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

Antonio Remiro-Azócar, PhD

Methods and Outreach, Novo Nordisk

“When worlds collide: Common methodological themes in meta-analysis, causal inference, and hybrid trial design”

ISCB46, 25th August 2025

Acknowledgements

This is joint work developed together with Harlan Campbell (University of British Columbia & Precision AQ)

Agenda

1. Context, estimands, assumptions
2. Estimators
3. Simulation study
4. Methodological extensions
5. Concluding remarks

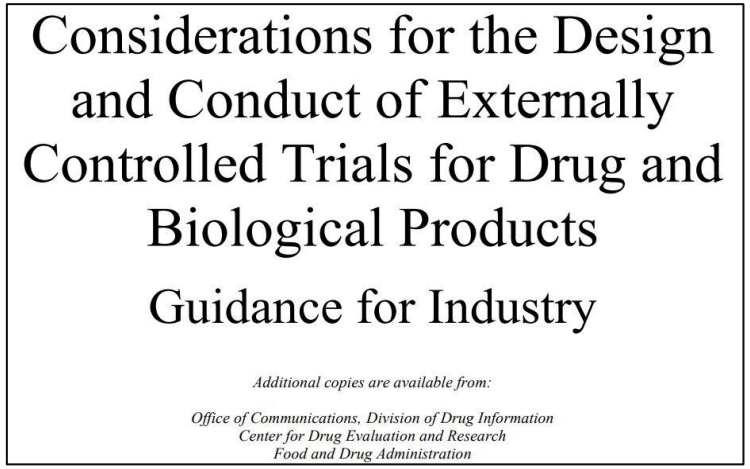
Context, estimands, assumptions

Single-arm trials (SATs)

Conducting RCTs might 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 special circumstances



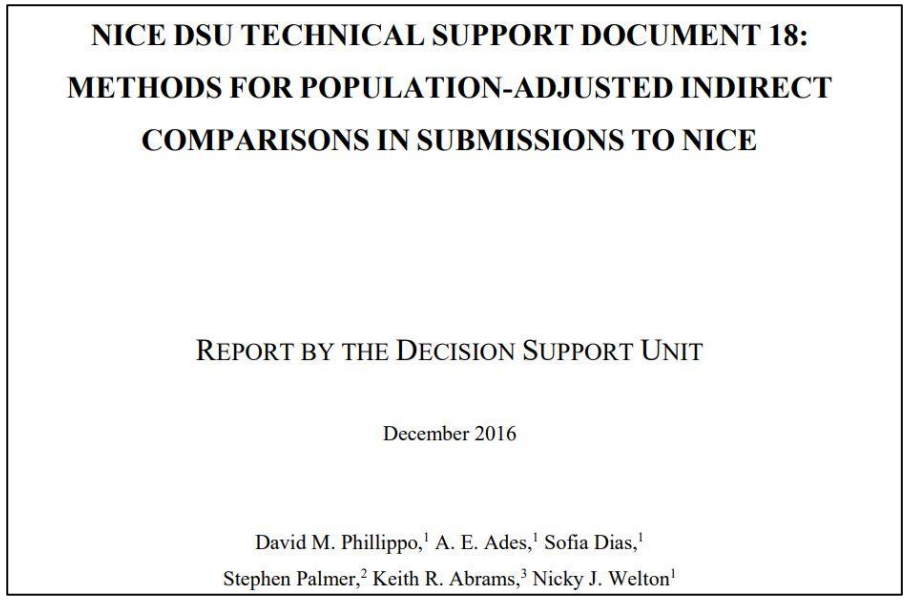
Regulatory submissions featuring externally controlled SATs are rising, mainly through accelerated approval pathways

Unanchored indirect treatment comparisons (ITCs)

Health technology assessment (HTA) requires comparisons versus all treatments in routine clinical practice

- 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
- Some therapeutic areas are rapidly evolving, with a changing comparator landscape and no single accepted standard of care
- It is not always possible to find compatible control arms with which to “anchor” an indirect treatment comparison

Unanchored indirect treatment comparisons may be required


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 a competitor-sponsored historical trial
- There may be limited access to subject-level data for the external control, only aggregate-level data from publications

Different **estimands** or **summary measures** can be targeted:

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

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

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

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

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

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

ATT or ATC?

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

Average treatment effect in the treated (ATT)

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

Attractive for the regulatory context…

- Consistent with the emulation of a randomized comparison in the registrational SAT
- The external control would aim to “mimic” the internal control arm of “pivotal” clinical trial
- Compatible with the mean absolute outcome targeted by the SAT, preserving the original SAT results
- Typically, the primary estimand for externally-controlled SATs seeking drug approval in the regulatory environment

Nevertheless…

- 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)

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

Typically, the target estimand in HTA…due to necessity as subject-level data are often unavailable for the external control

Potentially more desirable for external validity…

- External controls based on RWD or natural history studies: broad inclusion criteria targeting 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

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

Four critical assumptions

Based on Zhou et al (2024)

1. No direct effect of trial participation

- Trial participation does not affect the outcome except through treatment assignment itself (no Hawthorne effects)

2. Stable unit treatment value (SUTVA)

- No interference between subjects and treatment variation irrelevance (one well-defined version of each treatment)

$$Y_i = Y_i^1 T_i + Y_i^0 (1 - T_i)$$

3. Conditional ignorability of data source assignment

- ATT: Conditional on covariates, potential outcomes under the control are independent of the data source

$$Y_i^0 \perp S_i \mid \mathbf{X}_i$$

- ATC: Conditional on covariates, potential outcomes under the active intervention are independent of the data source

$$Y_i^1 \perp S_i \mid \mathbf{X}_i$$

STRONG IGNORABILITY

4. Overlap or positivity

- ATT: Support of the covariates in the SAT is contained within that of the external control $0 < \Pr(S = 0 \mid \mathbf{X} = \mathbf{x}) < 1, \forall \mathbf{x}\ f(\mathbf{x} \mid S = 1) > 0$
- ATC: Support of the covariates in the external control is contained within that of the SAT $0 < \Pr(S = 1 \mid \mathbf{X} = \mathbf{x}) < 1, \forall \mathbf{x}\ f(\mathbf{x} \mid S = 0) > 0$

These assumptions are unverifiable and potentially unreasonable in many practical scenarios…proceed with care (Senn 2025)

Estimators

Contextual differences

EXTERNALLY CONTROLLED SATs (REGULATORY)

- Modeling-based approach to odds 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 (MAIC): entropy balancing-based approach to odds 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 scores are estimated by maximizing the fit of a logistic regression
- Estimated weights do not produce adequate balance if the propensity score model is mis-specified
- Even a correctly specified propensity score model does not guarantee balance in finite samples
- 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
- Weights constrained to be positive and sample-bounded (interpolation as opposed to extrapolation)
- Minimally dispersed weights, which translates into larger effective sample sizes and precision
- Applicable where aggregate-level covariate moments are available for the external control

Modeling approach to weighting

Targeting the ATC

Inverse odds weights (IOW) defined as: $w_i = \frac{(1-e_i)S_i}{e_i} + (1-S_i),$ with $e_i = e(\mathbf{X}_i) = \Pr(S_i = 1 \mid \mathbf{X}_i)$

Weights are inverse conditional odds of SAT participation (conditional odds of external control participation)

A logistic regression is fitted to the concatenated SAT and external control subject-level data, typically 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 are plugged into the weight equation to derive weight estimates

The weighted average of observed outcomes under the active intervention is contrasted with the unweighted average of observed outcomes for the external control

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

weights normalized to sum to one to improve finite sample properties and provide more stable and precise estimation

Entropy balancing approach to weighting

Targeting the ATC

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

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

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

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

Bias-robustness considerations

Modeling approach is consistent if…

- The propensity score model for data source assignment is correctly specified
- That is, the logit of the propensity score (conditional probability of SAT participation) varies linearly with the covariate balance functions

Entropy balancing is consistent if…

- The logit of the conditional probability of external control participation (or SAT participation) **OR** the conditional outcome expectation under the active intervention varies linearly with the covariate balance functions
- For instance, mean-balancing ensures consistency if $\text{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \boldsymbol{\alpha}$ OR $\mathrm{E}(Y_i^1 \mid \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}$
- Mean- and variance-balancing ensures consistency if $\text{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \boldsymbol{\alpha_1} + (\mathbf{X}_i^2)^\top \boldsymbol{\alpha_2}$ OR $\mathrm{E}(Y_i^1 \mid \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}_1 + (\mathbf{X}_i^2)^\top \boldsymbol{\beta}_2$

Entropy balancing is **linear doubly robust**:

- “DR with respect to linear outcome regression and logistic propensity score regression” (Zhao and Percival, 2017)

Entropy balancing claimed more bias-robust as it is consistent under a greater number of distinct data-generating mechanisms

Is entropy balancing doubly robust?

It is rarely plausible that outcomes vary linearly with the covariates

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

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

- Increasing balancing constraints → more likely that covariate moments fall outside the convex hull of the observed covariate space
- Namely, feasible weighting solutions to the convex optimization problem do not exist (no set of positive weights can enforce balance)
- Increasing balancing constraints → further reductions in effective sample size and precision
- Aggregate data beyond means and variances (e.g., higher-order moments and means of transformed covariates) rarely reported

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

Augmented entropy balancing

Targeting the ATC

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:

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

Weighted G-computation

Targeting the ATC

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

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

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

Note: this is only doubly robust 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)

Simulation study

Data-generating mechanisms

KS1: propensity score and outcome model correctly specified

KS1: Y_i is generated from a Bernoulli distribution with

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

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

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

KS2: only propensity score model correctly specified

KS2: Y_i is generated from a Bernoulli distribution with

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

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

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

$Z_{i1} = \text{scale}(\exp(X_{i1}/2))$,

$Z_{i2} = \text{scale}(X_{i2}^2)$,

$Z_{i3} = \text{scale}((X_{i1}X_{i3} + 0.6)^3)$,

$Z_{i4} = \text{scale}((X_{i2} + X_{i4} + 20)^2)$

KS3: only outcome model correctly specified

KS3: Y_i is generated from a Bernoulli distribution with

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

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

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

KS4: propensity score and outcome model incorrectly specified

KS4: Y_i is generated from a Bernoulli distribution with

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

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

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

Target estimand will be the ATC

Variance estimation for all methods using non-parametric bootstrap

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

Discussion

- We hypothesized that entropy balancing weights can lead to more stable and precise ATC estimation than inverse odds modeling weights
- This is confirmed for the non-augmented estimators in the simulation study; entropy balancing 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 entropy balancing weights generally display enhanced precision compared to those using modeling 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 entropy balancing weights (these are the least biased and most precise estimators)

Methodological extensions

Targeting the ATT

› 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:

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

› 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

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

› 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:

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

› 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

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

Unavailable subject-level data for the control

› PRELIMINARY STEP (all methods except non-augmented entropy balancing)

- *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, random seed sensitivity
- 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
- Stack SAT subject-level covariate data with simulated subject-level covariate data for the external control

› WEIGHTING

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

› DR AUGMENTED WEIGHTING

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

› G-COMPUTATION

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

› 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 typical non-parametric bootstrap procedure
- Only bootstrap the SAT, as opposed to the concatenated SAT and external control
- 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)

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

Subject-level data are necessary

ATC/ATT/ATE depend on the full joint covariate distribution of the target (sub)population

In the absence of subject-level data for the external control:

- ATT not estimable or identifiable
- ATC typically not identifiable without further – implicit or explicit – assumptions about correlations, distributional forms, etc. in the external control
- Impossible to adjust for differences across data sources in “non-population” elements of the research question (misalignments in other estimand attributes or specification of “time zero” in target trial emulation!)
- Variance estimation issues: assuming the external control covariate distributional data as fixed leads to inflated Type I error rates, specially with small sample sizes (Josey et al 2021)

“In any situation with non-randomized data, such as observational evidence and single-arm trials, (…) complete access to the individual patient data is required” (Methodological Guideline for Quantitative Evidence Synthesis, HTA Coordination Group, 2024)

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 and causally-interpretable 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.
- 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.
- Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons. https://health.ec.europa.eu/latest-updates/methodological-guideline-quantitative-evidence-synthesis-direct-and-indirect-comparisons-2024-03-25_en; 2024. Accessed: 11-08-2025.
- Senn, S., 2025. Causal estimates for external controls. How reasonable are the assumptions?. Journal of the Royal Statistical Society Series A: Statistics in Society, p.qnaf127.
- 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.
- Zhao, Q. and Percival, D., 2017. Entropy balancing is doubly robust. Journal of causal inference, 5(1), p.20160010.
- Zhou, X., Zhu, J., Drake, C. and Pang, H., 2025. Causal estimators for incorporating external controls in randomized trials with longitudinal outcomes. Journal of the Royal Statistical Society Series A: Statistics in Society, 188(3), pp.791-818.