and that covariates and outcomes are defined and measured similarly across data sources.

The observed IPD consist of $(S_i, \mathbf{X}_i, T_i, Y_i)$, $i = 1, \dots, n_1 + n_0$, realizations of $(S, \mathbf{X}, T, Y)$ denoting the data source, baseline covariates, treatment assignment, and observed outcome for subject i . Here, the SAT and external data source have been stacked, with n_1 and n_0 as the sample sizes of the SAT and the external data source, respectively. It is assumed that all individuals in the SAT are under $T = 1$ and all individuals in the external data source are under $T = 0$, such that the control group is fully external. To be clear, we have $S_i = T_i$ for all $i = 1, \dots, n$, where $n = n_1 + n_0$; and we also have that $S_i = 1$ and $T_i = 1$ for all $i = 1, \dots, n_1$, and $S_i = 0$ and $T_i = 0$ for all $i = n_1 + 1, \dots, n$. We shall assume that there is no missingness or measurement error.

The observed outcome for subject i is $Y_i = Y_i^1 T_i + Y_i^0 (1 - T_i)$, where Y_i^t is the potential outcome had subject i been assigned treatment $t \in \{0, 1\}$, with $Y_i = Y_i^1$ if $i = 1, \dots, n_1$ and $Y_i = Y_i^0$ if $i = n_1 + 1, \dots, n$. Namely, the observed outcome for an individual in the SAT equals their potential outcome under the active intervention, and the observed outcome for an individual in the external data source equals their potential outcome under the control. Implicit in the notation is the stable unit treatment value assumption (SUTVA): that there is no interference between subjects and there is treatment version irrelevance, that is, one well-defined version of the active intervention and the control across all subjects and data sources.⁵⁴ Also implicit is that there is no direct effect of trial participation.⁵⁴ Namely, that trial participation—in the SAT or a historical trial, for that matter—does not affect the outcome except through treatment assignment itself, that is, there are no Hawthorne effects.^{55,56}

To estimate the ATC, we must construct estimators for $\mu_0^1 = E(Y^1 | S = 0)$ and $\mu_0^0 = E(Y^0 | S = 0)$. Outcomes for the subjects from the external data source have been generated under the control and we assume that there is no informative missingness or measurement error. Hence, unbiased estimation of μ_0^0 should be trivial using the sample mean, such that $\hat{\mu}_0^0 = \frac{1}{n_0} \sum_{i=n_1+1}^n Y_i$. Conversely, while the active intervention has been investigated in the SAT, its outcomes in the external control (sub) population are unobserved. Our challenge is therefore to produce a reliable estimate $\hat{\mu}_0^1$ of the mean absolute outcome μ_0^1 under the active intervention in the external control (sub) population, based on the observed data.

Two causal identification conditions, together known as *strong ignorability*, are required to construct a valid estimator of μ_0^1 . These ensure that the SAT and external control outcomes are comparable given adjustment for baseline covariates. The first assumption is *conditional data source ignorability*; formally, $Y_i^1 \perp S_i | \mathbf{X}_i$ for all $i = 1, \dots, n$. Namely, conditional on baseline covariates, the potential outcome under the active intervention is independent of the data source. This is akin to the conditional constancy or exchangeability of absolute outcomes invoked for unanchored ITCs, used to transport mean absolute outcomes under $T = 1$ from $S = 1$ to $S = 0$.^{25,28} Conditional ignorability is a strong assumption, resting on the SAT and the external control capturing all variables that are prognostic of outcome under the active intervention.

The second assumption is *positivity* or *overlap*. That is, the support of the baseline covariates in the external control is contained within that of the SAT. Mathematically, the probability of SAT

participation, conditional on the covariates necessary to ensure ignorability, should be bounded away from zero and one: $0 < \Pr(S = 1 \mid \mathbf{X} = \mathbf{x}) < 1$ for all $\mathbf{x}$ with positive density in the external control, that is, for all $\mathbf{x}$ such that $f(\mathbf{x} \mid S = 0) > 0$. Hence, it is possible to have SAT subjects in all regions of the covariate space in $S = 0$.^{54,57,58} Positivity violations can be deterministic or random. The former arise structurally, due to non-overlapping SAT and external control eligibility criteria. The latter arise empirically due to chance, particularly with small sample sizes.⁵⁹ To enforce positivity, analysts may subset the SAT based on the selection criteria of the external control.^{25,28} However, this further reduces the sample size of the SAT. Positivity is typically assessed by comparing the empirical distributions of the covariates in the SAT and the external control.⁶⁰ While outcome modeling-based approaches such as G-computation can overcome failures of positivity, they do so by potentially problematic and difficult-to-diagnose model-based extrapolation. Even minor model misspecification over the observed covariate space in the SAT may lead to poor extrapolation in unobserved regions of the covariate space.⁵⁰

Analogously, targeting the ATT would require constructing estimators for $\mu_1^1 = E(Y^1 \mid S = 1)$ and $\mu_1^0 = E(Y^0 \mid S = 1)$. Here, the challenge is the estimation of μ_1^0 because outcomes under the control have not been generated in the SAT. The conditional ignorability assumption would formally be $Y_i^0 \perp S_i \mid \mathbf{X}_i$ for all $i = 1, \dots, n$, and would rest on the SAT and the external control measuring all variables that are prognostic of outcome under the control. The positivity assumption would be $0 < \Pr(S = 0 \mid \mathbf{X} = \mathbf{x}) < 1$ for all $\mathbf{x}$ with positive density in the SAT, $f(\mathbf{x} \mid S = 1) > 0$, such that the support of the baseline covariates in the SAT is contained within that of the external control and it is possible to have external control subjects in all regions of the SAT covariate distribution. In this setting, analysts may apply the SAT selection criteria to the external control to guarantee that there is sufficient overlap. While the methods in the next sections target the ATC, we present any required modifications to target the ATT in the [Supplementary Material](#).

3.2. Inverse odds weighting

We first present a covariate adjustment method that models the data source assignment mechanism, conditional on baseline covariates, to estimate weights.^{34,35,41,61,62} Where the target estimand is the ATC, SAT subjects are weighted by their inverse conditional odds of SAT participation—their conditional odds of external control participation—to transport the SAT outcomes to the external control (sub) population. Such “inverse odds” weights (IOW) are defined as

$$w_i = \frac{(1 - e_i)S_i}{e_i} + (1 - S_i), \quad (1)$$

for subject $i = 1, \dots, n$, where the propensity score $e_i = e(\mathbf{X}_i) = \Pr(S_i = 1 \mid \mathbf{X}_i)$ denotes the conditional probability of SAT participation given covariates $\mathbf{X}_i$ for subject i . In Equation (1), note that the SAT subjects ($S_i = 1$) are weighted as $w_i = (1 - e_i)/e_i$, whereas the external control subjects ($S_i = 0$) are unweighted, that is, assigned a weight of $w_i = 1$.

In practice, the true propensity scores are unknown. Almost invariably, there are multiple baseline covariates and at least one of these is continuous, such that a data source assignment model is required to estimate the propensity scores. The model is often a logistic regression:

$$\text{logit}(e_i) = \alpha_0 + \mathbf{c}(\mathbf{X}_i)^\top \boldsymbol{\alpha}, \quad (2)$$

where $\text{logit}(e_i) = \ln((e_i)/(1 - e_i))$, $\alpha_0 = \ln(\Pr(S_i = 1 \mid \mathbf{c}(\mathbf{X}_i) = 0)/(\Pr(S_i = 0 \mid \mathbf{c}(\mathbf{X}_i) = 0)))$ is an intercept term, $\boldsymbol{\alpha}$ is a vector of regression parameters, and $\mathbf{c}(\mathbf{X}_i) = [c_1(\mathbf{X}_i), c_2(\mathbf{X}_i), \dots, c_k(\mathbf{X}_i)]^\top$ is a vector of covariate “balance functions” for subject $i = 1, \dots, n$. This is the set of functions containing the distributional features to be balanced between the SAT and the external control,^{41,63} potentially including sensible transformations of the covariates, for example, polynomials and interaction terms.

The logistic regression is typically fitted to the concatenated IPD using maximum-likelihood estimation, with the regression coefficient point estimates denoted by $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$, and model-based propensity

scores for subject $i = 1, \dots, n_1$, predicted by $\hat{e}_i = \text{logit}^{-1}(\hat{\alpha}_0 + \mathbf{c}(\mathbf{X}_i)^\top \hat{\boldsymbol{\alpha}}) = \text{expit}(\hat{\alpha}_0 + \mathbf{c}(\mathbf{X}_i)^\top \hat{\boldsymbol{\alpha}})$, where $\text{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. Weight estimates $\hat{w}_i$ for $i = 1, \dots, n_1$ are derived by plugging the corresponding propensity score predictions into Equation (1). With correct specification of the model in Equation (2), such that the log-odds of SAT participation are linear across the balance functions of the covariates, $\hat{e}_i$ and $\hat{w}_i$ consistently estimate the true conditional probability and inverse odds of SAT participation, respectively.

The ATC is estimated by contrasting the weighted average of observed outcomes under the active intervention with the unweighted average of observed outcomes for the external control. As per Section 2, mean absolute outcomes are converted to the additive scale imposed by link function $g(\cdot)$ prior to taking the difference between treatments on such scale, leading to the *IOW* estimator for the ATC:

$$\widehat{\text{ATC}} = g\left(\underbrace{\frac{1}{n_0} \sum_{i=1}^{n_1} \hat{w}_i Y_i}_{\hat{\mu}_0^1}\right) - g\left(\underbrace{\frac{1}{n_0} \sum_{i=n_1+1}^n Y_i}_{\hat{\mu}_0^0}\right), \quad (3)$$

The mean absolute outcome estimate for the active intervention can be bounded within its feasible range, for example, between 0 and 1 for probabilities, by normalizing or “stabilizing” the weights so that they sum to one.^{57,61} This results in the alternative *normalized IOW* estimator for the ATC:

$$\widehat{\text{ATC}} = g\left(\underbrace{\frac{\sum_{i=1}^{n_1} \hat{w}_i Y_i}{\sum_{i=1}^{n_1} \hat{w}_i}}_{\hat{\mu}_0^1}\right) - g\left(\underbrace{\frac{1}{n_0} \sum_{i=n_1+1}^n Y_i}_{\hat{\mu}_0^0}\right), \quad (4)$$

which should provide improved finite sample properties and more stable and precise estimation.^{64,65} Drawing an analogy with survey sampling, the estimator in Equation (3) is a Horvitz–Thompson-type estimator and that in Equation (4) is a Hajek-type estimator, with the latter typically considered to improve the performance of the former in the literature.⁶⁶

In expectation, if the model in Equation (2) is correctly specified, the estimated weights ($\hat{w}_i$, for i in $1, \dots, n_1$) will balance the covariate distribution of the SAT with respect to that of the external control, enabling consistent estimation of mean absolute outcome μ_0^1 and the ATC. A mathematical derivation showing the consistency of the *IOW* estimators is provided in the Appendix.

Unfortunately, this “modeling approach” to weighting,^{67,68} where propensity scores are explicitly modeled as a function of baseline covariates by a logistic regression, then estimated by maximizing the fit of such regression, has certain limitations:^{67,68}

- The resulting weights do not produce adequate covariate balance if the propensity score model is misspecified, and even a correctly specified model does not guarantee balance in finite samples.
- Propensity score predictions that are close to zero produce extreme and highly variable weights, which lead to unstable and imprecise ATC estimation, particularly where overlap is poor or the sample size of the SAT is small.
- There is limited applicability when covariate IPD for the external control are unavailable and only marginal summary moments from published tables of baseline characteristics are available.

3.3. Entropy balancing (matching-adjusted indirect comparison)

The limitations of the weighting methods in Section 3.2 motivate alternative “balancing” or “calibration” approaches to weighting. These estimate weights under the condition that covariates are balanced, viewing balance as an optimization problem, without explicitly modeling the propensity score. Generally, balancing approaches to weighting are: (1) less susceptible to bias by directly

enforcing covariate balance; (2) produce more stable weights, which translate into larger ESSs and more precise treatment effect estimation; and (3) are applicable where only aggregate-level marginal covariate moments are available for the external control.^{67–70}

Our focus here is on an entropy balancing approach³⁸ called MAIC,^{39–41} but see Chattopadhyay et al. and Filla et al. for details about similar balancing techniques.^{67,68} MAIC is the most popular balancing method in the context of externally controlled SATs. It has many features that are considered desirable: “linear double robustness,” minimally dispersed weights, and the estimation of odds weights that are guaranteed to be positive, resulting in increased interpretability and sample boundedness, that is, interpolating the observed data as opposed to extrapolating beyond its support.^{68,71} We review the main steps of MAIC, building on prior literature.^{39,70,72–75}

While MAIC does not explicitly model the propensity score as a function of baseline covariates, it implicitly assumes the following logistic model for data source assignment:

$$\ln(v_i) \propto \ln\left(\frac{(1 - e_i)}{e_i}\right) = \gamma_0 + \mathbf{c}(\mathbf{X}_i)^\top \boldsymbol{\gamma}, \quad (5)$$

where v_i is a weight proportional to the inverse conditional odds of SAT participation for subject $i = 1, \dots, n$, γ_0 is an intercept term parameter, and $\boldsymbol{\gamma}$ is a vector of model parameters. In Equation (5), it is the log-odds of external control participation, $\text{logit}(1 - e_i)$, that are linear across the covariate balance functions. Because $\text{logit}(1 - e_i) = -\text{logit}(e_i)$, this implies that the log-odds of SAT participation, $\text{logit}(e_i)$, also vary linearly with $\mathbf{c}(\mathbf{X}_i)$ as per Equation (2).

Signorovitch et al.³⁹ propose using the “method of moments” to estimate the model in Equation (5), such that:

$$\frac{\sum_{i=1}^{n_1} v_i \mathbf{c}(\mathbf{X}_i)}{\sum_{i=1}^{n_1} v_i} = \frac{1}{n_0} \sum_{i=n_1+1}^n \mathbf{c}(\mathbf{X}_i) \quad (6)$$

$$\frac{\exp(-\gamma_0) \sum_{i=1}^{n_1} \exp(\mathbf{c}(\mathbf{X}_i)^\top \boldsymbol{\gamma}) \mathbf{c}(\mathbf{X}_i)}{\exp(-\gamma_0) \sum_{i=1}^{n_1} \exp(\mathbf{c}(\mathbf{X}_i)^\top \boldsymbol{\gamma})} = \frac{1}{n_0} \underbrace{\sum_{i=n_1+1}^n \mathbf{c}(\mathbf{X}_i)}_{\hat{\boldsymbol{\theta}}} \quad (7)$$

$$\frac{\sum_{i=1}^{n_1} \exp(\mathbf{c}(\mathbf{X}_i)^\top \boldsymbol{\gamma}) \mathbf{c}(\mathbf{X}_i)}{\sum_{i=1}^{n_1} \exp(\mathbf{c}(\mathbf{X}_i)^\top \boldsymbol{\gamma})} = \hat{\boldsymbol{\theta}}, \quad (8)$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \dots, \hat{\boldsymbol{\theta}}_k)^\top$ is a vector of covariate balance function moments $j = 1, \dots, k$ for the external control sample, with $\hat{\boldsymbol{\theta}}_j = \frac{1}{n_0} \sum_{i=n_1+1}^n c_j(\mathbf{X}_i)$ assumed to be a consistent estimator for $\boldsymbol{\theta}_j$. Equation (6) is a constraint enforcing that the covariate distributional features of the weighted SAT subjects are exactly balanced with respect to those of the unweighted external control subjects. Equation (7) follows from introducing the assumed model in Equation (5) into the balancing constraint, and Equation (8) results from the exponentiated intercept terms canceling out.

Replacing $\boldsymbol{\gamma}$ with estimate $\hat{\boldsymbol{\gamma}}$ in Equation (8) and centering the SAT covariate balance functions on their external control means, one obtains

$$\frac{\sum_{i=1}^{n_1} \exp(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}) \mathbf{c}^*(\mathbf{X}_i)}{\sum_{i=1}^{n_1} \exp(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}})} = \mathbf{0}, \quad (9)$$

where $\mathbf{0}$ is a vector of zeros and $\mathbf{c}^*(\mathbf{X}_i) = \mathbf{c}(\mathbf{X}_i) - \hat{\boldsymbol{\theta}}$ is a vector of centered covariate balance functions for subject $i = 1, \dots, n_1$ in the SAT. Then, because the denominator is positive, Equation (9) is equal to $\sum_{i=1}^{n_1} \exp(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}) \mathbf{c}^*(\mathbf{X}_i) = \mathbf{0}$. Solving for $\hat{\boldsymbol{\gamma}}$ is equivalent to minimizing the objective function:

$$Q(\hat{\gamma}) = \sum_{i=1}^{n_1} \exp(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\gamma}), \quad (10)$$

as the derivative of $Q(\hat{\gamma})$ with respect to $\hat{\gamma}$ is $\sum_{i=1}^{n_1} \exp(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\gamma}) \mathbf{c}^*(\mathbf{X}_i)$. The objective function in Equation (10) is strictly convex and can be minimized using standard Newton-type convex optimization algorithms,³⁸ yielding a unique finite solution corresponding to the global minimum of $Q(\hat{\gamma})$. We have $v_i \propto \exp(\mathbf{c}(\mathbf{X}_i)^\top \gamma) \propto \exp((\mathbf{c}(\mathbf{X}_i) - \boldsymbol{\theta})^\top \gamma) = \exp(\mathbf{c}^*(\mathbf{X}_i)^\top \gamma)$. Subject to the normalization constraint $\sum_{i=1}^{n_1} \hat{v}_i = 1$, weights for each individual $i = 1, \dots, n_1$ in the SAT are estimated as

$$\hat{v}_i = \frac{\exp(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\gamma})}{\sum_{i=1}^{n_1} \exp(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\gamma})}. \quad (11)$$

Note that this definition of the entropy balancing weights coincides with the definitions presented by Jiang et al.⁷⁵ and by Jackson et al.⁷⁴

Similar to Equation (4), the MAIC estimator for the ATC contrasts absolute outcomes on the additive scale:

$$\widehat{\text{ATC}} = \underbrace{g\left(\sum_{i=1}^{n_1} \hat{v}_i Y_i\right)}_{\hat{\mu}_0^1} - \underbrace{g\left(\frac{1}{n_0} \sum_{i=n_1+1}^n Y_i\right)}_{\hat{\mu}_0^0}, \quad (12)$$

where the weights have already been normalized to sum to one. Alternatively, fitting a weighted univariable regression of outcome on treatment to the concatenated IPD (with weights $\hat{v}_i$ for $i = 1, \dots, n_1$ and unit weights for $i = n_1 + 1, \dots, n$) has been proposed, with the treatment coefficient of the fitted model yielding an ATC estimate.^{49,76–79} We discourage this approach. The weights estimated by Equation (11) are relative; their arbitrary rescaling by a constant of proportionality, for example, a normalization constant, will also balance the specified covariate functions and does not affect $\hat{\mu}_0^1$ and $\widehat{\text{ATC}}$ in Equation (12).^{72,74} Nevertheless, rescaling the weights while retaining weights of one for the external control subjects results in different fitted models.⁷⁴

The method-of-moments MAIC estimator gives the Lagrangian dual solution to an entropy balancing primal problem: minimizing the negative entropy of the weights.^{41,43,72} Namely, minimizing the objective function in Equation (10) is equivalent to minimizing the negative entropy $\sum_{i=1}^{n_1} v_i \ln(v_i)$ with Lagrange multipliers.^{43,72} The dual optimization problem is easier to solve than the primal³⁸ and has been formulated in different ways.^{41,72,75} These may perform differently computationally—for example, if minimization is performed on the log scale⁷²—but result in the same unique dual solution, up to numerical error, due to strict convexity. Because the negative entropy measures the distance of the weights from a uniform distribution, its minimization should produce less disperse weights than the modeling approach in Section 3.2.

MAIC (entropy balancing) has a “linear double robustness” property as it is consistent under two distinct underlying data-generating models^{41,43,44}: (1) the true log-odds of the propensity score are linear across the specified balance functions or (2) the true potential outcome under the active intervention is linear across the specified balance functions. Respectively, either $\text{logit}(e_i) = \alpha_0 + \mathbf{c}(\mathbf{X}_i)^\top \alpha$ for some parameters α_0 and α ; or, $E(Y_i^1 | \mathbf{X}_i) = \beta_0 + \mathbf{c}(\mathbf{X}_i)^\top \beta$ for some parameters β_0 and β .

For continuous covariates, it is common practice to only balance first-order sample moments (means) by setting $\mathbf{c}(\mathbf{X}_i) = c_1(\mathbf{X}_i) = \mathbf{X}_i$ and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_1 = \frac{1}{n_0} \sum_{i=n_1+1}^n \mathbf{X}_i$, in which case consistency is guaranteed as long as either the true propensity score model is $\text{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \alpha$; or, the potential outcome for the active intervention varies linearly with the covariates $\mathbf{X}_i$ that are mean-balanced: $E(Y_i^1 | \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \beta$. An alternative strategy is to also balance second-order sample moments (variances) by setting $\mathbf{c}(\mathbf{X}_i) = [c_1(\mathbf{X}_i), c_2(\mathbf{X}_i)]^\top = [\mathbf{X}_i, \mathbf{X}_i^2]^\top$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)^\top$, $\hat{\boldsymbol{\theta}}_1 = \frac{1}{n_0} \sum_{i=n_1+1}^n \mathbf{X}_i$ and $\hat{\boldsymbol{\theta}}_2 = \frac{1}{n_0} \sum_{i=n_1+1}^n \mathbf{X}_i^2$, in which case consistency is achieved if either the true propensity score model is $\text{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \alpha_1 + (\mathbf{X}_i^2)^\top \alpha_2$;

or, the potential outcome for the active intervention varies quadratically with $\mathbf{X}_i$, such that $E(Y_i^1 | \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \beta_1 + (\mathbf{X}_i^2)^\top \beta_2$. Balancing the means of the covariates and the squared covariates enforces that variances are balanced because $\text{Var}(X) = E(X^2) - E(X)^2$.

One could conjecture more flexible implicit data-generating models by considering other non-linear transformations of $\mathbf{X}_i$, for example, higher-order polynomial terms and flexible basis functions such as splines, and balancing on the means of the transformed covariates. Moreover, one could go beyond the marginal moments of individual covariates and also balance joint covariate distributions; for instance, by including quantiles of interactions for pairs or triples of covariates.⁸⁰ However, pursuing these more ambitious balancing strategies is often infeasible:

- As the number of balancing conditions increases, it is more likely that $\boldsymbol{\theta}$ falls outside the convex hull of $\mathbf{c}(\mathbf{X}_i)$ for $i = 1, \dots, n_1$.⁶⁰ This implies that a feasible weighting solution to the convex optimization problem does not exist: there is no set of positive weights that can enforce balance in the required distributional features and the numerical optimization algorithm will fail to converge.⁶⁷
- Increasing the number of balancing conditions leads to further reductions in ESS and precision, which are particularly problematic with low sample sizes and poor covariate overlap.
- Where covariate IPD for the external control are unavailable, only first- and second-order marginal moments from published tables of baseline characteristics are often available for balancing. Higher-order moments and the means of transformed covariates are rarely reported.

3.4. G-computation

An alternative to weighting involves postulating a model $m(\mathbf{X}_i; \boldsymbol{\beta})$ for the potential outcome expectation under the active intervention, conditional on covariates $\mathbf{X}_i$:

$$q(E(Y_i^1 | \mathbf{X}_i; \boldsymbol{\beta})) = m(\mathbf{X}_i; \boldsymbol{\beta}), \quad (13)$$

where $q(\cdot)$ denotes an appropriate link function and $\boldsymbol{\beta}$ is a vector of model parameters encoding the covariate–outcome relationships. For instance, a logistic regression could be used for binary outcomes, such that the model is binomial, the link function is the logit, and the potential outcome expectation is on the probability scale. We have assumed that the model for the conditional outcome expectation is parametric, but this need not necessarily be the case.⁸¹

The model in Equation (13) is fitted to the SAT participants $i = 1, \dots, n_1$ using maximum-likelihood estimation, such that the fitted model $m(\mathbf{X}_i; \hat{\boldsymbol{\beta}})$ with parameter estimates $\hat{\boldsymbol{\beta}}$ is an estimator of the (transformed) conditional expectation $q(E(Y_i^1 | \mathbf{X}_i, S_i = 1))$. Based on the model, potential outcomes under the active intervention are predicted for each subject in the external control:

$$\hat{Y}_i^1 = q^{-1}(m(\mathbf{X}_i; \hat{\boldsymbol{\beta}})). \quad (14)$$

In this case, $i = n_1 + 1, \dots, n$, and the predicted outcomes are counterfactual because the subjects in the external control have not received the active intervention.

By averaging (“marginalizing”) the potential outcome predictions generated by Equation (14) over the empirical covariate distribution of the external control, one obtains the *G-computation* estimator,^{35,45–48} also known as “regression standardization” or “plug-in G-formula,” for the mean potential outcome μ_0^1 had subjects in the external control received the active intervention:

$$\hat{\mu}_0^1 = \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1. \quad (15)$$

Consequently, mean absolute outcomes are converted to the additive scale imposed by link function $g(\cdot)$, prior to taking the difference between treatments on such scale to derive the G-computation

estimator for the ATC:

$$\widehat{\text{ATC}} = g \left(\underbrace{\frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1}_{\hat{\mu}_0^1} \right) - g \left(\underbrace{\frac{1}{n_0} \sum_{i=n_1+1}^n Y_i}_{\hat{\mu}_0^0} \right). \quad (16)$$

Note that the link function $g(\cdot)$ used for summarizing the treatment effect does not necessarily need to match the link function $q(\cdot)$ used for modeling.⁸²

The readership may be more familiar with the well-known G-computation estimator for the ATE, which contrasts averages of *predicted* potential outcomes between both treatment groups.³⁴ Our G-computation estimator in Equation (16) contrasts the average of predicted counterfactual outcomes under the active intervention with the average of *observed* factual outcomes for the external control group. To estimate the ATC, only potential outcomes for the SAT subjects need to be predicted because all potential outcomes under the control have already been observed for the subjects in the external control.⁴⁵

The G-computation estimators in Equations (15) and (16) rely on the outcome model in Equation (13) being correctly specified, in order to be consistent for the mean absolute outcome μ_0^1 and the ATC, respectively. Assuming correct model specification for all estimators, G-computation is more precise and efficient than weighting, particularly when poor overlap leads to large reductions in ESS.^{35,48} However, the increase in precision is achieved by implicit extrapolation into non-overlapping regions of the covariate space, hiding underlying failures of the positivity assumption.⁵⁰ Model misspecification bias is almost impossible to diagnose in extrapolated regions, and there is typically no inflation of the variance to reflect the extrapolation uncertainty.

3.5. Doubly robust augmented weighting estimators

The limitations of the methods in Sections 3.2–3.4 motivate the explicit augmentation of the weighting estimators,⁸³ allowing for a less restrictive outcome model that permits non-linear link functions and/or outcomes that depend on non-linear covariate transformations. Our proposed augmented weighting estimators will not use the outcome model to extrapolate, but to gain bias-robustness—and, potentially, some precision³⁶—with respect to their weighting counterparts. To construct such augmented estimators, we proceed as follows.

Firstly, based on Funk et al. and Shinozaki and Matsuyama,^{83,84} we propose combining the modeling approach to inverse odds weighting, described in Section 3.2, with G-computation. Specifically, suppose we have fitted the outcome model in Equation (13) to the SAT. We now let $\hat{Y}_i^1 = q^{-1} \left(m(\mathbf{X}_i; \hat{\beta}) \right)$ be a prediction of potential outcome Y_i^1 for the active intervention based on the fitted regression, not only for the external control subjects but for all subjects $i = 1, \dots, n$ in the SAT and the external control. The propensity score weights $\hat{w}_i$ derived for $i = 1, \dots, n_1$ are used to add an error-correcting term to the G-computation estimator. The resulting *DR augmented IOW* estimator for the mean absolute outcome μ_0^1 is

$$\begin{aligned} \hat{\mu}_0^1 &= \frac{1}{n_0} \left(\sum_{i=1}^{n_1} \hat{w}_i (Y_i - \hat{Y}_i^1) \right) + \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1 \\ &= \frac{1}{n_0} \left(\sum_{i=1}^{n_1} \hat{w}_i \epsilon_i^1 + \sum_{i=n_1+1}^n \hat{Y}_i^1 \right), \end{aligned} \quad (17)$$

where $\epsilon_i^1 = Y_i - \hat{Y}_i^1$ is a residual term for subject $i = 1, \dots, n_1$ in the SAT. Note that this estimator exactly corresponds to the DR estimator proposed by Shinozaki and Matsuyama,⁸⁴ except that their

estimator is for $E(Y^0|S = 1)$ while the estimator in Equation (17) is for $E(Y^1|S = 0)$. The estimand ultimately targeted by Shinozaki and Matsuyama is the ATT but ours is the ATC.

In Equation (17), the G-computation estimator has been augmented with a weighted average of the residuals for the SAT subjects. We shall explain later in this section why such a weighted average is an error-correcting term for the potential bias of the G-computation estimator.⁸⁵ The corresponding DR augmented IOW estimator for the ATC is

$$\widehat{\text{ATC}} = g \left(\underbrace{\frac{1}{n_0} \left(\sum_{i=1}^{n_1} \hat{w}_i \epsilon_i^1 + \sum_{i=n_1+1}^n \hat{Y}_i^1 \right)}_{\hat{\mu}_0^1} \right) - g \left(\underbrace{\frac{1}{n_0} \sum_{i=n_1+1}^n Y_i}_{\hat{\mu}_0^0} \right), \quad (18)$$

on the additive scale imposed by link function $g(\cdot)$. As for the non-augmented IOW estimators in Section 3.2, we can normalize or stabilize the weights so that they sum to one, thereby ensuring bounded estimates and improving the finite sample properties of the estimator.⁶¹ Using the normalized weights, we obtain the *DR augmented normalized IOW* estimators:

$$\hat{\mu}_0^1 = \frac{\sum_{i=1}^{n_1} \hat{w}_i \epsilon_i^1}{\sum_{i=1}^{n_1} \hat{w}_i} + \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1, \quad (19)$$

$$\widehat{\text{ATC}} = g \left(\underbrace{\frac{\sum_{i=1}^{n_1} \hat{w}_i \epsilon_i^1}{\sum_{i=1}^{n_1} \hat{w}_i} + \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1}_{\hat{\mu}_0^1} \right) - g \left(\underbrace{\frac{1}{n_0} \sum_{i=n_1+1}^n Y_i}_{\hat{\mu}_0^0} \right). \quad (20)$$

Equations (19) and (20) should provide more stable and precise estimation than Equations (17) and (18), respectively.⁶¹

Our main contribution is combining the entropy balancing-based MAIC approach, described in Section 3.3, with the G-computation estimator. Again, based on the outcome model fitted to the SAT, let $\hat{Y}_i^1 = q^{-1} \left(m(\mathbf{X}_i; \hat{\beta}) \right)$ be a prediction of the potential outcome Y_i^1 under the active intervention for all subjects $i = 1, \dots, n$ in the SAT and the external control. We now employ the MAIC weights $\hat{v}_i$ for $i = 1, \dots, n_1$ derived in Equation (11), instead of the “modeling” IOW weights, to estimate the error-correcting term used to augment the G-computation estimator. The resulting *DR augmented MAIC* estimator of the mean absolute outcome μ_0^1 is

$$\begin{aligned} \hat{\mu}_0^1 &= \sum_{i=1}^{n_1} \hat{v}_i \left(Y_i - \hat{Y}_i^1 \right) + \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1 \\ &= \sum_{i=1}^{n_1} \hat{v}_i \epsilon_i^1 + \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1, \end{aligned} \quad (21)$$

where the G-computation estimator has been augmented with a weighted average of the residuals $\epsilon_i^1 = Y_i - \hat{Y}_i^1$ for $i = 1, \dots, n_1$, but this time the weighted average has been computed using the MAIC weights.

The corresponding DR augmented MAIC estimator for the ATC is

$$\widehat{\text{ATC}} = g \left(\underbrace{\sum_{i=1}^{n_1} \hat{v}_i \epsilon_i^1 + \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1}_{\hat{\mu}_0^1} \right) - g \left(\underbrace{\frac{1}{n_0} \sum_{i=n_1+1}^n Y_i}_{\hat{\mu}_0^0} \right), \quad (22)$$

on the additive scale imposed by link function $g(\cdot)$. We conjecture that the DR augmented MAIC estimators in Equations (21) and (22) will perform better statistically than the augmented estimators based on the modeling approach to weighting, which could exhibit erratic performance with highly variable weights, particularly if these are combined with a misspecified outcome model.^{86,87} We expect the DR augmented MAIC estimators to inherit the more attractive properties of the entropy balancing weights: (1) lower susceptibility to bias by directly enforcing covariate balance and (2) greater stability, translating into larger ESSs after weighting and enhanced precision in estimation.

The augmented weighting estimators in Equations (17), (19), and (21) are DR for the mean potential outcome μ_0^1 . That is, they estimate μ_0^1 consistently as long as either the propensity score model for data source assignment or the outcome model is correctly specified, but not necessarily both. Note that, contrary to the approach that we shall describe in Section 3.6, this double robustness does not depend on using a canonical link function for the outcome model. In the Appendix, we provide an intuitive heuristic to demonstrate double robustness, which illustrates why the weighted average of residuals is an error-correcting term for the potential bias of G-computation. The heuristic is summarized as follows. Consider that all the augmented weighting estimators have the general form:

$$\hat{\mu}_0^1 = \sum_{i=1}^{n_1} \hat{u}_i (Y_i - \hat{Y}_i^1) + \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1, \quad (23)$$

for a generic weight estimate $\hat{u}_i$, where $\hat{u}_i = \hat{w}_i/n_0$ for the DR augmented IOW estimator in Equation (17), $\hat{u}_i = \hat{w}_i/\sum_{i=1}^{n_1} \hat{w}_i$ for the DR augmented normalized IOW estimator in Equation (19), and $\hat{u}_i = \hat{v}_i$ for the DR augmented MAIC estimator in Equation (21).

If the outcome model is correctly specified, the expectation of the first summation in Equation (23) converges to zero as $n_1 \rightarrow \infty$ because $\hat{Y}_i^1 \rightarrow Y_i$ and the terms inside the summation cancel out, irrespective of any postulated propensity score model for data source assignment. The second summation is equivalent to the G-computation estimator and is consistent for μ_0^1 because the outcome model is correct. Consequently, $\hat{\mu}_0^1 \rightarrow \mu_0^1$ and $\widehat{ATC} \rightarrow ATC$, under the assumption that $\hat{\mu}_0^0 = \frac{1}{n_0} \sum_{i=n_1+1}^n Y_i$ is consistent for μ_0^0 . Conversely, if the propensity score model is correctly specified but the outcome model is incorrect, the first summation consistently cancels out the bias produced by the G-computation estimator in the second summation and the remainder term is exactly equal to the non-augmented weighting estimator, which converges to μ_0^1 as $n_1 \rightarrow \infty$ because the propensity score model is correct. Similarly, $\widehat{ATC} \rightarrow ATC$ (assuming $\hat{\mu}_0^0 \rightarrow \mu_0^0$). Hence, the augmented weighting estimators in Equations (18), (20), and (22) are DR for the ATC.

While all the augmented weighting estimators described in this section are DR, the DR augmented MAIC estimators defined in Equations (21) and (22) are arguably more robust to model misspecification bias because they are consistent under a greater number of distinct underlying data-generating mechanisms. Namely, DR augmented MAIC is consistent as long as either: (1) the log-odds of the propensity score are linear across the covariate balance functions; (2) the potential outcome under the active intervention is linear across the covariate balance functions; or (3) the explicit augmentation model for the potential outcome under the active intervention is correctly specified. The estimation of the weights is consistent as long as either the first or the second condition holds. Conversely, the augmented estimators based on the modeling approach to weighting (defined in Equations (17)–(20)) are consistent as long as either the first or the third condition holds, with the first condition being necessary for consistent estimation of the weights.

3.6. Weighted G-computation

In Section 3.5, we proposed augmented estimators that combine the predictions of an unweighted outcome model with weights in a weighted average. Nevertheless, there are other ways of constructing augmented weighting estimators. One approach popularized by Park et al.⁷⁶ that has been claimed to

be DR consists of G-computation based on the predictions of a weighted outcome model.^{61,76} Where the target estimand is the ATC, this involves: (1) estimating weights using the methods described in Sections 3.2 and 3.3; (2) fitting a weighted model for the conditional outcome expectation to the SAT participants; and (3) marginalizing the outcome predictions of the weighted regression over the external control covariate distribution. The resulting estimator for the mean absolute outcome μ_0^1 would be

$$\hat{\mu}_0^1 = \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1 = \frac{1}{n_0} \sum_{i=n_1+1}^n q^{-1} \left(m(\mathbf{X}_i; \hat{\beta}_v) \right), \quad (24)$$

where $m(\mathbf{X}_i; \hat{\beta}_v)$ indexes the fitted weighted regression with vector $\hat{\beta}_v$ of parameter estimates. The ATC would be estimated by substituting Equation (24) into Equation (16).

Such estimators are only DR where the outcome model is a generalized linear model (GLM) with a canonical link function $q(\cdot)$.^{61,86,88,89} The estimator in Equation (24) and the corresponding ATC estimator would not be DR where the GLM link function is non-canonical,⁸⁸ or where the outcome model is a Cox proportional hazards model or a parametric survival model in the time-to-event setting.⁹⁰ Nevertheless, results by Gabriel et al. suggest asymptotic equivalence and similar finite sample performance to the augmented weighting estimators in Section 3.5 for GLMs with canonical link functions fitted via maximum-likelihood,⁸⁸ provided that the same weights are used and correct model specification. We note that the target of the investigations by Gabriel et al. is the ATE and the modeling approach to weighting.⁸⁸

3.7. Variance estimation

To estimate the variance and construct confidence intervals (CIs) for $\hat{\mu}_0^1$ and $\widehat{ATC}$, it is possible to use empirical sandwich-type (“robust”) variance estimators to account for the correlation induced by weighting.^{25,39} In the specific context of non-randomized comparisons, such as the externally controlled SATs and unanchored ITCs explored in this article, these estimators have exhibited either under-precision or over-precision for the ATT (or the ATC)^{91,92} and under-precision for the ATE.^{37,65} This is because most implementations ignore the estimation of the propensity score model or the weights, assuming the weights to be fixed quantities.^{91,92}

Analytic expressions that incorporate weight estimation could be derived,^{40,91,92} but we propose a practical alternative based on the ordinary non-parametric bootstrap,³⁷ explicitly accounting for uncertainty in the weight estimation. This involves resampling with replacement the concatenated IPD consisting of the SAT and external control data. In each bootstrap iteration, the weight estimation and/or outcome modeling procedures are performed, and μ_0^1 , μ_0^0 and ATC are re-estimated. Standard errors for $\hat{\mu}_0^1$, $\hat{\mu}_0^0$, and $\widehat{ATC}$ are given by the standard deviations across the bootstrap resamples. Subsequently, Wald-type CIs can be constructed. Alternatively, one can directly calculate CIs from the percentiles of the bootstrap resamples, for example, 2.5% and 97.5% for the 95% CI.

3.8. External controls with unavailable individual participant data

In the context of unanchored ITCs, the external control is often a historical trial for which IPD are unavailable, due to privacy or confidentiality reasons. In this case, only published AD are available for the external control.^{25,28,33} Such data consist of marginal summary moments $\hat{\theta}$ from reported tables of baseline characteristics, typically only including means and standard deviations (for continuous covariates), and an estimate $\hat{\mu}_0^0$ of the mean absolute outcome under the control in the external data source.^{25,28,39,70,74} An important shortfall of this scenario is the need to assume that $\hat{\theta}_j = \theta_j$ with zero variability for the covariate balance function moments $j = 1, \dots, k$, that is, that the external control covariate distributional data are fixed.⁴¹ While this may be reasonable with large sample sizes for the external control, it can otherwise result in overly precise inferences and inflated Type I error rates.⁴¹

In this setting, for all methods except (non-augmented) MAIC, one must simulate M individual-level covariate profiles from the assumed covariate distribution of the external control based on published summary statistics.^{48,82,93} The number M of hypothetical subject profiles should be relatively large, for example, $M = 1,000$, to minimize sampling variability and random seed sensitivity, and does not necessarily need to match the original sample size n_0 of the external control.^{48,82,93} Necessary information to infer the joint covariate distribution of the external control, for example, distributional forms and correlation structures, is rarely published. Hence, this must be borrowed from other data sources or selected based on theoretical properties, following recommendations in the literature.^{25,28,48,82,93}

The notation and procedures for Sections 3.2 and 3.4–3.6 change as follows. The observed IPD for the SAT are stacked with the simulated subject-level covariate data for the external control. The concatenated dataset is now $(S_i, \mathbf{X}_i, T_i, Y_i)$ for $i = 1, \dots, n_1, n_1 + 1, \dots, n_1 + M$. For the SAT subjects $i = 1, \dots, n_1$, we have $S_i = 1$ and $T_i = 1$, with $\mathbf{X}_i$ and Y_i corresponding to the actual covariate and outcome values observed in the trial. For the hypothetical external controls $i = n_1 + 1, \dots, n_1 + M$, we have $S_i = 0$ and $T_i = 0$, the values of $\mathbf{X}_i$ are simulated, and Y_i are unavailable but not required for the analysis. This is because the target estimand is the ATC and the outcomes for the external group under the control have already been observed factually, with the mean estimate $\hat{\mu}_0^0$ available from published results.

The general form of the IOW estimators for the ATC, described in Section 3.2, is now:

$$\widehat{\text{ATC}} = g \left(\underbrace{\frac{1}{K} \sum_{i=1}^{n_1} \hat{w}_i Y_i}_{\hat{\mu}_0^1} \right) - g \left(\hat{\mu}_0^0 \right),$$

where K is a constant. There is only a change in notation here given that individual-level outcomes under the control are now unavailable for the external data source, and cannot be included in the concatenated dataset.

For the G-computation estimator for the ATC, outlined in Section 3.4, we now have

$$\widehat{\text{ATC}} = g \left(\underbrace{\frac{1}{M} \sum_{i=n_1+1}^{n_1+M} \hat{Y}_i^1}_{\hat{\mu}_0^1} \right) - g \left(\hat{\mu}_0^0 \right), \quad (25)$$

where the potential outcome predictions $\hat{Y}_i^1$ under the active intervention are generated for each hypothetical external control subject $i = n_1 + 1, \dots, n_1 + M$, and averaged over the simulated covariate profiles.

The general form of the DR augmented weighting estimators for the ATC, proposed in Section 3.5, is now:

$$\widehat{\text{ATC}} = g \left(\underbrace{\frac{1}{K} \sum_{i=1}^{n_1} \hat{\omega}_i \epsilon_i^1 + \frac{1}{M} \sum_{i=n_1+1}^{n_1+M} \hat{Y}_i^1}_{\hat{\mu}_0^1} \right) - g \left(\hat{\mu}_0^0 \right),$$

where K is a constant, $\epsilon_i^1 = Y_i - \hat{Y}_i^1$ and $\hat{\omega}_i$ are a residual term and a weight estimate, respectively, for $i = 1, \dots, n_1$, and the potential outcome predictions $\hat{Y}_i^1$ under the active intervention are generated for all SAT subjects and hypothetical external controls $i = 1, \dots, n_1, n_1 + 1, \dots, n_1 + M$.

For the weighted G-computation estimator in Section 3.6, the outcome predictions would be averaged over the simulated covariates for the external control. The resulting estimator for the mean absolute outcome μ_0^1 is $\hat{\mu}_0^1 = \frac{1}{M} \sum_{i=n_1+1}^{n_1+M} \hat{Y}_i^1 = \frac{1}{M} \sum_{i=n_1+1}^{n_1+M} q^{-1} \left(m(\mathbf{X}_i; \hat{\beta}_v) \right)$, which would be input into Equation (25) for estimation of the ATC.

The unavailability of IPD for the external control entails some changes to the non-parametric bootstrap procedure described in Section 3.7 for variance estimation. In this case, only the SAT data, $(S_i, \mathbf{X}_i, T_i, Y_i)$ for $i = 1, \dots, n_1$, are resampled to re-estimate $g(\mu_0^1)$ in each bootstrap iteration, with the standard error, $\text{SE}(g(\hat{\mu}_0^1))$, computed as the standard deviation over the bootstrap resamples. Then, the decomposition

$$\text{SE}(\widehat{\text{ATC}}) = \sqrt{(\text{SE}(g(\hat{\mu}_0^1)))^2 + (\text{SE}(g(\hat{\mu}_0^0)))^2} \quad (26)$$

is used to estimate the standard error of the ATC, where $\text{SE}(g(\hat{\mu}_0^0))$ is derived from published aggregate-level results.^{48,79,93} A limitation of the above formula is that it assumes that the mean absolute outcomes are statistically independent. Moreover, while computing $\text{SE}(g(\hat{\mu}_0^0))$ is a trivial exercise for continuous and binary outcomes (e.g., there is a closed-form formula for the standard error of the log-odds using the Delta method), it can be challenging for other outcomes such as those in the time-to-event setting.^{79,93} Once $\text{SE}(\widehat{\text{ATC}})$ is computed, Wald-type CIs can be readily constructed.

4. Simulation study

4.1. Aims

We conducted a simulation study to evaluate the performance of various estimators under different conditions. The simulation study design was planned following the structured “ADEMP” approach outlined by Morris et al.,⁹⁴ to ensure reproducibility and meaningful conclusions. Specifically, we clearly defined research aims, data-generating mechanisms under controlled scenarios and estimands, and assessed the performance of several estimators using relevant performance measures: bias, empirical standard error (ESE), and coverage. All simulations and analyses were performed using R statistical software version 4.3.1.⁹⁵ The files and code required to run the simulations are publicly available on Github at <https://github.com/harlanhappydog/DRAWE->.

4.2. Data-generating mechanisms

We simulated data inspired by the data-generating mechanisms in a simulation study by Kang and Schafer.⁸⁷ Some modifications were required since Kang and Schafer considered continuous-valued outcomes,⁸⁷ while we consider binary outcomes. The simulated data consist of $(\mathbf{X}_i, \mathbf{Z}_i, T_i, S_i, Y_i)$ for $i = 1, \dots, n$, with the control group fully external such that $S_i = T_i$, with $n_1 = \sum_{i=1}^n S_i$, and $n_0 = \sum_{i=1}^n (1 - S_i)$, as detailed in Section 3.1. While $\mathbf{X}_i$ is observed, $\mathbf{Z}_i$ is unobserved. To generate the data, $\mathbf{X}_i$ is distributed as $\text{Normal}(0, I_4)$, for i in $1, \dots, n$, and $\mathbf{Z}_i$ is then obtained by applying the following transformations:

$$\begin{aligned} Z_{i1} &= \text{scale}(\exp(X_{i1}/2)), \\ Z_{i2} &= \text{scale}(X_{i2}^2), \\ Z_{i3} &= \text{scale}((X_{i1}X_{i3} + 0.6)^3), \\ Z_{i4} &= \text{scale}((X_{i2} + X_{i4} + 20)^2), \end{aligned}$$

where $\text{scale}(\cdot)$ indicates normalization such that Z_{i1}, Z_{i2}, Z_{i3} , and Z_{i4} each have mean of 0 and standard deviation of 1, that is, $Z = \text{scale}(f(X)) = (f(X) - \text{mean}(f(X)))/\text{sd}(f(X))$. Note that these transformations are similar to the ones detailed by Kang and Schafer⁸⁷ but not identical, with changes made to highlight the consequences of model misspecification.

We consider four different scenarios. For each, we generated 10,000 simulated datasets of size $n = 200$ and $n = 1,000$. Note that, in all four scenarios, the distribution of S is approximately balanced such that $n_1 \approx n_0$. The four scenarios are defined as:

- **KS1:** Y_i is generated from a Bernoulli distribution with $\Pr(Y_i = 1 \mid \mathbf{X}_i, T_i) = \text{expit}(X_{1i} - 1.50X_{2i} + 0.5X_{3i} - 0.5X_{4i} + 1.50T_i - 0.50T_iX_{1i})$, where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with $\Pr(S_i = 1 \mid \mathbf{X}_i) = \text{expit}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.5X_{i4})$. The distribution of the covariates is such that overlap between the two groups is relatively high, with overlap proportions of 0.68, 0.85, 0.92, and 0.85 for X_1 , X_2 , X_3 , and X_4 , respectively (see Figure 3 in the Supplementary Material).
- **KS2:** Y_i is generated from a Bernoulli distribution with $\Pr(Y_i = 1 \mid \mathbf{Z}_i, T_i) = \text{expit}(Z_{1i} - 1.50Z_{2i} + 0.5Z_{3i} - 0.5Z_{4i} + 1.50T_i - 0.50T_iZ_{1i})$, where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with $\Pr(S_i = 1 \mid \mathbf{X}_i) = \text{expit}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.5X_{i4})$. The relevant covariate adjustment approaches would fit an outcome model to the observed $\mathbf{X}_i$, as the $\mathbf{Z}_i$ used for the true outcome-generating process are unobserved. The distribution of the covariates is such that overlap between the two groups is relatively high, with overlap proportions of 0.68, 0.85, 0.92, and 0.85 for X_1 , X_2 , X_3 , and X_4 , respectively (see Figure 3 in the Supplementary Material).
- **KS3:** Y_i is generated from a Bernoulli distribution with $\Pr(Y_i = 1 \mid \mathbf{X}_i, T_i) = \text{expit}(X_{1i} - 1.50X_{2i} + 0.5X_{3i} - 0.5X_{4i} + 1.50T_i - 0.50T_iX_{1i})$, where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with $\Pr(S_i = 1 \mid \mathbf{Z}_i) = \text{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.5Z_{i4})$. The relevant covariate adjustment approaches would balance or fit a propensity score model to the observed $\mathbf{X}_i$, as the $\mathbf{Z}_i$ used for the true data source assignment process are unobserved. The distribution of the covariates is such that overlap between the two groups is relatively high, with overlap proportions of 0.71, 0.84, 0.99, and 0.89 for X_1 , X_2 , X_3 , and X_4 , respectively (see Figure 4 in the Supplementary Material).
- **KS4:** Y_i is generated from a Bernoulli distribution with $\Pr(Y_i = 1 \mid \mathbf{Z}_i, T_i) = \text{expit}(Z_{1i} - 1.50Z_{2i} + 0.5Z_{3i} - 0.5Z_{4i} + 1.50T_i - 0.50T_iZ_{1i})$, where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with $\Pr(S_i = 1 \mid \mathbf{Z}_i) = \text{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.5Z_{i4})$. The distribution of the covariates is such that overlap between the two groups is relatively high, with overlap proportions of 0.71, 0.84, 0.99, and 0.89 for X_1 , X_2 , X_3 , and X_4 , respectively (see Figure 4 in the Supplementary Material).

We assume that there is unlimited access to subject-level data for the SAT and the external control, such that $(\mathbf{X}_i, T_i, S_i, Y_i)$ are observed for all $i = 1, \dots, n$.

4.3. Estimands

The estimand of interest is the ATC, as defined in Section 2. We adopt the logit link function $g(p) = \ln(p/(1-p))$ for marginal outcome probability p , such that the ATC is on the marginal log-odds ratio scale. The values of the ATC estimands were calculated numerically, by simulating 10 million binary outcomes using the true data-generating mechanisms outlined in Section 4.2. Data-generating mechanisms KS1, KS2, KS3, and KS4 correspond to true ATCs of 1.116, 1.215, 1.068, and 1.181, respectively.

4.4. Methods

We compared 16 estimators:

1. The naïve estimator, which does not perform covariate adjustment:

$$\widehat{\text{ATC}} = g\left(\underbrace{\frac{1}{n_1} \sum_{i=1}^{n_1} Y_i}_{\hat{\mu}_1^1}\right) - g\left(\underbrace{\frac{1}{n_0} \sum_{i=n_1+1}^n Y_i}_{\hat{\mu}_0^0}\right). \quad (27)$$

2. The *IOW* estimator as per Equation (3) (Section 3.2), with weights derived using the “modeling” approach with a logistic regression data source assignment model as per Equation (2).
3. The *normalized IOW* estimator with normalized weights as per Equation (4) (Section 3.2) derived using the “modeling” approach with a logistic regression data source assignment model as per Equation (2).
4. The *MAIC* (entropy balancing) estimator, as per Equation (12) (Section 3.3).
5. The *G-computation* estimator, as per Equation (16) (Section 3.4) with the outcome model defined with a logistic link function.
6. The *DR augmented IOW* estimator, as per Equation (18) (see Section 3.5) with the outcome model defined with a logistic link function.
7. The *DR augmented normalized IOW* estimator, as per Equation (20) (see Section 3.5) with the outcome model defined with a logistic link function.
8. The *DR augmented MAIC* (augmented entropy balancing) estimator, as per Equation (22) (our main contribution, see Section 3.5) with the outcome model defined with a logistic link function.
9. The *weighted G-computation (normalized IOW weights)* estimator described in Section 3.6 with the outcome model defined with a logistic link function, and using the (normalized) “modeling” *IOW* weights.
10. The *weighted G-computation (MAIC weights)* estimator described in Section 3.6 with the outcome model defined with a logistic link function, but using the *MAIC* (entropy balancing) weights instead of the “modeling” *IOW* weights.
- 11–16. Estimators 11–16 are the same as Estimators 5–10 but using a non-canonical Cauchit link function, $q(x) = \tan(\pi(x - 0.5))$. See Morgan and Smith⁹⁶ for an example of model fitting with the Cauchit.

To be clear, for Estimators 1–10, both the outcome model and the propensity score model for data source assignment will be correctly specified in KS1. For Estimators 11–16, the propensity score model for data source assignment will be correctly specified in KS1, but not the outcome model. In KS2, the outcome model will be incorrectly specified for all estimators fitting an outcome model (Estimators 5–16). Finally, in KS3, the propensity score model will be incorrectly specified for all estimators considering a propensity score model, and in KS4, both the outcome model and the propensity score model will be incorrectly specified.

For all estimators, we computed 95% CIs using the non-parametric bootstrap approach described in Section 3.7. Specifically, we used $B = 200$ bootstrap resamples of the concatenated SAT and external control IPD to approximate the standard error of $\widehat{ATC}$ and subsequently constructed Wald-type CIs.

4.5. Performance measures

To assess the performance of the estimators in our simulation study, we computed several key metrics: bias, ESE, and 95% CI coverage. Bias was calculated as the difference between the average of the point estimates across simulations and the true estimand value, providing a measure of systematic error. The ESE was computed as the standard deviation of the point estimates across simulations, reflecting the precision or variability of the different estimators. The 95% CI coverage was determined as the proportion of simulated datasets in which the constructed 95% CI contained the true estimand value, evaluating the quality of interval estimation. In addition, we estimated Monte Carlo standard errors (MCSEs) using the formulas provided by Morris et al. to quantify the uncertainty in the performance measures due to using a finite number of simulations.⁹⁴

4.6. Results

The complete results for the simulation study are displayed in Tables 1–4. Note that, for the $n = 200$ simulations, the MCSEs are less than 0.018 for bias, less than 0.013 for ESE, and less than 0.006 for 95% CI coverage; for the $n = 1,000$ simulations, the MCSEs are less than 0.011 for bias, less than 0.008 for ESE, and less than 0.005 for 95% CI coverage.

Under $n = 1,000$, all methods perform as expected in terms of bias. The naïve estimator appears biased in all four scenarios (but in KS3, the bias is very small), the DR estimators appear unbiased when either model is correct (Scenarios KS1–KS3), and the singly robust estimators avoid bias only when the corresponding propensity score or outcome model is correct. Unlike the three DR methods, the weighted G-computation estimators show bias in Scenario KS2 when the Cauchit link function is used in the outcome model (but not when the canonical logistic link is used). This suggests that the weighted G-computation estimators may be DR for the ATC when the canonical logit link is used, but not otherwise. The trends observed for the bias are similar under $n = 200$ with the caveat that the (augmented and non-augmented) weighting estimators that appeared unbiased under $n = 1,000$ exhibit some small-sample bias in the corresponding scenarios. This is particularly notable in Scenario KS2 and, to a lesser extent, in Scenario KS1, and is probably due to small ESSs after weighting.

Under correct specification of the outcome model (KS1 and KS3), G-computation is the most precise covariate-adjusted estimator, but the augmented estimators are almost as precise (e.g., compare the G-computation estimator which obtains $ESE = 0.150$ to the DR augmented MAIC estimator which obtains $ESE = 0.170$ for KS1 with $n = 1,000$; see Table 1). Moreover, the augmented estimators generally produce precision gains versus their respective non-augmented weighting counterparts. When both the outcome model and the propensity score model are correctly specified (KS1), all augmented estimators have increased precision compared to the non-augmented weighting estimators based on modeling weights, but not necessarily against MAIC (any increase in precision for $n = 1,000$ is modest). When only the propensity score model is correctly specified (KS2), outcome model misspecification does not induce any meaningful loss in precision for the augmented estimators compared to their non-augmented weighting counterparts.

There have been some concerns in the literature about DR augmented estimators amplifying bias and variance when misspecified weights are combined with a misspecified outcome model.⁸⁷ Such amplification is not apparent in our simulation study. In KS4, our proposed DR augmented MAIC estimator is the least biased of all estimators and the most precise of the augmented and non-augmented weighting estimators. Under $n = 1,000$, this advantage is modest when the logit link is used for the outcome model (bias of 0.482 vs. 0.512 for G-computation), but more pronounced when the Cauchit link is used (bias of 0.350 vs. 0.524 for G-computation).

In Section 3.3, we hypothesized that entropy balancing weights, like those employed by MAIC, can lead to more stable and precise ATC estimation than inverse odds modeling weights. This appears to be confirmed for the non-augmented estimators in our simulation study; MAIC exhibits greater precision than the approaches using (normalized or non-normalized) IOW modeling weights in all simulation scenarios. Additionally, the precision gains have been inherited by the augmented approaches. For the methods highlighted in Section 3.5, estimators using MAIC weights display enhanced precision compared to those using IOW modeling weights in all simulation scenarios, while producing similar levels of bias, even lower bias under dual model misspecification.

Assuming unbiasedness, interval estimation is appropriate if the coverage is approximately equal to 0.95; poor coverage can arise due to bias or to inadequate variance/interval estimation. Coverage is generally close to 0.95 for all covariate adjustment methods in the simulation scenarios under which they are unbiased, which suggests that our proposed non-parametric bootstrap approach for variance estimation is adequate. In the cases in which the DR methods are unbiased (i.e., KS1–KS3), coverage rates are between 0.935 and 0.947 across $n = 1,000$ scenarios. Interestingly, coverage rates seem to increase for these methods under $n = 200$ despite the small-sample bias, lying between 0.942 and 0.965. Note that, due to computational limitations, the non-parametric bootstrap approach in the simulation

Table 1. Results from Scenario KS1, where both the logit-link outcome model and the propensity score model are correctly specified.

Method	Bias	ESE	95% CI coverage	Average 95% CI width
<i>n</i> = 200, ATC = 1.116				
1. The naive estimator	0.623	0.326	0.539	1.295
2. IOW	0.027	0.516	0.944	1.979
3. Normalized IOW	0.055	0.456	0.945	1.761
4. MAIC	0.037	0.413	0.962	2.276
With logit link used for outcome model:				
5. G-computation	0.019	0.351	0.956	1.431
6. DR augmented IOW	0.029	0.424	0.955	1.663
7. DR augmented normalized IOW	0.029	0.403	0.949	1.605
8. DR augmented MAIC	0.031	0.406	0.955	1.718
9. Weighted G-computation (normalized IOW weights)	0.029	0.401	0.942	1.582
10. Weighted G-computation (MAIC weights)	0.026	0.402	0.946	1.770
With Cauchit link used for outcome model:				
11. G-computation	0.006	0.383	0.965	1.689
12. DR augmented IOW	0.026	0.432	0.954	1.682
13. DR augmented normalized IOW	0.026	0.413	0.951	1.635
14. DR augmented MAIC	0.028	0.408	0.957	1.717
15. Weighted G-computation (normalized IOW weights)	0.063	0.914	0.964	2.696
16. Weighted G-computation (MAIC weights)	0.052	0.415	0.963	1.857
<i>n</i> = 1,000, ATC = 1.116				
1. The naive estimator	0.604	0.142	0.008	0.562
2. IOW	0.008	0.205	0.949	0.810
3. Normalized IOW	0.012	0.197	0.942	0.750
4. MAIC	0.008	0.172	0.944	0.660
With logit link used for outcome model:				
5. G-computation	0.005	0.150	0.950	0.593
6. DR augmented IOW	0.007	0.175	0.948	0.679
7. DR augmented normalized IOW	0.006	0.173	0.947	0.669
8. DR augmented MAIC	0.007	0.170	0.942	0.653
9. Weighted G-computation (normalized IOW weights)	0.006	0.170	0.940	0.648
10. Weighted G-computation (MAIC weights)	0.007	0.171	0.939	0.649
With Cauchit link used for outcome model:				
11. G-computation	−0.020	0.158	0.947	0.641
12. DR augmented IOW	0.006	0.178	0.946	0.683
13. DR augmented normalized IOW	0.006	0.175	0.945	0.673
14. DR augmented MAIC	0.006	0.172	0.940	0.658
15. Weighted G-computation (normalized IOW weights)	−0.003	0.180	0.948	0.781
16. Weighted G-computation (MAIC weights)	−0.004	0.173	0.942	0.677

Table 2. Results from Scenario KS2, where the outcome model is incorrectly specified and the propensity score model is correctly specified.

Method	Bias	ESE	95% CI coverage	Average 95% CI width
<i>n</i> = 200, ATC = 1.215				
1. The naive estimator	0.230	0.324	0.909	1.279
2. IOW	0.023	0.642	0.929	2.393
3. Normalized IOW	0.060	0.511	0.940	1.952
4. MAIC	0.061	0.517	0.965	2.778
With logit link used for outcome model:				
5. G-computation	0.088	0.434	0.952	1.739
6. DR augmented IOW	0.050	0.532	0.949	2.093
7. DR augmented normalized IOW	0.053	0.526	0.942	2.007
8. DR augmented MAIC	0.048	0.489	0.953	2.042
9. Weighted G-computation (normalized IOW weights)	0.056	0.480	0.937	1.868
10. Weighted G-computation (MAIC weights)	0.036	0.480	0.949	2.241
With Cauchit link used for outcome model:				
11. G-computation	0.222	0.506	0.951	2.085
12. DR augmented IOW	0.046	0.517	0.955	2.053
13. DR augmented normalized IOW	0.046	0.505	0.950	1.974
14. DR augmented MAIC	0.039	0.474	0.958	1.990
15. Weighted G-computation (normalized IOW weights)	0.225	1.950	0.964	5.768
16. Weighted G-computation (MAIC weights)	0.132	0.486	0.972	2.478
<i>n</i> = 1,000, ATC = 1.215				
1. The naive estimator	0.226	0.141	0.658	0.557
2. IOW	0.020	0.283	0.950	1.085
3. Normalized IOW	0.016	0.220	0.939	0.834
4. MAIC	0.015	0.205	0.936	0.779
With logit link used for outcome model:				
5. G-computation	0.075	0.188	0.937	0.736
6. DR augmented IOW	0.016	0.228	0.942	0.866
7. DR augmented normalized IOW	0.016	0.224	0.938	0.849
8. DR augmented MAIC	0.014	0.205	0.935	0.776
9. Weighted G-computation (normalized IOW weights)	0.017	0.203	0.936	0.778
10. Weighted G-computation (MAIC weights)	0.014	0.203	0.936	0.770
With Cauchit link used for outcome model:				
11. G-computation	0.228	0.284	0.817	1.012
12. DR augmented IOW	0.016	0.224	0.942	0.854
13. DR augmented normalized IOW	0.016	0.220	0.940	0.836
14. DR augmented MAIC	0.012	0.199	0.937	0.756
15. Weighted G-computation (normalized IOW weights)	0.163	1.270	0.935	2.820
16. Weighted G-computation (MAIC weights)	0.114	0.189	0.905	0.743

study was conducted with only $B = 200$ resamples. This may have impacted the observed coverage rates and we suspect that coverage might be more appropriate when using a larger number of resamples.

While some covariate adjustment methods display bias-induced undercoverage in the scenarios under which they are biased (e.g., MAIC in KS3 under $n = 1,000$ or all estimators in KS4), they

Table 3. Results from Scenario KS3, where the propensity score model is incorrectly specified. The logit-link outcome model is correctly specified; however, when the Cauchit link is used for the outcome model, both models are incorrectly specified.

Method	Bias	ESE	95% CI coverage	Average 95% CI width
<i>n</i> = 200, ATC = 1.068				
1. The naive estimator	−0.032	0.307	0.950	1.211
2. IOW	0.147	0.583	0.958	2.246
3. Normalized IOW	−0.021	0.397	0.950	1.539
4. MAIC	0.132	0.353	0.955	1.561
With logit link used for outcome model:				
5. G-computation	0.015	0.286	0.962	1.180
6. DR augmented IOW	0.028	0.344	0.964	1.404
7. DR augmented normalized IOW	0.027	0.335	0.958	1.338
8. DR augmented MAIC	0.024	0.313	0.962	1.296
9. Weighted G-computation (normalized IOW weights)	0.020	0.305	0.958	1.245
10. Weighted G-computation (MAIC weights)	0.018	0.305	0.959	1.300
With Cauchit link used for outcome model:				
11. G-computation	−0.031	0.296	0.969	1.316
12. DR augmented IOW	0.008	0.351	0.964	1.412
13. DR augmented normalized IOW	0.005	0.332	0.960	1.348
14. DR augmented MAIC	0.010	0.316	0.965	1.323
15. Weighted G-computation (normalized IOW weights)	0.018	0.701	0.972	2.176
16. Weighted G-computation (MAIC weights)	−0.001	0.320	0.970	1.448
<i>n</i> = 1,000, ATC = 1.068				
1. The naive estimator	−0.039	0.135	0.937	0.530
2. IOW	0.118	0.231	0.969	0.917
3. Normalized IOW	−0.044	0.168	0.933	0.645
4. MAIC	0.104	0.148	0.894	0.574
With logit link used for outcome model:				
5. G-computation	0.005	0.124	0.949	0.489
6. DR augmented IOW	0.007	0.144	0.945	0.558
7. DR augmented normalized IOW	0.007	0.142	0.943	0.547
8. DR augmented MAIC	0.006	0.133	0.943	0.517
9. Weighted G-computation (normalized IOW weights)	0.006	0.130	0.943	0.508
10. Weighted G-computation (MAIC weights)	0.006	0.130	0.944	0.506
With Cauchit link used for outcome model:				
11. G-computation	−0.053	0.124	0.926	0.496
12. DR augmented IOW	−0.015	0.144	0.942	0.557
13. DR augmented normalized IOW	−0.017	0.141	0.939	0.546
14. DR augmented MAIC	−0.010	0.134	0.941	0.520
15. Weighted G-computation (normalized IOW weights)	−0.044	0.131	0.937	0.606
16. Weighted G-computation (MAIC weights)	−0.045	0.130	0.933	0.518

may also display adequate coverage (e.g., normalized IOW with *n* = 200) because of excessively large standard errors, probably due to low ESSs after weighting. As observed for KS4, bias-induced undercoverage tends to worsen with higher sample sizes, as interval estimates around the wrong target

Table 4. Results from Scenario KS4, where both the outcome model and the propensity score model are incorrectly specified.

Method	Bias	ESE	95% CI coverage	Average 95% CI width
<i>n</i> = 200, ATC = 1.181				
1. The naive estimator	0.521	0.334	0.685	1.330
2. IOW	0.807	0.778	0.912	2.793
3. Normalized IOW	0.582	0.470	0.760	1.840
4. MAIC	0.613	0.463	0.790	2.032
With logit link used for outcome model:				
5. G-computation	0.536	0.376	0.750	1.551
6. DR augmented IOW	0.586	0.464	0.767	1.846
7. DR augmented normalized IOW	0.580	0.446	0.753	1.770
8. DR augmented MAIC	0.517	0.407	0.779	1.671
9. Weighted G-computation (normalized IOW weights)	0.546	0.415	0.760	1.692
10. Weighted G-computation (MAIC weights)	0.545	0.423	0.786	1.808
With Cauchit link used for outcome model:				
11. G-computation	0.555	0.363	0.799	1.677
12. DR augmented IOW	0.455	0.430	0.850	1.767
13. DR augmented normalized IOW	0.456	0.415	0.837	1.704
14. DR augmented MAIC	0.400	0.386	0.878	1.654
15. Weighted G-computation (normalized IOW weights)	0.575	1.291	0.871	4.635
16. Weighted G-computation (MAIC weights)	0.515	0.405	0.882	2.135
<i>n</i> = 1,000, ATC = 1.181				
1. The naive estimator	0.495	0.146	0.069	0.575
2. IOW	0.778	0.350	0.340	1.420
3. Normalized IOW	0.514	0.197	0.256	0.768
4. MAIC	0.546	0.190	0.174	0.739
With logit link used for outcome model:				
5. G-computation	0.512	0.161	0.110	0.636
6. DR augmented IOW	0.536	0.185	0.183	0.729
7. DR augmented normalized IOW	0.534	0.183	0.176	0.716
8. DR augmented MAIC	0.482	0.170	0.196	0.667
9. Weighted G-computation (normalized IOW weights)	0.524	0.179	0.159	0.714
10. Weighted G-computation (MAIC weights)	0.533	0.181	0.162	0.708
With Cauchit link used for outcome model:				
11. G-computation	0.524	0.144	0.050	0.578
12. DR augmented IOW	0.395	0.173	0.368	0.679
13. DR augmented normalized IOW	0.400	0.171	0.347	0.666
14. DR augmented MAIC	0.350	0.158	0.403	0.620
15. Weighted G-computation (normalized IOW weights)	0.488	0.165	0.226	1.089
16. Weighted G-computation (MAIC weights)	0.479	0.158	0.155	0.646

value become narrower. The naïve estimator displays discernible undercoverage in KS1, KS2, and KS4 (particularly under $n = 1,000$), not only due to bias but also due to overprecise standard errors that do not account for covariate differences.

Table 5. Summary statistics of the four baseline covariates identified as imbalanced prognostic factors, before and after weighting using MAIC (entropy balancing) and normalized inverse odds weighting (IOW). The standard deviation of age in the weighted columns is $\sqrt{\sum_{i=1}^{n_1} v_i (X_{1,i} - \sum_{i=1}^{n_1} v_i X_{1,i})^2}$, where $X_{1,i}$ and v_i are the age and the weight, respectively, for subject $i = 1, \dots, n_1$ in the intervention SAT.

Covariate	Intervention SAT ($n_1 = 500$)	External control ($n_0 = 300$)	Normalized IOW weighted SAT (ESS = 153.42)	MAIC weighted SAT (ESS = 157.07)
Age in years (mean; SD)	59.85; 9.01	50.06; 3.24	49.53; 3.18	50.06; 3.24
Sex (proportion male)	0.38	0.49	0.45	0.49
ECOG (proportion status 1)	0.41	0.35	0.30	0.35
Smoking (proportion smokers)	0.32	0.19	0.20	0.19

5. Applied example

We now demonstrate the application of some of the methods outlined in Section 3 to synthetic lung cancer clinical trial data. The data were obtained from the “MAIC” R package, implemented by researchers from the pharmaceutical industry.⁹⁷ Our objective is to compare the objective response, a binary outcome Y , under two treatments: the active “intervention” ($T = 1$) and the external “control” ($T = 0$). The data consist of IPD from a SAT ($S = 1$) with $n_1 = 500$ subjects under the “intervention,” and AD from an external historical SAT ($S = 0$) of $n_0 = 300$, which makes up the “control.” The unavailability of IPD for the external control allows us to illustrate the methodological extensions described in Section 3.8. R code to reproduce the applied example is provided in the [Supplementary Material](#).

The target estimand is the ATC on the marginal log-odds ratio scale. Four baseline covariates, one continuous—age—and three binary—sex, the Eastern Cooperative Oncology Group (ECOG) performance status and smoking status—have been identified as prognostic factors under the intervention, and are imbalanced between the intervention SAT and external control samples. There are no missing values for baseline characteristics and outcomes. Subjects in the intervention SAT are, on average, somewhat older, less likely to be male, more likely to be physically restricted (as indicated by ECOG performance status), and more likely to be smokers, relative to subjects in the external control (Table 5). In addition, the age of subjects in the intervention SAT has substantially greater variance than that of subjects in the external control.

We consider the naïve estimator first, which does not perform covariate adjustment (Equation (27)). In the intervention SAT, 390 of $n_1 = 500$ subjects attained objective response, which equates to a 78% response rate, $\hat{\mu}_1^1 = 0.78$. This implies a log-odds of response of $g(\hat{\mu}_1^1) = 1.266$, where $g(\cdot) = \text{logit}(\cdot)$. In the external control, 120 of $n_0 = 300$ subjects attained objective response, which equates to a 40% response rate, $\hat{\mu}_0^0 = 0.40$. This implies a log-odds of response of $g(\hat{\mu}_0^0) = -0.405$. A naïve estimate is obtained by simple subtraction: $\widehat{ATC}_{naive} = 1.266 - (-0.405) = 1.671$ and using the Delta method, we obtain: $SE(g(\hat{\mu}_0^0)) = 0.118$ and $SE(g(\hat{\mu}_1^1)) = 0.108$, and a Wald-type 95% CI of (1.358, 1.984).

Next, consider the *normalized IOW* estimator. A logistic regression model for the probability of SAT participation, conditional on age, sex, ECOG performance status, smoking status, and age-squared is fitted to the concatenated individual-level data comprising the intervention SAT and $M = 10,000$ simulated covariate profiles for the external control. The estimated propensity scores are used to derive IOW, which are subsequently normalized to sum to one, yielding the Hajek-type estimator in Equation (4). Table 5 shows that the IOW-weighted covariate means are close to, but do not exactly match, those of the external control. The ESS of the IOW-weighted intervention SAT is 153.42 and standardized mean differences are substantially reduced after weighting (e.g., from 1.445 to -0.079 for age), indicating that the weights achieve adequate, though not exact, balance. Using the

normalized IOW weights, the ATC is estimated as: $\widehat{ATC}_{IOW} = 1.333$. Having previously calculated $SE(g(\widehat{\mu}_0^0)) = 0.118$, we use the non-parametric bootstrap with $B = 10,000$ resamples to obtain $SE(g(\widehat{\mu}_0^1)) = 0.183$, and per Equation (26), we then obtain $SE(\widehat{ATC}_{IOW}) = 0.216$ and a Wald-type 95% CI of (0.911, 1.756).

We now consider the MAIC estimator. The positivity assumption is assessed using a method proposed by Glimm and Yau, which verifies whether covariate AD from the external control lie within the convex hull of the SAT covariate space, and whether the MAIC numerical optimization algorithm can converge.⁶⁰ This method is implemented using the “maicLP” R function in the “maicChecks” R package,⁹⁸ which confirms that a feasible weighting solution to the MAIC convex optimization problem exists, that is, that there is a set of positive weights that can enforce covariate balance between the intervention SAT and the external control, and that the MAIC numerical optimization algorithm can converge.

MAIC is performed using the procedure described in Section 3.3. We choose to weight the intervention SAT such that the means of all four baseline covariates and the variance of age are exactly balanced with respect to the external control. Following the notation in Section 3.3, we have

$$\mathbf{c}(\mathbf{X}) = [\text{Age}, \text{Sex}, \text{ECOG}, \text{Smoking}, \text{Age}^2]^\top = \begin{pmatrix} 45 & 71 & \cdots & 58 \\ 1 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ 2025 & 5041 & \cdots & 3364 \end{pmatrix},$$

where $\mathbf{c}(\mathbf{X})$ is a 5-by-500 matrix, with the rows representing the age, sex, ECOG performance status, smoking status, and age-squared for subjects in the intervention SAT. Using the BFGS convex optimization algorithm to minimize the objective function in Equation (10), we obtain $\hat{\gamma} = (3.542, 0.589, -0.698, -0.048, -0.036)$, and weights are calculated subject to the constraint that they sum to one. Figure 1 shows a histogram illustrating the empirical distribution of the resulting MAIC weights compared to the normalized IOW weights. The ESS—that is, the number of independent non-weighted observations that would be required to give an estimate with approximately the same precision as the weighted sample estimate—of the intervention SAT after weighting is 157.07. Using the MAIC weights, the ATC is estimated as: $\widehat{ATC}_{MAIC} = 1.331$. Having previously calculated $SE(g(\widehat{\mu}_0^0)) = 0.118$, we use the non-parametric bootstrap with $B = 10,000$ resamples to obtain $SE(g(\widehat{\mu}_0^1)) = 0.177$, and per Equation (26), we then obtain $SE(\widehat{ATC}_{MAIC}) = 0.212$ and a Wald-type 95% CI of (0.915, 1.748).

To perform *G-computation* with unavailable IPD for the external control, we first simulate $M = 10,000$ individual-level covariate profiles from an assumed joint covariate distribution for the external control, as per the approach outlined by Remiro-Azócar et al.⁴⁸ or the “infinite population” STC method described by Zhang et al.⁹⁹ We proceed by assuming that the pairwise correlation structure of the four covariates in the external control is equal to that observed in the intervention SAT:

$$\begin{pmatrix} 1.00 & 0.03 & 0.00 & 0.00 \\ 0.03 & 1.00 & -0.14 & -0.02 \\ 0.00 & -0.14 & 1.00 & -0.01 \\ 0.00 & -0.02 & -0.01 & 1.00 \end{pmatrix},$$

with the rows/columns in the order: age, sex, ECOG performance status, and smoking status. For age, we assumed a Normal(50.06, 3.24) marginal distribution; for sex, ECOG performance status and smoking status, we assumed Bernoulli(0.49), Bernoulli(0.35), and Bernoulli(0.19) marginal distributions, respectively, based on the summary statistics of the external control in Table 5. Individual-level covariates were ultimately simulated from a Gaussian copula characterized by the aforementioned

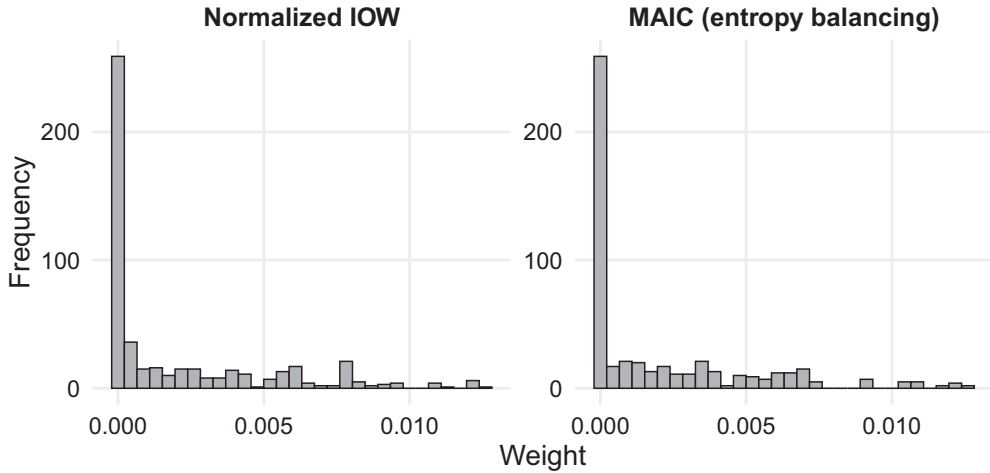


Figure 1. Histograms of the normalized IOW weights (left) and MAIC (entropy balancing) weights (right).

marginal distributions and pairwise correlation structure, using the “add_integration” function from the multinma R package.¹⁰⁰

Subsequently, a logistic-link binomial GLM for the outcome expectation under the intervention, conditional on baseline covariates, was postulated. This relates objective response Y_i^1 under the intervention $T = 1$ to baseline covariates $\mathbf{X}_i = (\text{Age}_i, \text{Sex}_i, \text{ECOG}_i, \text{Smoking}_i)^\top$ as

$$\text{logit} \left(E(Y_i^1 \mid \mathbf{X}_i; \boldsymbol{\beta}) \right) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Sex}_i + \beta_3 \text{ECOG}_i + \beta_4 \text{Smoking}_i + \beta_5 \text{Age}_i^2,$$

for i in $1, \dots, n_1$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\top$ are regression coefficients. The model was fitted to the intervention SAT using maximum-likelihood estimation with Fisher scoring, and we obtained regression coefficient estimates of $\hat{\beta}_0 = 5.72$, $\hat{\beta}_1 = -0.20$, $\hat{\beta}_2 = 0.12$, $\hat{\beta}_3 = 0.13$, $\hat{\beta}_4 = 0.01$, and $\hat{\beta}_5 = 0.00$. Then, following the G-computation procedure outlined in Section 3.4, the ATC is estimated as: $\widehat{\text{ATC}}_{Gcomp} = 1.325$. Using the non-parametric bootstrap with $B = 10,000$ resamples, we obtain $\text{SE}(g(\hat{\mu}_0^1)) = 0.164$, and, as per Equation (26), then obtain $\text{SE}(\widehat{\text{ATC}}_{Gcomp}) = 0.202$, and a Wald-type 95% CI of (0.929, 1.722).

Finally, our *DR augmented MAIC* estimator proposed in Section 3.5 produces the estimate: $\text{ATC}_{DR} = 1.332$. Using the non-parametric bootstrap with $B = 10,000$ resamples, we obtain $\text{SE}(g(\hat{\mu}_0^1)) = 0.179$. Then, as per Equation (26), we have $\text{SE}(\widehat{\text{ATC}}_{DR}) = 0.214$ and a Wald-type 95% CI of (0.912, 1.751).

Figure 2 shows the point estimates obtained using the five different estimators alongside their 95% CIs. When comparing the covariate-adjusted approaches to the naïve approach, we observe that covariate adjustment shifts the point estimate toward the null considerably. When comparing *normalized IOW*, *MAIC*, *G-computation*, and our proposed *DR augmented MAIC* estimator, results across the four methods seem consistent. Despite the shift toward the null, results suggest that the intervention improves objective response versus the control, statistically significantly at the 5% level. In this case, the DR point estimate is not meaningfully different than the MAIC or G-computation point estimates, and the DR approach offers slightly increased standard errors and wider CIs than G-computation. Nevertheless, this loss of precision seems a relatively minor price to pay, compensated for by greater reassurance in our results due to increased protection against misspecification of the outcome model.

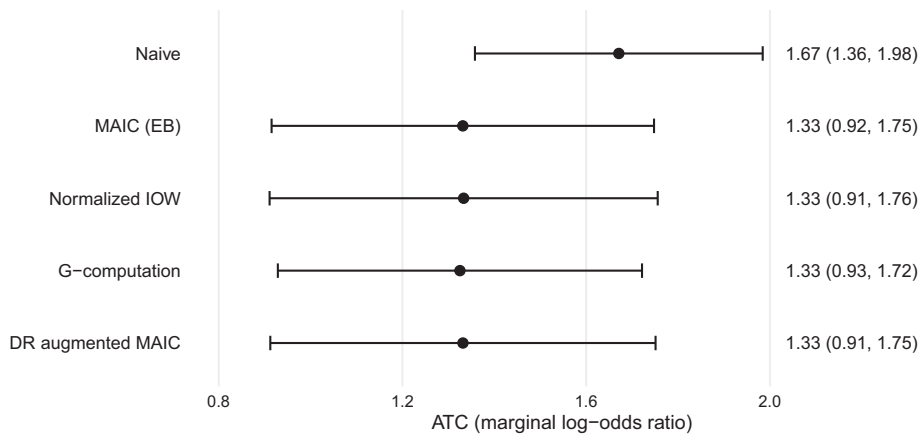


Figure 2. Point estimates with 95% CIs of the ATC (marginal log-odds ratio of objective response) for the different estimators in the applied example. DR denotes doubly robust and EB denotes entropy balancing.

6. Discussion

The objective of this article was to clarify existing approaches for DR estimation in the context of externally controlled SATs and to propose an estimator that augments approaches based on MAIC or entropy balancing via an outcome model. We described and illustrated an extension of this estimator specifically tailored to unanchored ITCs, for the setting with unavailable external control IPD, which is commonly encountered in practice. In a simulation study and applied example, we evaluated the performance and demonstrated the use of different DR augmented estimators, highlighting their merits with respect to the more popular non-augmented singly robust estimators.

Our findings reinforce the understanding that “balancing” approaches to weighting, such as MAIC or entropy balancing, can enhance performance relative to standard “modeling” approaches, but lack the DR property for non-linear outcome models. Conversely, augmented MAIC demonstrates double robustness and exhibits higher precision than non-augmented weighting estimators when the outcome model is correctly specified. Moreover, it generally achieves near-identical precision to G-computation, which offers the lowest variance under correct specification of the outcome model, but may exhibit notable bias where the outcome model is misspecified.

A potential concern about augmented estimators has been possible bias and variance amplification where both the propensity score and the outcome model are incorrectly specified. Nevertheless, such amplification under dual model misspecification was not observed in our simulation study, and one can argue that risks are mitigated by employing “balancing” instead of “modeling” weights (for reasons outlined in the introduction to Section 3.3), but further theoretical work and simulation studies are required to fully support this claim. The results of our simulation study motivate the routine application of DR augmented estimators, particularly those based on MAIC or entropy balancing weights, in practical settings. It is unfortunate to see that virtually all applications of MAIC do not consider augmentation via an outcome model, and that most practical uses of augmented estimators apply “modeling” instead of “balancing” weights.¹⁰¹

The weighted G-computation estimator described in Section 3.6 exhibited double robustness for the ATC when a canonical logistic link function was defined for the outcome model, but not when the link function was non-canonical, which coincides with what has been suggested previously.⁸⁸ Future simulation studies should consider non-binary outcomes, including survival outcomes with censoring, and other summary effect measures beyond the log-odds ratio.

Our simulation study and applied example considered scenarios with relatively low sample sizes in the SAT and the external control, corresponding to typical settings in rare disease and late-stage hematological or solid tumor oncology, where the number of subjects enrolled in SATs can be 100 or 200, but may also consist of several hundreds. The number of external controls can be equally small; as such, our findings are potentially applicable where the target estimand is the ATT instead of the ATC, which requires weighting the external control as opposed to the SAT. Somewhat worryingly, all augmented and non-augmented weighting estimators displayed some small-sample bias in our simulation study under a total sample size of $n = 200$, even if modeling assumptions were correct. This is consistent with previous research that has also noted small-sample bias.^{102,103}

The level of (deterministic) overlap between the SAT and external control covariate distributions in our simulation study was relatively strong. The performance of augmented and non-augmented weighting estimators with respect to G-computation will likely worsen as overlap decreases, particularly in conjunction with small sample sizes. Nevertheless, we conjecture that the performance of balancing-based approaches will suffer less than that of their corresponding modeling-based counterparts, due to generating more stable and less extreme weights. This is unless a complete lack of overlap results in the absence of a solution to the convex optimization balancing problem, in which case extrapolating via G-computation might be the only option.

Finally, it is important to emphasize that all of the covariate-adjusted estimators we considered, including the DR augmented approaches, require the important and untestable assumption of no unmeasured prognostic factors. This is typically the main credibility concern of externally controlled SATs. In practice, important prognostic factors may be unknown or unavailable in at least one of the SAT or external control data sources. Researchers should always consider expert knowledge (e.g., consult clinicians, epidemiologists, or statisticians with relevant expertise) and review the relevant literature to evaluate the plausibility of the no unmeasured prognostic factors assumption. An important area of future research is the development of sensitivity analysis or quantitative bias analysis methods to help explore the sensitivity of results to unmeasured prognostic factors, in the specific context of externally controlled SATs and unanchored ITCs.¹⁰⁴

Author contributions. H.C. and A.R.-A. conceived the idea for this project, contributed to the design of the study and to the interpretation of the results. H.C. and A.R.-A. wrote the first draft of the manuscript and approved the final version of the article.

Competing interest statement. H.C. is employed by Precision AQ, a life sciences consultancy company, and A.R.-A. is employed by Novo Nordisk, a pharmaceutical company. No competing interests are declared as this research is purely methodological.

Data availability statement. The files required to generate the data, run the simulations, and reproduce the results of the simulation study and reproduce the applied example are available at <https://github.com/harlanhappydog/DRAWE->, DOI: [10.5281/zenodo.20451858](https://doi.org/10.5281/zenodo.20451858).

Funding statement. The authors declare that no specific funding has been received for this article.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/rsm.2026.10106>.

AI disclosure. During the preparation of this work, the authors used AI tools (Claude) for editing purposes. The authors have reviewed and edited the final content.

References

- [1] Jahanshahi M, Gregg K, Davis G, et al. The use of external controls in FDA regulatory decision making. *Ther Innov Regul Sci.* 2021;55(5): 1019–1035.
- [2] Mishra-Kalyani P, Kordestani LA, Rivera D, et al. External control arms in oncology: current use and future directions. *Ann Oncol.* 2022;33(4): 376–383.
- [3] Horton DB, Blum MD, Burcu M. Real-world evidence for assessing treatment effectiveness and safety in pediatric populations. *J Pediatr.* 2021;238(1): 312–316.
- [4] Beaulieu-Jones BK, Finlayson SG, Yuan W, et al. Examining the use of real-world evidence in the regulatory process. *Clin Pharmacol Ther.* 2020;107(4): 843–852.

- [5] Flynn R, Plueschke K, Quinten C, et al. Marketing authorization applications made to the European medicines agency in 2018–2019: what was the contribution of real-world evidence? *Clin Pharmacol Ther.* 2022;111(1): 90–97.
- [6] Zou KH, Vigna C, Talwai A, et al. The next horizon of drug development: external control arms and innovative tools to enrich clinical trial data. *Ther Innov Regul Sci.* 2024;58(3): 443–455.
- [7] Curtis LH, Sola-Morales O, Heidt J, et al. Regulatory and HTA considerations for development of real-world data derived external controls. *Clin Pharmacol Ther.* 2023;114(2): 303–315.
- [8] Gray CM, Grimson F, Layton D, Pocock S, Kim J. A framework for methodological choice and evidence assessment for studies using external comparators from real-world data. *Drug Saf.* 2020;43(7): 623–633.
- [9] Sola-Morales O, Curtis LH, Heidt J, et al. Effectively leveraging RWD for external controls: a systematic literature review of regulatory and HTA decisions. *Clin Pharmacol Ther.* 2023;114(2): 325–355.
- [10] FDA Guidance for Industry. Considerations for the design and conduct of externally controlled trials for drug and biological products. 2023. Accessed: 2024-09-22. <https://www.fda.gov/media/164960/download>.
- [11] Committee for Medicinal Products for Human Use (CHMP). Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation application. 2024. Accessed: 2024-09-25. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence-marketing-authorisation-application_en.pdf.
- [12] Bakker E, Plueschke K, Jonker CJ, Kurz X, Starokozhko V, Mol PG. Contribution of real-world evidence in European medicines agency's regulatory decision making. *Clin Pharmacol Ther.* 2023;113(1): 135–151.
- [13] Goring S, Taylor A, Müller K, et al. Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. *BMJ Open.* 2019;9(2): e024895.
- [14] National Institute for Health and Care. Excellence health technology evaluations: the manual. 2022. Accessed: 2024-09-23. <https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741>.
- [15] CADTH. Procedures for reimbursement reviews. 2024. Accessed: 2024-09-23. https://www.cda-amc.ca/sites/default/files/Drug_Review_Process/CADTH%20Drug%20Reimbursement%20Review%20Procedures.pdf.
- [16] IQWiG. General methods version 7.0. 2023. Accessed: 2024-09-23. https://www.iqwig.de/methoden/general-methods_version-7-0.pdf.
- [17] Patel D, Grimson F, Mihaylova E, et al. Use of external comparators for health technology assessment submissions based on single-arm trials. *Value Health.* 2021;24(8): 1118–1125.
- [18] National Institute for Health and Care. Excellence real-world evidence framework. 2022. Accessed: 2024-09-23. <https://www.nice.org.uk/corporate/ecd9/resources/nice-realworld-evidence-framework-pdf-1124020816837>.
- [19] Institute for Quality and Efficiency in Health Care (IQWiG). Concepts for the generation of routine practice data and their analysis for the benefit assessment of drugs according to §35a Social Code Book V. 2020. Accessed: 2024-09-23. https://www.iqwig.de/download/a19-43_routine-practice-data-for-the-benefit-assessment-of-drugs_extract-of-rapid-report_v1-0.pdf.
- [20] Jaksa A, Louder A, Maksymiuk C, et al. A comparison of 7 oncology external control arm case studies: critiques from regulatory and health technology assessment agencies. *Value Health.* 2022;25(12): 1967–1976.
- [21] Paul JE, Trueman P. 'Fourth hurdle reviews', NICE, and database applications. *Pharmacoepidemiol Drug Saf.* 2001;10(5): 429–438.
- [22] EUnetHTA. Joint action 3—WP1: A future model of HTA cooperation. 2021. Accessed: 2024-09-24. <https://www.eunetha.eu/wp-content/uploads/2021/09/FMC-HTA-WHITE-PAPER-FOR-PUBLICATION.pdf>.
- [23] Vreman RA, Mantel-Teeuwisse AK, Hövels AM, Leufkens HG, Goettsch WG. Differences in health technology assessment recommendations among European jurisdictions: the role of practice variations. *Value Health.* 2020;23(1): 10–16.
- [24] Sutton A, Ades A, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics.* 2008;26(9): 753–767.
- [25] Phillippo D, Ades T, Dias S, Palmer S, Abrams KR, Welton N. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE. NICE Decision Support Unit, 2016.
- [26] EUnetHTA. 21 methods guideline D4.3.2 direct and indirect comparison, V1.0. 2022. Accessed: 2024-09-24. <https://www.eunetha.eu/wp-content/uploads/2022/08/EUnetHTA-21-Deliverable-D4.3.2-Methodological-Guideline-on-Direct-and-indirect-comparisons-V1.0.pdf>.
- [27] EUnetHTA. 21—Individual practical guideline document D4.3.1: Direct and indirect comparisons, V 1.0. 2022. Accessed: 2024-09-24. <https://www.eunetha.eu/wp-content/uploads/2022/12/EUnetHTA-21-D4.3.1-Direct-and-indirect-comparisons-v1.0.pdf>.
- [28] Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Mak.* 2018;38(2): 200–211.
- [29] Goring S, Gustafson P, Liu Y, Saab S, Cline S, Platt R. Disconnected by design: analytic approach in treatment networks having no common comparator. *Res Synth Methods.* 2016;7(4): 420–432.
- [30] Stevens JW, Fletcher C, Downey G, Sutton A. A review of methods for comparing treatments evaluated in studies that form disconnected networks of evidence. *Res Synth Methods.* 2018;9(2): 148–162.

- [31] Serret-Larmande A, Zenati B, Dechartres A, Lambert J, Hajage D. A methodological review of population-adjusted indirect comparisons reveals inconsistent reporting and suggests publication bias. *J Clin Epidemiol.* 2023;163: 1–10.
- [32] Truong B, Tran LAT, Le TA, Pham TT, Vo TT. Population adjusted-indirect comparisons in health technology assessment: a methodological systematic review. *Res Synth Methods.* 2023;14(5): 660–670.
- [33] Lambert J, Lengliné E, Porcher R, Thiébaut R, Zohar S, Chevret S. Enriching single-arm clinical trials with external controls: possibilities and pitfalls. *Blood Adv.* 2023;7(19): 5680–5690.
- [34] Ren J, Cislo P, Cappelleri JC, Hlavacek P, DiBonaventura M. Comparing g-computation, propensity score-based weighting, and targeted maximum likelihood estimation for analyzing externally controlled trials with both measured and unmeasured confounders: a simulation study. *BMC Med Res Methodol.* 2023;23(1): 1–18.
- [35] Loiseau N, Trichelair P, He M, et al. External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning. *BMC Med Res Methodol.* 2022;22(1): 1–13.
- [36] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med.* 2004;23(19): 2937–2960.
- [37] Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med.* 2016;35(30): 5642–5655.
- [38] Hainmueller J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Polit Anal.* 2012;20(1): 25–46.
- [39] Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics.* 2010;28: 935–945.
- [40] Cheng D, Ayyagari R, Signorovitch J. The statistical performance of matching-adjusted indirect comparisons: estimating treatment effects with aggregate external control data. *Ann Appl Stat* 2020;14(4): 1806–1833.
- [41] Josey KP, Berkowitz SA, Ghosh D, Raghavan S. Transporting experimental results with entropy balancing. *Stat Med.* 2021;40(19): 4310–4326.
- [42] Amusa L, Zewotir T, North D. Examination of entropy balancing technique for estimating some standard measures of treatment effects: a simulation study. *Electron J Appl Stat Anal.* 2019;12(2): 491–507.
- [43] Cheng D, Tchetgen ET, Signorovitch J. On the double-robustness and semiparametric efficiency of matching-adjusted indirect comparisons. *Res Synth Methods.* 2023;14(3): 438–442.
- [44] Zhao Q, Percival D. Entropy balancing is doubly robust. *J Causal Infer.* 2017;5(1): 1–19.
- [45] Wang A, Nianogo RA, Arah OA. G-computation of average treatment effects on the treated and the untreated. *BMC Med Res Methodol* 2017;17(1): 3.
- [46] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model.* 1986;7(9–12): 1393–1512.
- [47] Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. *Epidemiology.* 2014;25(6): 889–897.
- [48] Remiro-Azócar A, Heath A, Baio G. Parametric G-computation for compatible indirect treatment comparisons with limited individual patient data. *Res Synth Methods.* 2022;13(6): 716–744.
- [49] Ishak KJ, Proskorovsky I, Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics.* 2015;33(6): 537–549.
- [50] Vo TT. A cautionary note on the use of G-computation in population adjustment. *Res Synth Methods.* 2023;14(3): 338–341.
- [51] Vanier A, Fernandez J, Kelley S, et al. Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for health. *BMJ Evid-Based Med.* 2024;29(1): 1–5.
- [52] Siu DH, Lin FP, Cho D, et al. Framework for the use of external controls to evaluate treatment outcomes in precision oncology trials. *JCO Precis Oncol.* 2024;8: e2300317.
- [53] Polito L, Liang Q, Pal N, et al. Applying the estimand and target trial frameworks to external control analyses using observational data: a case study in the solid tumor setting. *Front Pharmacol.* 2024;15: 1223858.
- [54] Zhou X, Zhu J, Drake C, Pang H. Causal estimators for incorporating external controls in randomized trials with longitudinal outcomes. *J Royal Stat Soc Ser A Stat Soc* 2025;188(3): 791–818.
- [55] Braunholtz DA, Edwards SJ, Lilford RJ. Are randomized clinical trials good for us (in the short term)? Evidence for a “trial effect”. *J Clin Epidemiol.* 2001;54(3): 217–224.
- [56] Dahabreh IJ, Robins JM, Haneuse SJ, Hernán MA. Generalizing causal inferences from randomized trials: counterfactual and graphical identification. Preprint, [arXiv:1906.10792](https://arxiv.org/abs/1906.10792), 2019.
- [57] Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics.* 2019;75(2): 685–694.
- [58] Buchanan AL, Hudgens MG, Cole SR, et al. Generalizing evidence from randomized trials using inverse probability of sampling weights. *J Royal Stat Soc Ser A Stat Soc.* 2018;181(4): 1193–1209.
- [59] Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol.* 2010;171(6): 674–677.
- [60] Glimm E, Yau L. Geometric approaches to assessing the numerical feasibility for conducting matching-adjusted indirect comparisons. *Pharm Stat.* 2022;21(5): 974–987.
- [61] Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernan MA. Extending inferences from a randomized trial to a new target population. *Stat Med.* 2020;39(14): 1999–2014.

- [62] Colnet B, Mayer I, Chen G, et al. Causal inference methods for combining randomized trials and observational studies: a review. *Stat Sci.* 2024;39(1): 165–191.
- [63] Josey KP, Yang F, Ghosh D, Raghavan S. A calibration approach to transportability and data-fusion with observational data. *Stat Med.* 2022;41(23): 4511–4531.
- [64] Busso M, DiNardo J, McCrary J. New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Rev Econ Stat.* 2014;96(5): 885–897.
- [65] Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11(5): 561–570.
- [66] Särndal CE, Swensson B, Wretman J. *Model Assisted Survey Sampling*. Springer Science & Business Media; 2003.
- [67] Chattopadhyay A, Hase CH, Zubizarreta JR. Balancing vs modeling approaches to weighting in practice. *Stat Med.* 2020;39(24): 3227–3254.
- [68] Filla T, Schwender H, Kuss O. Balancing versus modelling in weighted analysis of non-randomised studies with survival outcomes: a simulation study. *Stat Med.* 2024;43(17): 3140–3163.
- [69] Hirshberg DA, Zubizarreta JR. On two approaches to weighting in causal inference. *Epidemiology.* 2017;28(6): 812–816.
- [70] Wang J. On matching-adjusted indirect comparison and calibration estimation. Preprint, [arXiv:2107.11687](https://arxiv.org/abs/2107.11687), 2021.
- [71] Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika.* 2010;97(3): 661–682.
- [72] Phillippo DM, Dias S, Ades A, Welton NJ. Equivalence of entropy balancing and the method of moments for matching-adjusted indirect comparison. *Res Synth Methods.* 2020;11(4): 568–572.
- [73] Remiro-Azócar A. Two-stage matching-adjusted indirect comparison. *BMC Med Res Methodol.* 2022;22(1): 217.
- [74] Jackson D, Rhodes K, Ouwens M. Alternative weighting schemes when performing matching-adjusted indirect comparisons. *Res Synth Methods.* 2021;12(3): 333–346.
- [75] Jiang Z, Cappelleri JC, Gamalo M, Chen Y, Thomas N, Chu H. A comprehensive review and shiny application on the matching-adjusted indirect comparison. *Res Synth Methods.* 2024;15: 671–686.
- [76] Park JE, Campbell H, Towle K, et al. Unanchored population-adjusted indirect comparison methods for time-to-event outcomes using inverse odds weighting, regression adjustment, and doubly robust methods with either individual patient or aggregate data. *Value Health.* 2024;27(3): 278–286.
- [77] Jiang Y, Ni W. Performance of unanchored matching-adjusted indirect comparison (MAIC) for the evidence synthesis of single-arm trials with time-to-event outcomes. *BMC Med Res Methodol.* 2020;20: 1–9.
- [78] Hatswell AJ, Freemantle N, Baio G. The effects of model misspecification in unanchored matching-adjusted indirect comparison: results of a simulation study. *Value Health.* 2020;23(6): 751–759.
- [79] Chandler CO, Proskorovsky I. Uncertain about uncertainty in matching-adjusted indirect comparisons? A simulation study to compare methods for variance estimation. *Res Synth Methods.* 2024;15(6): 1094–1110.
- [80] Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. *J Am Stat Assoc.* 2015;110(511): 910–922.
- [81] Kern HL, Stuart EA, Hill J, Green DP. Assessing methods for generalizing experimental impact estimates to target populations. *J Res Educ Effect.* 2016;9(1): 103–127.
- [82] Remiro-Azócar A. Transportability of model-based estimands in evidence synthesis. *Stat Med.* 2024;43(22): 4217–4249.
- [83] Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol.* 2011;173(7): 761–767.
- [84] Shinozaki T, Matsuyama Y. Brief report: doubly robust estimation of standardized risk difference and ratio in the exposed population. *Epidemiology.* 2015;26(6): 873–877.
- [85] Benkeser D, Carone M, Laan MVD, Gilbert PB. Doubly robust nonparametric inference on the average treatment effect. *Biometrika.* 2017;104(4): 863–880.
- [86] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when “inverse probability” weights are highly variable. *Stat Sci.* 2007;22(4): 544–559.
- [87] Kang JD, 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.
- [88] Gabriel EE, Sachs MC, Martinussen T, et al. Inverse probability of treatment weighting with generalized linear outcome models for doubly robust estimation. *Stat Med.* 2024;43(3): 534–547.
- [89] Wooldridge JM. Inverse probability weighted estimation for general missing data problems. *J Econ.* 2007;141(2): 1281–1301.
- [90] Gabriel EE, Sachs MC, Waernbaum I, et al. Propensity weighting plus adjustment in proportional hazards model is not doubly robust. *Biometrics.* 2024;80(3): ujae069.
- [91] Reifeis SA, Hudgens MG. On variance of the treatment effect in the treated when estimated by inverse probability weighting. *Am J Epidemiol.* 2022;191(6): 1092–1097.
- [92] Kostouraki A, Hajage D, Rachet B, et al. On variance estimation of the inverse probability-of-treatment weighting estimator: a tutorial for different types of propensity score weights. *Stat Med.* 2024;43(13): 2672–2694.
- [93] Ren S, Ren S, Welton NJ, Strong M. Advancing unanchored simulated treatment comparisons: a novel implementation and simulation study. *Res Synth Methods.* 2024;15(4): 657–670.
- [94] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med.* 2019;38(11): 2074–2102.

- [95] R Core Team. R: a language and environment for statistical computing, version R-3.4. 1. Vienna, Austria: R Foundation for Statistical Computing. software, 2017.
- [96] Morgan BJ, Smith D. A note on Wadley's problem with overdispersion. *J Royal Stat Soc Ser C Appl Stat.* 1992;41(2): 349–354.
- [97] Bennett I. Roche/MAIC: v0.3.0—Validate release. 2022. <https://doi.org/10.5281/zenodo.6624152>.
- [98] Yau L. maicChecks v0.2.0. 2025. <https://doi.org/10.1002/pst.2210>.
- [99] Zhang L, Bujkiewicz S, Jackson D. Four alternative methodologies for simulated treatment comparison: how could the use of simulation be re-invigorated? *Res Synth Methods.* 2024;15(2): 227–241.
- [100] Phillippo DM. Multinma: a comprehensive R package for Bayesian network meta-analysis with aggregate data, individual patient data, or a mixture of both. Abstract from Society for Research Synthesis Methodology Annual Meeting 2024, Amsterdam, Netherlands.
- [101] Webster-Clark M, Stürmer T, Wang T, et al. Using propensity scores to estimate effects of treatment initiation decisions: state of the science. *Stat Med.* 2021;40(7): 1718–1735.
- [102] Zetterqvist J, Vermeulen K, Vansteelandt S, Sjölander A. Doubly robust conditional logistic regression. *Stat Med.* 2019;38(23): 4749–4760.
- [103] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Double-robust and efficient methods for estimating the causal effects of a binary treatment. Preprint. [arXiv:2008.00507](https://arxiv.org/abs/2008.00507), 2020.
- [104] Gupta A, Hsu G, Kent S, et al. Quantitative bias analysis for single-arm trials with external control arms. *JAMA Netw Open.* 2025;8(3): e252152.

Appendix

Consistency of weighting estimators

To see that the IOW estimators are consistent under a correctly specified propensity score model for data source assignment, consider a simple scenario with a binary outcome, Y , and a single discrete covariate, X , such that:

$$\begin{aligned}\mu_0^1 &= E(Y|T = 1, S = 0) = 1 \times \Pr(Y = 1|T = 1, S = 0) + 0 \times \Pr(Y = 0|T = 1, S = 0) \\ &= \Pr(Y = 1|T = 1, S = 0).\end{aligned}$$

Basic probability rules imply that the marginal risk is the weighted average of the stratum-specific risks:

$$= \sum_x \Pr(Y = 1|T = 1, S = 0, X = x) \Pr(X = x|S = 0, T = 1).$$

Then, due to the assumption of conditional data source ignorability, we have

$$\begin{aligned}&= \sum_x \Pr(Y = 1|T = 1, X = x) \Pr(X = x|S = 0) \\ &= \sum_x \Pr(Y = 1|T = 1, X = x) \Pr(X = x|S = 0) \frac{\Pr(X = x|S = 1)}{\Pr(X = x|S = 1)} \\ &= \sum_x \Pr(Y = 1|T = 1, X = x) \Pr(X = x|S = 1) \frac{\Pr(X = x|S = 0)}{\Pr(X = x|S = 1)},\end{aligned}$$

and from Bayes' rule, we have

$$\begin{aligned}&= \sum_x \Pr(Y = 1|T = 1, X = x) \Pr(X = x|S = 1) \frac{\Pr(S = 0|X = x)}{\Pr(S = 1|X = x)} \frac{\Pr(S = 1)}{\Pr(S = 0)} \\ &= \underbrace{\frac{\Pr(S = 1)}{\Pr(S = 0)}}_A \underbrace{\sum_x \Pr(Y = 1|T = 1, X = x) \Pr(X = x|S = 1) \frac{\Pr(S = 0|X = x)}{\Pr(S = 1|X = x)}}_B.\end{aligned}\tag{A.1}$$

Since $\sum_{i=1}^n \frac{S_i}{n} = n_1/n \rightarrow \Pr(S = 1)$ and $\sum_{i=1}^n (1 - S_i)/n = n_0/n \rightarrow \Pr(S = 0)$, we can consistently estimate $A \approx n_1/n_0$. We can also consistently estimate B from the sample since the covariate

distribution for the SAT, $(X|S = 1)$, is observed (and since, within the sample, $T = 1 \iff S = 1$):

$$B \approx \sum_{i=1}^{n_1} \Pr(Y = 1|S = 1, X = x_i) \frac{\Pr(S = 0|X = x_i)}{\Pr(S = 1|X = x_i)} \approx \sum_{i=1}^{n_1} \frac{Y_i}{n_1} \frac{\Pr(S = 0|X = x_i)}{\Pr(S = 1|X = x_i)}.$$

Finally, if the propensity score model is correctly specified, the inverse odds weights consistently estimate the true inverse odds, such that, for i in $1, \dots, n_1$:

$$\widehat{w}_i \rightarrow \frac{\Pr(S = 0|X = x_i)}{\Pr(S = 1|X = x_i)}. \quad (\text{A.2})$$

Therefore, the IOW estimator is consistent:

$$\begin{aligned} \widehat{\mu}_0^1 &= \underbrace{\frac{n_1}{n_0}}_A \underbrace{\sum_{i=1}^{n_1} \frac{Y_i}{n_1} \widehat{w}_i}_B \\ &= \frac{1}{n_0} \sum_{i=1}^{n_1} Y_i \widehat{w}_i \rightarrow \mu_0^1. \end{aligned} \quad (\text{A.3})$$

Note that $E(S_i w_i) = n_0/n_1$ implies that $E\left(\frac{1}{n_1} \sum_{i=1}^{n_1} w_i\right) = n_0/n_1$, which implies that $E\left(\sum_{i=1}^{n_1} w_i\right) = n_0$. As such, the normalized IOW estimator in Equation (4) is also consistent for the ATC.

While the entropy balancing MAIC weights ($\widehat{v}_i$) defined in Equation (11) will be different than the IOW weights ($\widehat{w}_i$) obtained from maximum-likelihood estimation of the logistic regression model in Equation (2), $\widehat{v}_i = n_0 \widehat{w}_i$ will consistently estimate the true inverse odds if the logistic regression model is correctly specified. To be clear, if the logistic regression model correctly specifies the true propensity score model, then we have both: $\widehat{w}_i \rightarrow \frac{\Pr(S=0|X=x_i)}{\Pr(S=1|X=x_i)}$ (i.e., the IOW weights are consistent) and $\widehat{v}_i \rightarrow \frac{\Pr(S=0|X=x_i)}{\Pr(S=1|X=x_i)}$ (i.e., the entropy balancing weights are consistent), for i in $1, \dots, n_1$ (see Zhao and Percival for details). Therefore, following the same logic as detailed in Section 3.2 for the IOW estimators, the MAIC estimator, as defined in Equation (12), is also consistent if the implied propensity score model is correctly specified.

Double robustness of the augmented weighting estimators

Note that the following derivation does not depend on using the canonical link function for the outcome model. Consider the simple scenario where we have a binary outcome, Y , and a single discrete covariate, X , and the augmented weighting estimator in Equation (21) is re-written as

$$\widehat{\mu}_0^1 = \sum_{i=1}^{n_1} \widehat{v}_i (Y_i - \widehat{Y}_i^1) + \frac{1}{n_0} \sum_{i=n_1+1}^n \widehat{Y}_i^1. \quad (\text{A.4})$$

If the outcome model is correctly specified, the expectation of the first summation in Equation (A.4) converges to zero as $n_1 \rightarrow \infty$ because $\widehat{Y}_i^1 \rightarrow Y_i$ and the terms inside the summation cancel out, irrespective of any postulated propensity score model. The second summation is equivalent to the G-computation estimator and is consistent for μ_0^1 because the outcome model is correct. Consequently, $\widehat{\mu}_0^1 \rightarrow \mu_0^1$, and $\widehat{\text{ATC}} \rightarrow \text{ATC}$ (assuming $\widehat{\mu}_0^0 \rightarrow \mu_0^0$).

Conversely, if the propensity score model is correctly specified but the outcome model is incorrect, the first summation consistently cancels out the bias produced by the G-computation estimator in the second summation and the remainder term is exactly equal to the non-augmented weighting estimator,

which converges to μ_0^1 as $n_1 \rightarrow \infty$ because the propensity score model is correct. To illustrate this, consider rearranging Equation (A.4) to

$$\begin{aligned}\hat{\mu}_0^1 &= \sum_{i=1}^{n_1} \hat{v}_i Y_i - \sum_{i=1}^{n_1} \hat{v}_i \hat{Y}_i^1 + \frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1 \\ &= \underbrace{\sum_{i=1}^{n_1} \hat{v}_i Y_i}_C + \underbrace{\left(\frac{1}{n_0} \sum_{i=n_1+1}^n \hat{Y}_i^1 - \sum_{i=1}^{n_1} \hat{v}_i \hat{Y}_i^1 \right)}_D.\end{aligned}\quad (\text{A.5})$$

First, C is equivalent to the MAIC estimator and is consistent for μ_0^1 because the propensity score model is correct. Expanding the summations in D over $i = 1, \dots, n$, we have

$$D = \left(\sum_{i=1}^n \frac{(1 - S_i) \hat{Y}_i^1}{n_0} - \frac{S_i \hat{v}_i \hat{Y}_i^1}{n_0} \right) \quad (\text{A.6})$$

$$= \frac{1}{n_0} \left(\sum_{i=1}^n \hat{Y}_i^1 ((1 - S_i) - S_i \hat{v}_i) \right), \quad (\text{A.7})$$

where $\hat{v}_i = n_0 \hat{v}_i$, for i in $1, \dots, n$. Then, suppose that the outcome model for Y^1 is independent of S and that Y_i^{1*} is the large sample limit of $\hat{Y}_i^1$. Also, suppose that the large sample limit of $\hat{v}_i$ is $\frac{\Pr(S=0|X=x_i)}{\Pr(S=1|X=x_i)}$, since the propensity score model is correctly specified. Then:

$$\begin{aligned}\mathbb{E} \left(\hat{Y}_i^1 ((1 - S_i) - S_i \hat{v}_i) \right) &= \mathbb{E} \left(\mathbb{E} \left(\hat{Y}_i^1 ((1 - S_i) - S_i \hat{v}_i) \mid X \right) \right) \\ &= \mathbb{E} \left(\mathbb{E}(\hat{Y}_i^1 \mid X) \times \mathbb{E}(((1 - S_i) - S_i \hat{v}_i) \mid X) \right) \\ &\rightarrow \mathbb{E} \left(Y_i^{1*} \sum_x \Pr(S = 0|X = x) \Pr(X = x) - \Pr(S = 1|X = x) \Pr(X = x) \right. \\ &\quad \left. \frac{\Pr(S = 0|X = x)}{\Pr(S = 1|X = x)} \right) \\ &= \mathbb{E} \left(Y_i^{1*} \sum_x \Pr(S = 0|X = x) \Pr(X = x) - \Pr(X = x) \Pr(S = 0|X = x) \right) \\ &= \mathbb{E} \left(Y_i^{1*} \times 0 \right) \\ &= 0.\end{aligned}\quad (\text{A.8})$$

Consequently, $D \rightarrow 0$ which implies that $\hat{\mu}_0^1 \rightarrow \mu_0^1$, and $\widehat{\text{ATC}} \rightarrow \text{ATC}$ having assumed $\hat{\mu}_0^0 \rightarrow \mu_0^0$.