A common framework for externally-controlled single-arm trials and unanchored comparisons

ENTROPY BALANCING AND AUGMENTED WEIGHTING ESTIMATORS

Antonio Remiro-Azócar, PhD (joint work with Harlan Campbell from UBC & Precision AQ)

Methods and Outreach, Novo Nordisk

ASA BIOP Webinar Series

20th June 2025

Agenda

1. Context and estimands
2. Estimation of the average treatment effect in the control (ATC)
3. Methodological extensions
4. Simulation study
5. Concluding remarks

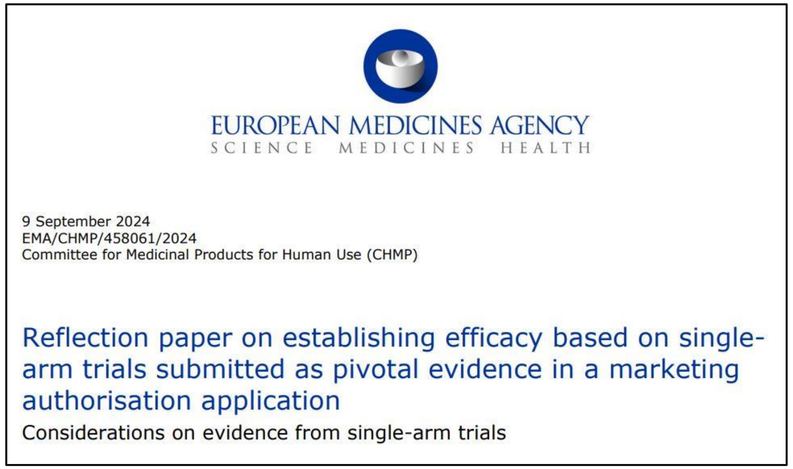
Context and estimands

Single-arm trials (SATs)

Conducting RCTs may not be possible:

- Where recruitment to RCTs is unfeasible due to small populations, e.g., rare diseases with orphan designation
- For life-threatening conditions with high unmet need and no standard of care, e.g., last-line of therapy in solid tumor oncology
- Where enrolling patients to placebo is unethical, e.g., pediatric trials for treatments with proven efficacy in adults

Regulators recognize that externally controlled SATs might be required in these special circumstances

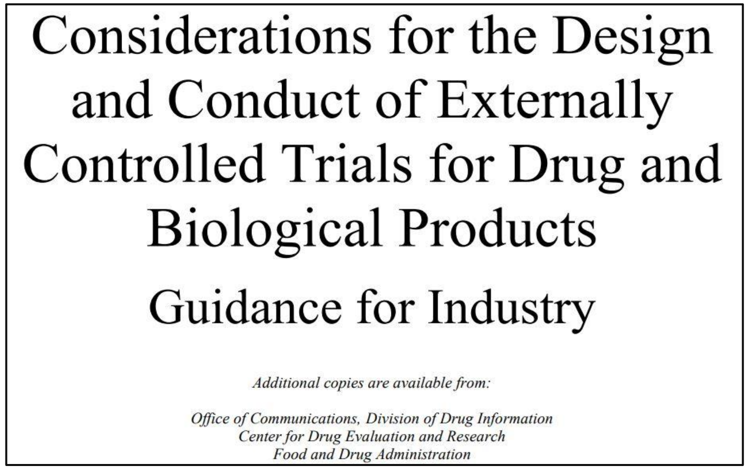


EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

9 September 2024
EMA/CHMP/458061/2024
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

Considerations on evidence from single-arm trials

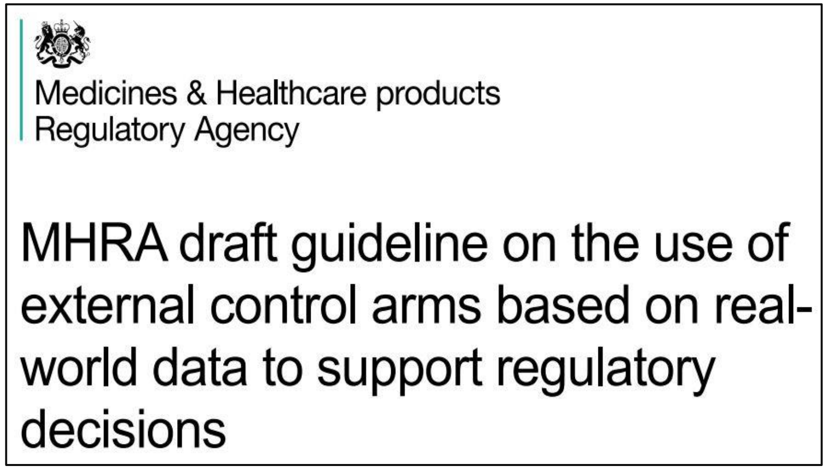


Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products

Guidance for Industry

Additional copies are available from:

Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration



Medicines & Healthcare products
Regulatory Agency

MHRA draft guideline on the use of external control arms based on real-world data to support regulatory decisions

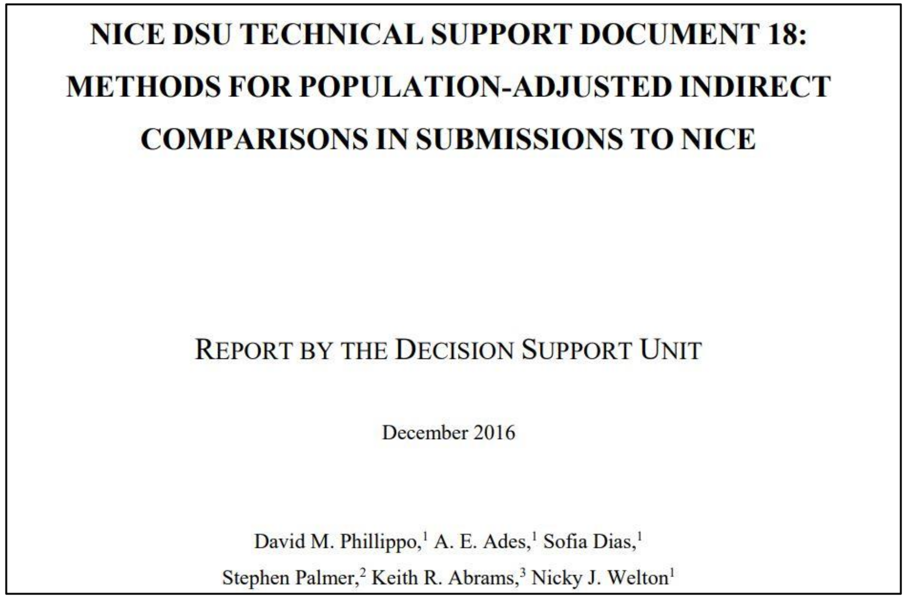
Marketing authorization applications featuring externally controlled SATs continue to rise

Unanchored indirect treatment comparisons (ITCs)

Health technology assessment (HTA) requires comparing the clinical and cost-effectiveness of new health technologies against all existing alternatives on the market:

- The scope of assessments depends on the policy question and is not always driven by the available data
- RCTs cannot have all desired treatment arms, given the number of jurisdictions and variations in clinical practice
- It is not always possible to find compatible control arms to “anchor” indirect comparisons, e.g., in rapidly evolving therapeutic areas with a changing comparator landscape and no single accepted standard of care

Unanchored indirect treatment comparisons may be required in these circumstances



NICE DSU TECHNICAL SUPPORT DOCUMENT 18:
METHODS FOR POPULATION-ADJUSTED INDIRECT COMPARISONS IN SUBMISSIONS TO NICE

REPORT BY THE DECISION SUPPORT UNIT

December 2016

David M. Phillippo,[1] A. E. Ades,[1] Sofia Dias,[1]
Stephen Palmer,[2] Keith R. Abrams,[3] Nicky J. Welton[1]

Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons

Adopted on 8 March 2024 by the HTA CG pursuant to Article 3(7), point (d), of Regulation (EU) 2021/2282 on Health Technology Assessment

A common framework

Unanchored ITCs are externally controlled SATs with two “special” characteristics:

- The external control is often a competitor-sponsored historical trial
- Limited access to subject-level data for the external control, only aggregate-level data from publications

Nevertheless, different **estimands** or **summary measures** can be targeted:

- **Average treatment effect (ATE)** among the combined SAT and external control…somewhat ambiguous here

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

- **Average treatment effect in the treated (ATT)** among those participating in the SAT

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

- **Average treatment effect in the control (ATC)** among those in the external control group

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

ATT or ATC?

Difference between summary measures driven by them targeting different (sub) populations or “analysis sets”

Average treatment effect in the treated (ATT)

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

- Would be the primary estimand for drug approval purposes in the regulatory environment
- 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
- Compatible with the mean absolute outcome targeted by the SAT, preserving the original SAT results
- Potentially unappealing where generalizability to routine clinical practice is a priority: SAT populations are often highly selected and may lack representativeness with respect to “real-world” populations

ATT or ATC?

Average treatment effect in the control (ATC)

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

- Often the target estimand in HTA: due to necessity as subject-level data may be unavailable for the external control
- Potentially more desirable for external validity, e.g., external controls based on RWD or natural history studies, with broad inclusion criteria and heterogeneous populations
- But not necessarily so…historical controls from past clinical trials may not reflect the current standard of care, RWD-derived external controls based on a single country are not necessarily transferable across jurisdictions, etc.

Statistical considerations (sample size, precision) may also play a role in the estimand choice, e.g., when weighting

We shall assume that there is access to subject-level data but that the ATC is the primary target of estimation

"In any situation with non-randomized data, such as observational evidence and single-arm trials, (…) complete access to the IPD is required" (Methodological Guideline for Quantitative Evidence Synthesis, HTA CG)

Estimation of the ATC

Available methodology

EXTERNALLY-CONTROLLED SATs

- Odds weighting: modeling-based approach to weighting
- Outcome modeling: G-computation
- Doubly robust (DR) methods are well established: augmented approaches, TMLE, etc.
- Use of data-adaptive (machine learning) estimators has been explored within a DR framework
- Methodologies assume full access to subject-level data
- Target is typically the ATT

UNANCHORED ITCS (HTA)

- Matching-adjusted indirect comparison: balancing-based approach to weighting
- Outcome modeling: Simulated treatment comparison
- Doubly robust augmented approaches, TMLE, yet to be leveraged
- Reliance on the correct specification of a single parametric model
- Methodologies developed under limited access to subject-level data
- Target is typically the ATC

Weighting: modeling versus balancing

MODELING

- Explicitly models the propensity score as a function of baseline covariates
- Propensity score/estimated by maximizing the fit of a logistic regression
- Weights may not produce adequate balance, e.g., if propensity score model is mis-specified
- Propensity score predictions that are close to zero produce extreme weights, which lead to imprecision
- Limited applicability with unavailable subject-level covariates for the external control

ENTROPY BALANCING

- Does not explicitly model the propensity score, but implicitly assumes a logistic propensity score model
- Covariate balance viewed as a convex optimization problem
- Less susceptible to bias by directly enforcing covariate balance
- Minimally dispersed weights, which translates into larger effective sample sizes and precision
- Applicable where aggregate-level covariate moments are available for the external control

(Inverse) odds weighting

Westreich et al (2017), Dahabreh et al (2020)

”modeling” approach to weighting

› DESCRIPTION

- Models the data source assignment mechanism, conditional on covariates, to estimate weights
- SAT subjects weighted by their inverse conditional odds of SAT participation – conditional odds of external control participation – to transport SAT outcomes to the external control (sub) population
- Inverse odds weights defined as:

$$w_i = \frac{(1 - e_i)S_i}{e_i} + (1 - S_i), \text{ with } \; e_i = e(\mathbf{X}_i) = \Pr(S_i = 1 \mid \mathbf{X}_i)$$

- A logistic regression is fitted to the concatenated SAT and external control subject-level data, using maximum-likelihood estimation to estimate model-based propensity scores

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

$$\hat{e}_i = \text{logit}^{-1}\left(\hat{\alpha}_0 + \mathbf{c}(\mathbf{X}_i)^\top \hat{\boldsymbol{\alpha}}\right)$$

- Propensity score predictions plugged into weight equation to derive weight estimates
- ATC estimated by contrasting the weighted average of observed outcomes under the active intervention with the unweighted average of observed outcomes for the external control

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

› MODELING ASSUMPTION

- Correct specification of propensity score model
- Logit of the propensity score (conditional probability of SAT participation) assumed to vary linearly with the covariate balance functions

› RECOMMENDATIONS

- Normalize/stabilize the weights so that they sum to one

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

- This bounds the mean absolute outcome estimate for the active intervention within its feasible range
- Should improve finite sample properties and provide more stable and precise estimation

› LIMITATIONS

- Estimated weights do not produce adequate covariate balance if the propensity score model is mis-specified
- Even a correctly specified model does not guarantee balance in finite simples
- Propensity score predictions that are close to zero produce extreme and highly variable weights, which may lead to unstable and imprecise ATC estimation
- Limited applicability with unavailable subject-level covariates for the external control

Entropy balancing (matching-adjusted indirect comparison)

Signorovitch et al (2010), Josey et al (2021)

"balancing" or "calibration" approach to weighting

› MOTIVATION

- Less susceptible to bias by directly **enforcing covariate balance**
- **Minimally dispersed weights**, which translates into larger effective sample sizes and precision
- Applicable when **aggregate-level marginal covariate moments** available for the external control
- "**Linear double robustness**": consistent under two distinct underlying data-generating models
- Weights **constrained to be positive**, sample-boundeness (interpolation as opposed to extrapolation)

› DESCRIPTION

- Propensity score is not explicitly modeled, but logistic model for data source assignment is assumed

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

- Weights proportional to the inverse conditional odds of SAT participation

- "Method of moments" to estimate the model while enforcing covariate balance constraint

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

$$\mathbf{c}^*(\mathbf{X}_i) = \mathbf{c}(\mathbf{X}_i) - \hat{\boldsymbol{\theta}}$$

- Solve by minimizing objective function using convex optimization algorithm $Q(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n_1} \exp\left(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}\right)$

- Weights for SAT estimated as

$$\hat{\omega}_i = \frac{\exp\left(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}\right)}{\sum_{i=1}^{n_1} \exp\left(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}\right)}$$

- ATC estimated as

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

› MODELING ASSUMPTION

- **Linear double robustness** with linear outcome regression and propensity score logistic regression
- Logit of conditional probability of external control participation (or SAT participation) **or** conditional outcome expectation under the active intervention varies linearly with the covariate balance functions

- For instance, mean-balancing ensures consistency if:

$$\mathrm{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \boldsymbol{\alpha} \quad \text{OR} \quad \mathrm{E}(Y_i^1 \mid \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}$$

- Mean- and variance-balancing ensures consistency if:

$$\mathrm{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \boldsymbol{\alpha}_1 + (\mathbf{X}_i^2)^\top \boldsymbol{\alpha}_2 \quad \text{OR} \quad \mathrm{E}(Y_i^1 \mid \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}_1 + (\mathbf{X}_i^2)^\top \boldsymbol{\beta}_2$$

› LIMITATIONS

Standard balancing strategies do not enable to conjecture an implicit outcome model that is flexible enough for DR:

- As the number of balancing constraints increases, it becomes less likely to find a feasible weighting solution
- As the number of balancing constraints increases, effective sample size and precision drop
- Aggregate data on anything other than means and variances are not typically reported

Is entropy balancing doubly robust?

Doubly robust “augmented” methods have rarely been applied to unanchored ITCs, despite being preferred by decision-makers and their development being recommended by HTA agencies

One reason might be a misunderstanding that MAIC, the most popular approach, is always doubly robust

MAIC (entropy balancing) is **linear doubly robust**:

- “Doubly robust with respect to linear outcome regression and logistic propensity score regression” (Zhao and Percival, 2017)
- Consistent if logit of conditional probability of external control participation (or SAT participation) **or** conditional outcome expectation under the active intervention varies linearly with the covariate balance functions
- For instance, mean-balancing ensures consistency if $\operatorname{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \boldsymbol{\alpha}$ OR $\mathrm{E}(Y_i^1 \mid \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}$
- Mean- and variance-balancing ensure consistency if $\operatorname{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \boldsymbol{\alpha}_1 + (\mathbf{X}_i^2)^\top \boldsymbol{\alpha}_2$ OR $\mathrm{E}(Y_i^1 \mid \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}_1 + (\mathbf{X}_i^2)^\top \boldsymbol{\beta}_2$

However, it is rarely plausible that outcomes vary linearly with the covariates (e.g., due to non-linear link functions or outcomes that depend on other non-linear transformations of the covariates)

The need for augmentation

Suggested balancing strategies are insufficient for consistency with more complex underlying outcome models

One could consider balancing other non-linear covariate transformations and interactions, but this is rarely feasible:

- As the number of balancing constraints increases, it is more likely that the covariate moments fall outside the convex hull of the observed covariate space
- 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
- Increasing the number of balancing conditions leads to further reductions in effective sample size and precision
- Aggregate data on higher order moments and transformed covariate means is rarely reported

This motivates the explicit augmentation of the weighting estimators, allowing for a less restrictive outcome model

Augmented entropy balancing

Campbell and Remiro-Azócar (2025)

- Postulate a model for the conditional outcome expectation under the active intervention and fit it to the SAT

$$q\left(\mathrm{E}(Y_i^1 \mid \mathbf{X}_i; \boldsymbol{\beta})\right) = m\left(\mathbf{X}_i; \boldsymbol{\beta}\right)$$

- Predict potential outcomes for the active intervention for all subjects in the SAT and the external control

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

- The G-computation estimator is augmented with a weighted average of residuals, but using **entropy balancing weights**; the weighted average is the “one-step” correction term for the potential bias of G-computation

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

- Estimator for the ATC:

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

- Variance estimation: non-parametric bootstrap, resampling with replacement the concatenated subject-level data

Augmented entropy balancing

Reasons for augmentation using the entropy balancing weights:

- Augmented estimators inherit attractive properties: lower susceptibility to bias by directly enforcing balance
- Greater weight stability, translating into larger effective sample sizes after weighting and enhanced precision
- Consistency under a greater number of distinct underlying data-generating mechanisms

What about “weighted G-computation”?

- Another augmented estimator claimed to be doubly robust consists of G-computation based on the predictions of a weighted outcome model

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

- Note: this is only DR where the outcome model is a GLM with canonical link function! (Gabriel et al 2024)
- Results suggest asymptotic equivalence and similar finite-sample performance to the augmented weighting estimators previously described **for GLMs with canonical link functions** (Gabriel et al 2024, Sloczynski et al 2023)

Methodological extensions

Unavailable subject-level data for the control

Campbell and Remiro-Azócar (2025)

› PRELIMINARY STEP (all methods except MAIC)

- *M* individual-level covariate profiles simulated from the assumed covariate distribution of the external control based on published summary statistics
- Number of hypothetical subject profiles should be relatively large (e.g., *M*=1000) to minimize sampling variability and random seed sensitivity
- Necessary information to infer the joint covariate distribution of the external control, e.g., distributional forms and correlation structures, is rarely published
- This must be borrowed from other data sources or selected based on theoretical properties, following recommendations in the literature
- Stack SAT subject-level covariate data with simulated subject-level covariate data for the external control

› WEIGHTING

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

› DR AUGMENTED WEIGHTING

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

› G-COMPUTATION

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

› WEIGHTED G-COMPUTATION

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

› VARIANCE ESTIMATION

- Some changes to the non-parametric bootstrap procedure
- Only bootstrap the SAT
- Assumes that mean absolute outcomes are statistically independent (overconservativeness)
- Assumes the external control covariate distributional data are fixed, potentially unreasonable with small sample sizes for the external control (overprecision)

$$\mathrm{SE}\left(\widehat{\mathrm{ATC}}\right) = \sqrt{\left(\mathrm{SE}\left(g(\hat{\mu}_0^1)\right)\right)^2 + \left(\mathrm{SE}\left(g(\hat{\mu}_0^0)\right)\right)^2}$$

Targeting the ATT

Campbell and Remiro-Azócar (2025)

› WEIGHTING

- External control subjects weighted by their conditional odds of SAT participation
- Objective is to balance the external control covariate distribution with respect to that of the SAT
- General form of estimators:

$$\widehat{\mathrm{ATT}} = \underbrace{g\left(\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i\right)}_{\hat{\mu}_1^1} - \underbrace{g\left(\frac{1}{K}\sum_{i=n_1+1}^{n} \hat{\upsilon}_i Y_i\right)}_{\hat{\mu}_1^0}$$

› G-COMPUTATION

- Model for the conditional outcome expectation postulated under the control, not under the active intervention
- Potential outcome under the control predicted for all SAT subjects

$$\widehat{\mathrm{ATT}} = \underbrace{g\left(\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i\right)}_{\hat{\mu}_1^1} - \underbrace{g\left(\frac{1}{n_1}\sum_{i=1}^{n_1} \hat{Y}_i^0\right)}_{\hat{\mu}_1^0}, \qquad \hat{Y}_i^0 = q^{-1}\left(m(\mathbf{X}_i; \hat{\boldsymbol{\beta}})\right)$$

› DR AUGMENTED WEIGHTING

- Model for the conditional outcome expectation fitted to the external control
- Potential outcomes under the control predicted for all SAT and external control subjects

$$\hat{Y}_i^0 = q^{-1}\left(m(\mathbf{X}_i; \hat{\boldsymbol{\beta}})\right)$$

- The potential outcome predictions are augmented with a weighted average of residuals for the external control subjects
- General form of estimators:

$$\widehat{\mathrm{ATT}} = \underbrace{g\left(\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i\right)}_{\hat{\mu}_1^1} - \underbrace{g\left(\frac{1}{K}\sum_{i=n_1+1}^{n} \hat{\upsilon}_i \epsilon_i^0 + \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{Y}_i^0\right)}_{\hat{\mu}_1^0}, \qquad \epsilon_i^0 = Y_i - \hat{Y}_i^0$$

› WEIGHTED G-COMPUTATION

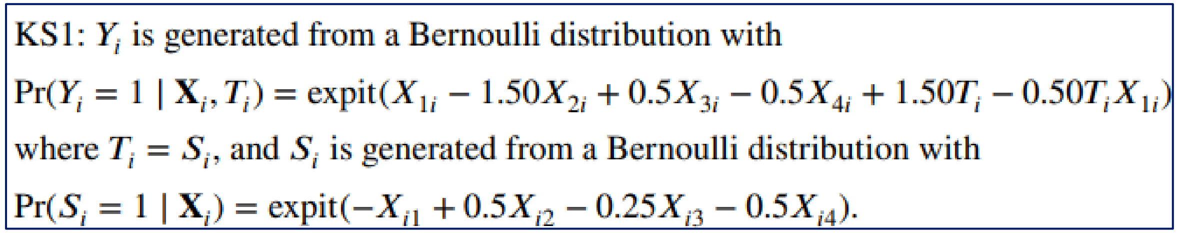
- Estimate weights for the odds of SAT participation, fit a weighted model for the conditional outcome expectation to the external controls, and average outcome predictions of the weighted regression under the SAT covariate distribution

$$\widehat{\mathrm{ATT}} = \underbrace{g\left(\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i\right)}_{\hat{\mu}_1^1} - \underbrace{g\left(\frac{1}{n_1}\sum_{i=1}^{n_1} \hat{Y}_i^0\right)}_{\hat{\mu}_1^0}, \quad \hat{\mu}_1^0 = \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} \hat{Y}_i^0 = \tfrac{1}{n_1}\sum_{i=1}^{n_1} q^{-1}\left(m(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_\upsilon)\right)$$

Simulation study

Data-generating mechanisms

KS1: propensity score and outcome model correctly specified

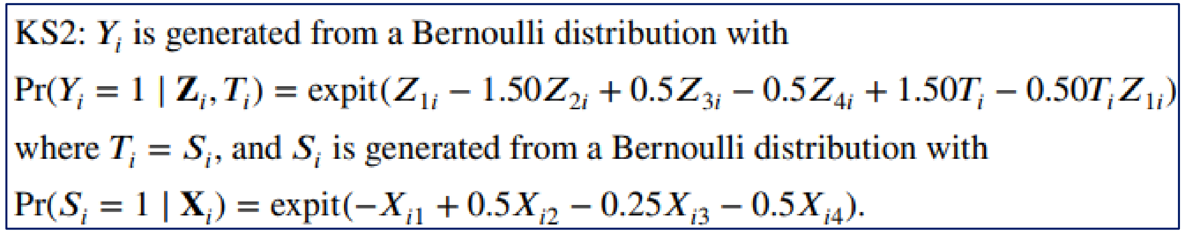

KS1: Y_i is generated from a Bernoulli distribution with

$$\Pr(Y_i = 1 \mid \mathbf{X}_i, T_i) = \text{expit}(X_{1i} - 1.50X_{2i} + 0.5X_{3i} - 0.5X_{4i} + 1.50T_i - 0.50T_i X_{1i})$$

where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with

$$\Pr(S_i = 1 \mid \mathbf{X}_i) = \text{expit}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.5X_{i4}).$$

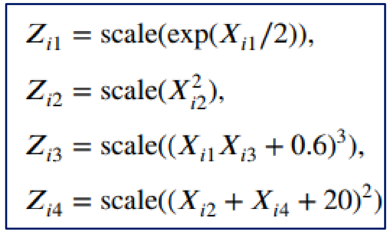
KS2: only propensity score model correctly specified


KS2: Y_i is generated from a Bernoulli distribution with

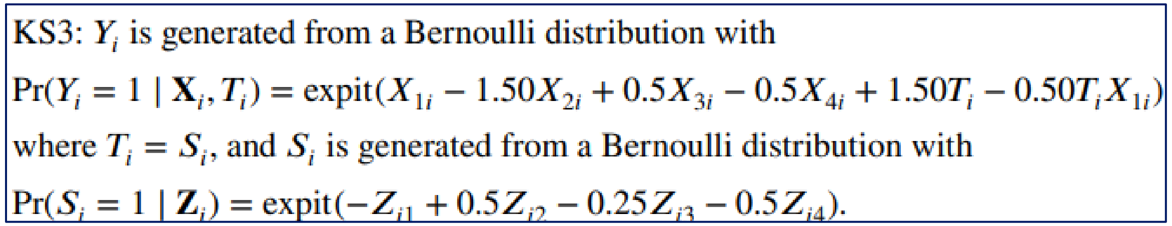
$$\Pr(Y_i = 1 \mid \mathbf{Z}_i, T_i) = \text{expit}(Z_{1i} - 1.50Z_{2i} + 0.5Z_{3i} - 0.5Z_{4i} + 1.50T_i - 0.50T_i Z_{1i})$$

where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with

$$\Pr(S_i = 1 \mid \mathbf{X}_i) = \text{expit}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.5X_{i4}).$$


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

KS3: only outcome model correctly specified

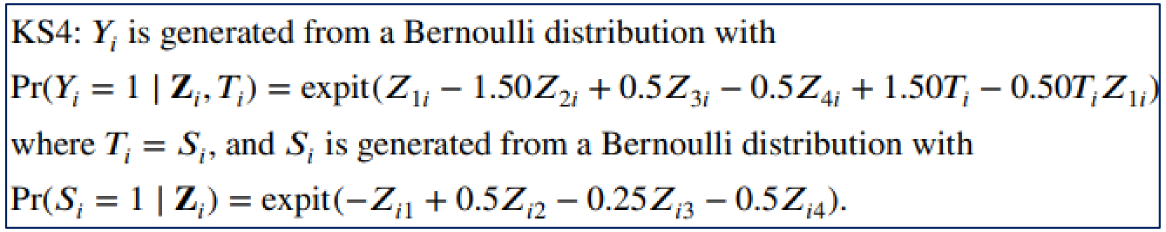

KS3: Y_i is generated from a Bernoulli distribution with

$$\Pr(Y_i = 1 \mid \mathbf{X}_i, T_i) = \text{expit}(X_{1i} - 1.50X_{2i} + 0.5X_{3i} - 0.5X_{4i} + 1.50T_i - 0.50T_i X_{1i})$$

where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with

$$\Pr(S_i = 1 \mid \mathbf{Z}_i) = \text{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.5Z_{i4}).$$

KS4: propensity score and outcome model incorrectly specified


KS4: Y_i is generated from a Bernoulli distribution with

$$\Pr(Y_i = 1 \mid \mathbf{Z}_i, T_i) = \text{expit}(Z_{1i} - 1.50Z_{2i} + 0.5Z_{3i} - 0.5Z_{4i} + 1.50T_i - 0.50T_i Z_{1i})$$

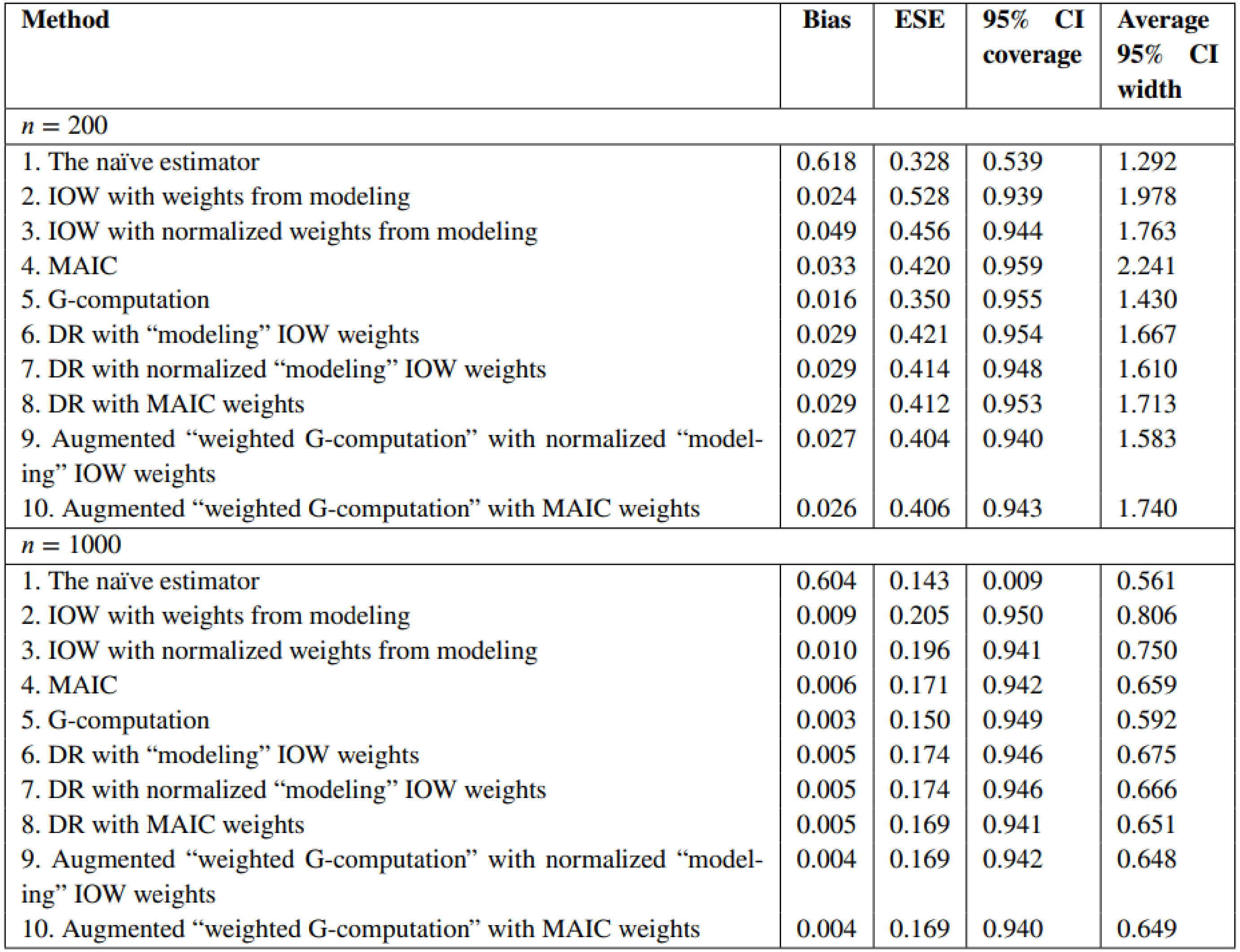
where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with

$$\Pr(S_i = 1 \mid \mathbf{Z}_i) = \text{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.5Z_{i4}).$$

Target estimand will be the ATC

KS1: both models correctly specified

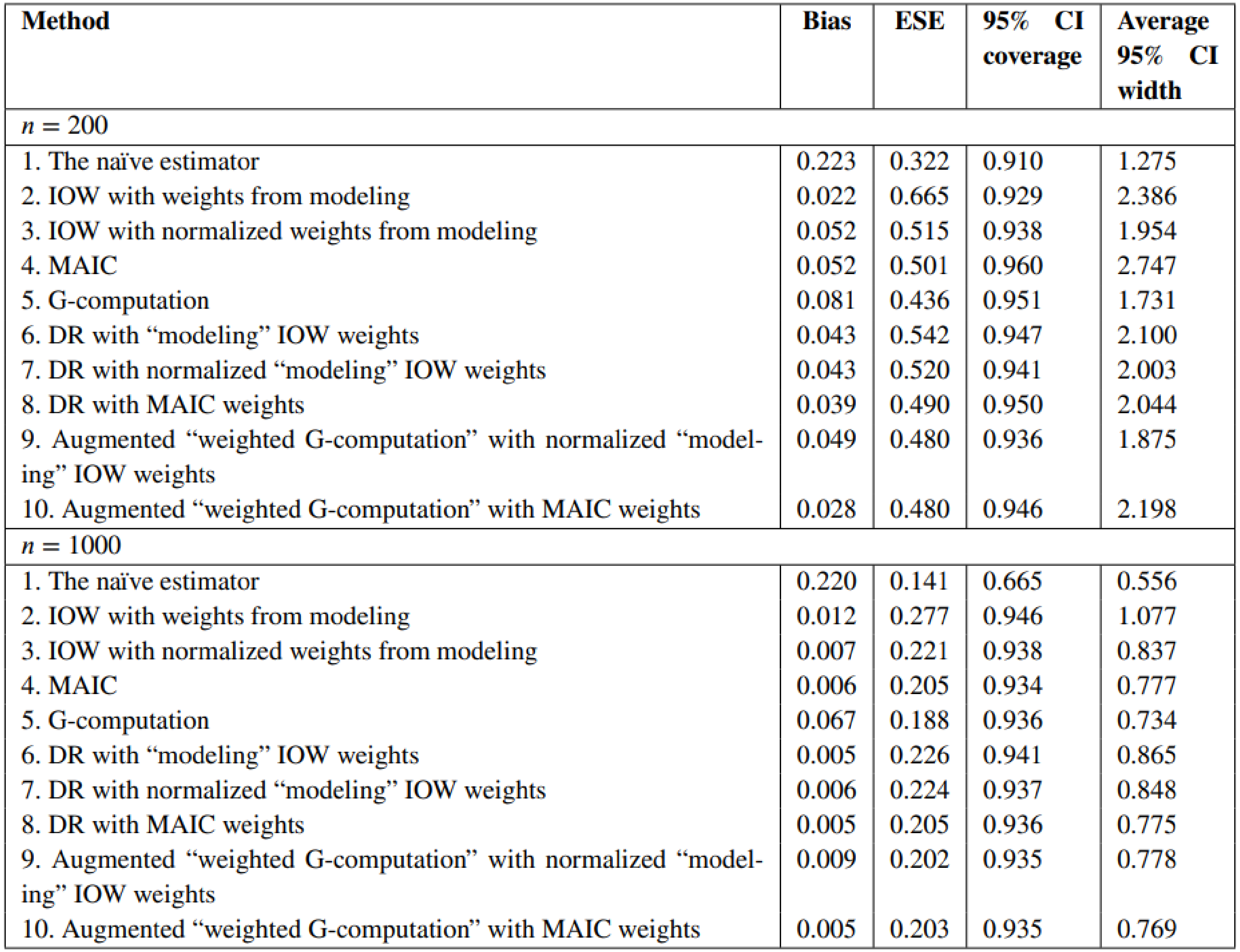
- The naïve estimator is biased
- All covariate-adjusted estimators are virtually unbiased under *n*=1000
- Some small-sample bias, even for theoretically consistent estimators, under *n*=200
- G-computation exhibits the greatest precision, but augmented weighting estimators are almost as precise



Method	**Bias**	**ESE**	**95% CI coverage**	**Average 95% CI width**
n = 200				
1. The naïve estimator	0.618	0.328	0.539	1.292
2. IOW with weights from modeling	0.024	0.528	0.939	1.978
3. IOW with normalized weights from modeling	0.049	0.456	0.944	1.763
4. MAIC	0.033	0.420	0.959	2.241
5. G-computation	0.016	0.350	0.955	1.430
6. DR with "modeling" IOW weights	0.029	0.421	0.954	1.667
7. DR with normalized "modeling" IOW weights	0.029	0.414	0.948	1.610
8. DR with MAIC weights	0.029	0.412	0.953	1.713
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.027	0.404	0.940	1.583
10. Augmented "weighted G-computation" with MAIC weights	0.026	0.406	0.943	1.740
n = 1000				
1. The naïve estimator	0.604	0.143	0.009	0.561
2. IOW with weights from modeling	0.009	0.205	0.950	0.806
3. IOW with normalized weights from modeling	0.010	0.196	0.941	0.750
4. MAIC	0.006	0.171	0.942	0.659
5. G-computation	0.003	0.150	0.949	0.592
6. DR with "modeling" IOW weights	0.005	0.174	0.946	0.675
7. DR with normalized "modeling" IOW weights	0.005	0.174	0.946	0.666
8. DR with MAIC weights	0.005	0.169	0.941	0.651
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.004	0.169	0.942	0.648
10. Augmented "weighted G-computation" with MAIC weights	0.004	0.169	0.940	0.649

KS2: PS model correctly specified

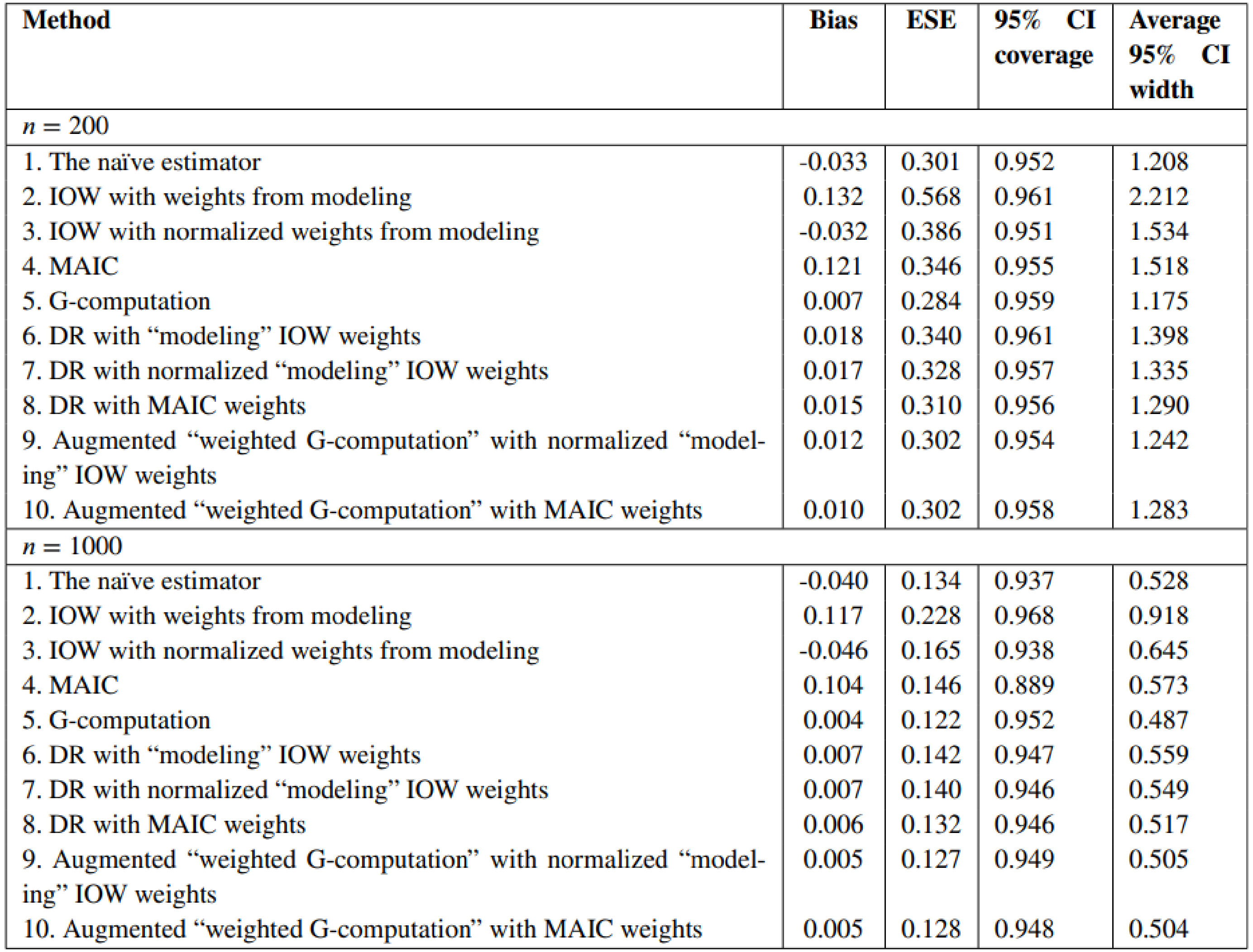
- G-computation exhibits bias
- Non-augmented and augmented weighting estimators are unbiased for *n*=1000 (weight normalization improves precision)
- Some small-sample bias, even for theoretically consistent weighting estimators, under *n*=200
- Outcome model misspecification does not induce a loss of precision for the augmented estimators compared to their non-augmented counterparts



Method	**Bias**	**ESE**	**95% CI coverage**	**Average 95% CI width**
n = 200				
1. The naïve estimator	0.223	0.322	0.910	1.275
2. IOW with weights from modeling	0.022	0.665	0.929	2.386
3. IOW with normalized weights from modeling	0.052	0.515	0.938	1.954
4. MAIC	0.052	0.501	0.960	2.747
5. G-computation	0.081	0.436	0.951	1.731
6. DR with “modeling” IOW weights	0.043	0.542	0.947	2.100
7. DR with normalized “modeling” IOW weights	0.043	0.520	0.941	2.003
8. DR with MAIC weights	0.039	0.490	0.950	2.044
9. Augmented “weighted G-computation” with normalized “modeling” IOW weights	0.049	0.480	0.936	1.875
10. Augmented “weighted G-computation” with MAIC weights	0.028	0.480	0.946	2.198
n = 1000				
1. The naïve estimator	0.220	0.141	0.665	0.556
2. IOW with weights from modeling	0.012	0.277	0.946	1.077
3. IOW with normalized weights from modeling	0.007	0.221	0.938	0.837
4. MAIC	0.006	0.205	0.934	0.777
5. G-computation	0.067	0.188	0.936	0.734
6. DR with “modeling” IOW weights	0.005	0.226	0.941	0.865
7. DR with normalized “modeling” IOW weights	0.006	0.224	0.937	0.848
8. DR with MAIC weights	0.005	0.205	0.936	0.775
9. Augmented “weighted G-computation” with normalized “modeling” IOW weights	0.009	0.202	0.935	0.778
10. Augmented “weighted G-computation” with MAIC weights	0.005	0.203	0.935	0.769

KS3: Outcome model correctly specified

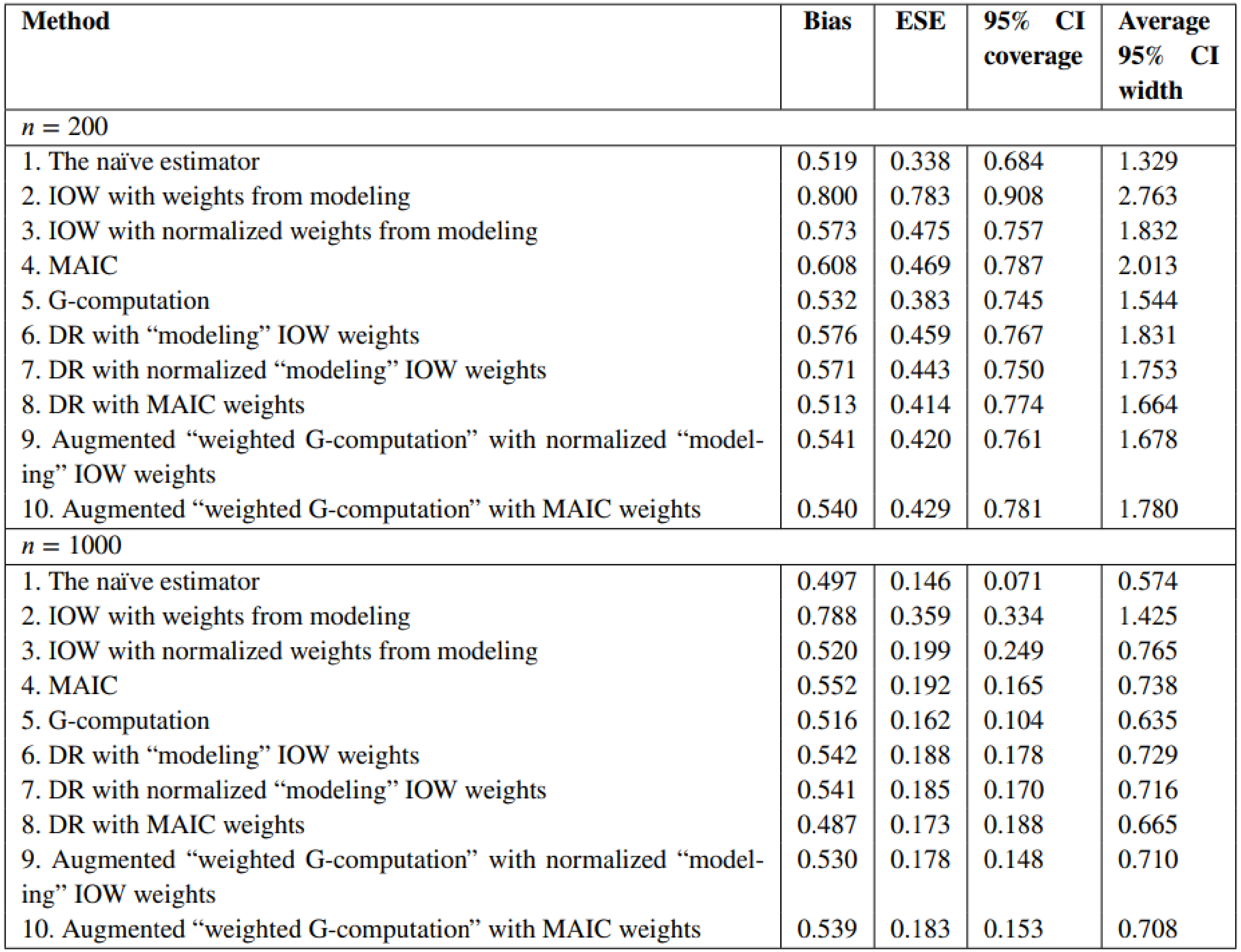
- Non-augmented weighting estimators exhibit bias; including the MAIC (entropy balancing) approach
- MAIC (entropy balancing) is not doubly robust with a logistic outcome model
- Augmented weighting estimators are generally more precise than their non-augmented weighting counterparts
- G-computation exhibits the greatest precision, but augmented weighting estimators are almost as precise



Method	Bias	ESE	95% CI coverage	Average 95% CI width
n = 200				
1. The naïve estimator	-0.033	0.301	0.952	1.208
2. IOW with weights from modeling	0.132	0.568	0.961	2.212
3. IOW with normalized weights from modeling	-0.032	0.386	0.951	1.534
4. MAIC	0.121	0.346	0.955	1.518
5. G-computation	0.007	0.284	0.959	1.175
6. DR with "modeling" IOW weights	0.018	0.340	0.961	1.398
7. DR with normalized "modeling" IOW weights	0.017	0.328	0.957	1.335
8. DR with MAIC weights	0.015	0.310	0.956	1.290
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.012	0.302	0.954	1.242
10. Augmented "weighted G-computation" with MAIC weights	0.010	0.302	0.958	1.283
n = 1000				
1. The naïve estimator	-0.040	0.134	0.937	0.528
2. IOW with weights from modeling	0.117	0.228	0.968	0.918
3. IOW with normalized weights from modeling	-0.046	0.165	0.938	0.645
4. MAIC	0.104	0.146	0.889	0.573
5. G-computation	0.004	0.122	0.952	0.487
6. DR with "modeling" IOW weights	0.007	0.142	0.947	0.559
7. DR with normalized "modeling" IOW weights	0.007	0.140	0.946	0.549
8. DR with MAIC weights	0.006	0.132	0.946	0.517
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.005	0.127	0.949	0.505
10. Augmented "weighted G-computation" with MAIC weights	0.005	0.128	0.948	0.504

KS4: Dual model misspecification

- All approaches are biased
- Augmentation via an outcome model does not protect against the simultaneous misspecification of two models
- There is no bias or variance amplification for the augmented estimators under dual model misspecification!



Method	Bias	ESE	95% CI coverage	Average 95% CI width
$n = 200$				
1. The naïve estimator	0.519	0.338	0.684	1.329
2. IOW with weights from modeling	0.800	0.783	0.908	2.763
3. IOW with normalized weights from modeling	0.573	0.475	0.757	1.832
4. MAIC	0.608	0.469	0.787	2.013
5. G-computation	0.532	0.383	0.745	1.544
6. DR with “modeling” IOW weights	0.576	0.459	0.767	1.831
7. DR with normalized “modeling” IOW weights	0.571	0.443	0.750	1.753
8. DR with MAIC weights	0.513	0.414	0.774	1.664
9. Augmented “weighted G-computation” with normalized “modeling” IOW weights	0.541	0.420	0.761	1.678
10. Augmented “weighted G-computation” with MAIC weights	0.540	0.429	0.781	1.780
$n = 1000$				
1. The naïve estimator	0.497	0.146	0.071	0.574
2. IOW with weights from modeling	0.788	0.359	0.334	1.425
3. IOW with normalized weights from modeling	0.520	0.199	0.249	0.765
4. MAIC	0.552	0.192	0.165	0.738
5. G-computation	0.516	0.162	0.104	0.635
6. DR with “modeling” IOW weights	0.542	0.188	0.178	0.729
7. DR with normalized “modeling” IOW weights	0.541	0.185	0.170	0.716
8. DR with MAIC weights	0.487	0.173	0.188	0.665
9. Augmented “weighted G-computation” with normalized “modeling” IOW weights	0.530	0.178	0.148	0.710
10. Augmented “weighted G-computation” with MAIC weights	0.539	0.183	0.153	0.708

Additional remarks

- We hypothesized that entropy balancing weights, like those employed by MAIC, can lead to more stable and precise ATC estimation than inverse odds modelling weights
- This is confirmed for the non-augmented estimators in our simulation study; MAIC exhibits greater precision than (normalized or non-normalized) modelling weighting approaches in all scenarios
- The precision gains have been inherited by the augmented approaches; estimators using MAIC weights generally display enhanced precision compared to those using modelling weights
- The augmented “weighted G-computation” estimators are also doubly robust for the ATC, noting that the logistic outcome model has a canonical link function
- The augmented “weighting G-computation” estimators offer similar performance than our proposed doubly robust augmented estimator with MAIC weights (these are the least biased and most precise estimators)

Concluding remarks

Concluding remarks

- Externally controlled single-arm trials and unanchored ITCs are different versions of the same problem
- The use of modern causal inference methods, e.g., DR methods, data-adaptive estimation, remains underexploited in HTA and unanchored ITCs (and evidence synthesis in general)
- HTA and unanchored ITCs should start considering doubly robust augmented approaches
- Entropy balancing approaches to weighting have desirable properties, regardless of subject-level data availability
- This is not the only common methodological theme across regulatory vs. HTA: transportability vs. anchored indirect comparisons, causal meta-analysis, etc.

References

- Campbell, H. and Remiro-Azócar, A., 2025. Doubly robust augmented weighting estimators for the analysis of externally controlled single-arm trials and unanchored indirect treatment comparisons. arXiv preprint arXiv:2505.00113.
- Dahabreh, I.J., Robertson, S.E., Steingrimsson, J.A., Stuart, E.A. and Hernan, M.A., 2020. Extending inferences from a randomized trial to a new target population. Statistics in medicine, 39(14), pp.1999-2014.
- Gabriel, E.E., Sachs, M.C., Martinussen, T., Waernbaum, I., Goetghebeur, E., Vansteelandt, S. and Sjölander, A., 2024. Inverse probability of treatment weighting with generalized linear outcome models for doubly robust estimation. Statistics in Medicine, 43(3), pp.534-547.
- Josey, K.P., Berkowitz, S.A., Ghosh, D. and Raghavan, S., 2021. Transporting experimental results with entropy balancing. Statistics in medicine, 40(19), pp.4310-4326.
- Słoczyński, T., Uysal, S.D. and Wooldridge, J.M., 2023. Covariate balancing and the equivalence of weighting and doubly robust estimators of average treatment effects. arXiv preprint arXiv:2310.18563.
- Signorovitch, J.E., Wu, E.Q., Yu, A.P., Gerrits, C.M., Kantor, E., Bao, Y., Gupta, S.R. and Mulani, P.M., 2010. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. Pharmacoeconomics, 28, pp.935-945.
- Westreich, D., Edwards, J.K., Lesko, C.R., Stuart, E. and Cole, S.R., 2017. Transportability of trial results using inverse odds of sampling weights. American journal of epidemiology, 186(8), pp.1010-1014.
- Zhao, Q. and Percival, D., 2017. Entropy balancing is doubly robust. Journal of causal inference, 5(1), p.20160010.