

# Protocol for a Simulation Study Comparing Alternative Weighting Approaches for Matching-Adjusted Indirect Comparisons

true

January 2026

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Background and Rationale . . . . .	2
1.2	Objectives . . . . .	3
<b>2</b>	<b>Methods</b>	<b>3</b>
2.1	Overview of Matching-Adjusted Indirect Comparison . . . . .	3
2.2	Method of Moments Approach . . . . .	3
2.3	Entropy Balancing Approach . . . . .	4
2.4	Non-Uniform Base Weights Framework . . . . .	5
2.5	Alternative Loss Functions: The Cressie-Read Divergence Family . . . . .	5
2.6	Arm-Separate Weighting Schemes . . . . .	5
2.7	Data Generation Model . . . . .	6
2.8	Simulation Scenarios . . . . .	6
2.9	Weighting Methods to be Compared . . . . .	7
2.10	Estimands and Treatment Effect Estimation . . . . .	7
2.11	Performance Metrics . . . . .	7
2.12	Statistical Analysis Plan . . . . .	8
2.13	Software and Reproducibility . . . . .	8
<b>3</b>	<b>Timeline</b>	<b>8</b>
<b>4</b>	<b>Ethics Statement</b>	<b>8</b>
<b>5</b>	<b>Funding</b>	<b>8</b>
	<b>References</b>	<b>8</b>

# 1 Introduction

## 1.1 Background and Rationale

Health technology assessment agencies and regulatory bodies frequently require comparative effectiveness evidence to inform decision-making regarding the reimbursement and approval of new interventions. In many therapeutic areas, however, head-to-head randomized controlled trials comparing all relevant treatment options are unavailable. This evidence gap creates a methodological challenge: how can reliable treatment comparisons be made when direct comparative data do not exist?

The traditional approach to this problem is the indirect treatment comparison, first formalized by Bucher and colleagues in 1997 (Bucher et al. 1997). The Bucher method enables comparison of treatments B and C through a common comparator A by computing the difference between relative effects observed in separate trials (AB and AC). Specifically, if  $d_{AB}$  represents the relative treatment effect of B versus A and  $d_{AC}$  represents the effect of C versus A, then the indirect comparison of B versus C is computed as  $d_{BC} = d_{AB} - d_{AC}$ . This approach relies on a critical assumption: that the relative treatment effects are consistent across the trial populations, an assumption often referred to as the similarity or transitivity assumption.

In practice, the transitivity assumption is frequently violated because patient populations enrolled in different trials may differ systematically in their baseline characteristics. These cross-trial differences become particularly problematic when the differing characteristics are effect modifiers—variables that modify the magnitude or direction of treatment effects. When effect modification is present, the relative treatment effect observed in one trial population may not be applicable to another population with different covariate distributions.

Matching-adjusted indirect comparison (MAIC), introduced by Signorovitch and colleagues in 2010 (Signorovitch et al. 2010), addresses this limitation by reweighting individual patient data (IPD) from one trial to match the aggregate covariate distribution reported in another trial. By aligning the populations through propensity score-like weighting, MAIC aims to reduce or eliminate bias arising from cross-trial differences in effect modifiers. The method has gained widespread adoption in health technology assessment submissions to agencies such as NICE in the United Kingdom and has become a standard tool in the evidence synthesis toolkit.

Despite its popularity, several methodological questions regarding MAIC implementation remain inadequately addressed. The original MAIC method, based on the method of moments, estimates weights by solving a system of equations that equate weighted covariate means in the IPD to the corresponding means in the aggregate data. Phillippo and colleagues (Phillippo et al. 2020) recently established a fundamental theoretical result: the method of moments approach is mathematically equivalent to entropy balancing, an approach that finds weights minimizing the Kullback-Leibler divergence from uniform weights subject to the same moment-matching constraints. This equivalence, proven through convex optimization duality theory, opens new methodological possibilities that have not been systematically evaluated.

First, the entropy balancing formulation naturally accommodates non-uniform base weights, enabling the combination of MAIC with other adjustment methods. For instance, base weights could incorporate inverse probability of censoring weights (IPCW) to adjust for treatment switching, or inverse probability of treatment weights (IPTW) for nonparametric covariate adjustment aimed at variance reduction (Williamson et al. 2014). Second, the optimization framework suggests that alternative loss functions from the Cressie-Read divergence family could be employed instead of entropy, potentially offering different bias-variance tradeoffs (Cressie and Read 1984). The empirical likelihood loss function, a special case of the Cressie-Read family, is particularly interesting given its established statistical properties (Owen 2001).

Additionally, Petto and colleagues (Petto et al. 2019) proposed arm-separate weighting schemes that match covariates separately for treatment and control arms rather than matching the total population. When arm-specific aggregate data is available, this approach may preserve within-arm balance better than total-population matching. However, the relative performance of these approaches across different data scenarios has not been comprehensively evaluated.

## 1.2 Objectives

The primary objective of this simulation study is to compare the statistical performance of alternative MAIC weighting methods across a range of data-generating scenarios relevant to health technology assessment. Specifically, this study aims to evaluate bias, precision, coverage probability, and effective sample size for the following methods: standard method of moments MAIC, entropy balancing MAIC, empirical likelihood MAIC, and arm-separate entropy balancing. Secondary objectives include empirically confirming the theoretical equivalence of method of moments and entropy balancing approaches, and identifying scenario characteristics that favor particular weighting methods.

## 2 Methods

### 2.1 Overview of Matching-Adjusted Indirect Comparison

Matching-adjusted indirect comparison addresses the problem of comparing treatments when individual patient data are available from one trial but only aggregate summary statistics are available from another. Consider a scenario in which we wish to compare treatments B and C, but no head-to-head trial exists. Instead, we have access to IPD from a trial comparing A versus B (the AB trial) and aggregate data from a published trial comparing A versus C (the AC trial).

The fundamental challenge is that patients enrolled in the AB trial may differ systematically from those in the AC trial. Let  $X$  denote a vector of baseline covariates that may modify treatment effects. If the distribution of  $X$  differs between trials, the relative treatment effects  $d_{AB}$  and  $d_{AC}$  may not be directly comparable because they were estimated in different target populations. MAIC addresses this by reweighting the IPD from the AB trial to create a pseudo-population that matches the covariate distribution of the AC trial.

Formally, let  $\{(X_i, T_i, Y_i)\}_{i=1}^n$  denote the IPD from the AB trial, where  $X_i$  is the covariate vector,  $T_i \in \{0, 1\}$  is the treatment indicator (0 for A, 1 for B), and  $Y_i$  is the outcome. Let  $\bar{X}^{AC}$  denote the vector of published covariate means from the AC trial. The goal of MAIC is to find weights  $w_i \geq 0$  such that the weighted covariate means in the IPD equal the AC trial means:

$$\sum_{i=1}^n w_i X_i = \bar{X}^{AC}$$

where the weights are normalized to sum to one. After obtaining these weights, the treatment effect  $d_{AB}$  is estimated using weighted regression or a weighted estimator appropriate for the outcome type. The indirect comparison is then computed as  $d_{BC} = d_{AB}^{weighted} - d_{AC}$ .

### 2.2 Method of Moments Approach

The original MAIC method proposed by Signorovitch et al. (Signorovitch et al. 2010, 2012) obtains weights through the method of moments. The key insight is that propensity score-style weights can be derived by modeling the odds of trial membership. If we consider a hypothetical combined population of the AB and AC trials, the odds that a patient belongs to the AC trial given their covariates can be modeled using a logistic function:

$$\frac{P(S = AC|X)}{P(S = AB|X)} = \exp(\alpha^T X)$$

where  $S$  denotes trial membership and  $\alpha$  is a vector of parameters. The weights for the IPD patients are then defined as  $w_i \propto \exp(\alpha^T X_i)$ , normalized to sum to one.

The parameter vector  $\alpha$  is estimated by solving the moment conditions that equate weighted means to the target values. Specifically,  $\alpha$  is found such that:

$$\frac{\sum_{i=1}^n \exp(\alpha^T X_i) X_i}{\sum_{i=1}^n \exp(\alpha^T X_i)} = \bar{X}^{AC}$$

This system of  $p$  equations in  $p$  unknowns (where  $p$  is the number of covariates) can be solved using standard numerical optimization methods. In practice, the equations are typically centered by defining  $\tilde{X}_i = X_i - \bar{X}^{AC}$  and solving for  $\alpha$  such that the weighted mean of the centered covariates equals zero.

## 2.3 Entropy Balancing Approach

Entropy balancing, introduced by Hainmueller (Hainmueller 2012) in the causal inference literature, provides an alternative perspective on the weight estimation problem. Rather than deriving weights from a propensity score model, entropy balancing directly seeks weights that minimize a measure of distance from a set of base weights subject to the covariate balance constraints.

The standard entropy balancing problem is formulated as:

$$\min_w \sum_{i=1}^n w_i \log \left( \frac{w_i}{w_i^{(0)}} \right)$$

subject to:

$$\begin{aligned} \sum_{i=1}^n w_i X_i &= \bar{X}^{AC} \\ \sum_{i=1}^n w_i &= 1 \\ w_i &\geq 0 \text{ for all } i \end{aligned}$$

where  $w_i^{(0)}$  denotes the base weights, typically set to uniform weights  $1/n$ . The objective function is the Kullback-Leibler divergence (or relative entropy) between the weight distribution  $w$  and the base weight distribution  $w^{(0)}$ .

This constrained optimization problem can be solved using the method of Lagrange multipliers. The Lagrangian is:

$$L(w, \alpha, \lambda) = \sum_{i=1}^n w_i \log \left( \frac{w_i}{w_i^{(0)}} \right) + \alpha^T \left( \bar{X}^{AC} - \sum_{i=1}^n w_i X_i \right) + \lambda \left( 1 - \sum_{i=1}^n w_i \right)$$

Taking the derivative with respect to  $w_i$  and setting it equal to zero yields the optimal weights in terms of the Lagrange multipliers:

$$w_i = w_i^{(0)} \exp(\alpha^T X_i - 1 - \lambda)$$

After normalization, this simplifies to:

$$w_i = \frac{w_i^{(0)} \exp(\alpha^T X_i)}{\sum_{j=1}^n w_j^{(0)} \exp(\alpha^T X_j)}$$

Phillippo et al. (Phillippo et al. 2020) proved that when the base weights are uniform, the optimal  $\alpha$  from entropy balancing is identical to the  $\alpha$  obtained from the method of moments. This equivalence follows from convex duality: the method of moments objective function is precisely the dual of the entropy balancing primal problem. Consequently, the two approaches yield identical weights and treatment effect estimates.

## 2.4 Non-Uniform Base Weights Framework

The equivalence result of Phillippo et al. (Phillippo et al. 2020) extends beyond uniform base weights. When non-uniform base weights  $w_i^{(0)}$  are specified, entropy balancing finds weights that minimize the divergence from these base weights while satisfying the balance constraints. This generalization enables the combination of MAIC with other weighting-based adjustment methods.

One important application is variance reduction through nonparametric covariate adjustment (NPCA), as described by Williamson et al. (Williamson et al. 2014). In a randomized trial, although treatment assignment is independent of baseline covariates by design, adjusting for prognostic covariates can improve precision. NPCA uses inverse probability of treatment weights (IPTW) to create balance not just on the treatment assignment mechanism (which is already balanced by randomization) but on the entire covariate distribution. When combined with MAIC, the base weights could incorporate IPTW adjustment for prognostic covariates, potentially reducing variance while MAIC adjusts for population differences.

Another application is adjustment for treatment switching using inverse probability of censoring weights (IPCW). In oncology trials, patients randomized to the control arm may switch to the experimental treatment upon disease progression, confounding the treatment effect estimate. IPCW methods model the probability of switching and use inverse probability weights to adjust for this informative censoring. By incorporating IPCW weights as base weights in entropy balancing MAIC, both treatment switching and population differences can be addressed simultaneously.

## 2.5 Alternative Loss Functions: The Cressie-Read Divergence Family

While entropy (Kullback-Leibler divergence) is the most common choice for balancing weights, alternative divergence measures may offer different statistical properties. The Cressie-Read family of divergence measures provides a spectrum of options parameterized by a scalar  $\gamma$  (Cressie and Read 1984):

$$D_\gamma(w||w^{(0)}) = \frac{1}{\gamma(\gamma+1)} \sum_{i=1}^n w_i^{(0)} \left[ \left( \frac{w_i}{w_i^{(0)}} \right)^{\gamma+1} - 1 \right]$$

Different values of  $\gamma$  yield different divergence measures:  $\gamma = 0$  corresponds to the Kullback-Leibler divergence (entropy),  $\gamma = -1$  yields the empirical likelihood,  $\gamma = 1$  gives the chi-squared divergence, and  $\gamma = -0.5$  produces the Hellinger distance.

The empirical likelihood case ( $\gamma = -1$ ) is particularly interesting from a statistical theory perspective. Owen (Owen 2001) established that empirical likelihood has appealing asymptotic properties, including Bartlett correctability and automatic determination of confidence region shapes. In the context of MAIC, empirical likelihood minimizes:

$$D_{-1}(w||w^{(0)}) = -2 \sum_{i=1}^n w_i^{(0)} \log \left( \frac{w_i}{w_i^{(0)}} \right)$$

subject to the same balance and normalization constraints. Whether these theoretical advantages translate to practical benefits in MAIC applications is an empirical question that this simulation study aims to address.

## 2.6 Arm-Separate Weighting Schemes

Standard MAIC matches the total IPD population to the total aggregate data population. Petto et al. (Petto et al. 2019) proposed an alternative approach in which covariates are matched separately for treatment and control arms. This arm-separate weighting is motivated by the observation that even in randomized trials, the treatment and control arms may have slightly different covariate distributions due to random imbalance. When arm-specific aggregate data is available from the comparator trial, matching each arm separately may preserve arm-specific balance better than total-population matching.

Let the aggregate data include arm-specific covariate means  $\bar{X}_{T=1}^{AC}$  for the treatment arm and  $\bar{X}_{T=0}^{AC}$  for the control arm. The arm-separate entropy balancing approach solves two separate optimization problems: one for the treated patients in the IPD (matching to  $\bar{X}_{T=1}^{AC}$ ) and one for the control patients (matching to  $\bar{X}_{T=0}^{AC}$ ).

An extension of this approach, denoted EbArmILD, additionally balances covariates that are available only in the IPD (individual-level data covariates) between the treatment and control arms within the IPD. This may be useful when strong prognostic factors are measured in the IPD but not reported in the aggregate data publications.

## 2.7 Data Generation Model

The simulation study will generate data from a two-trial scenario reflecting the typical MAIC application. The AB trial provides individual patient data with  $n_{AB}$  patients randomized 1:1 to treatments A and B. The AC trial is summarized only through aggregate statistics from  $n_{AC} = 500$  patients randomized to A and C.

Covariates will be generated from multivariate normal distributions:

$$X_{AB} \sim N(\mu_{AB}, I_p)$$

$$X_{AC} \sim N(\mu_{AC}, I_p)$$

where  $\mu_{AB}$  and  $\mu_{AC}$  are  $p$ -dimensional mean vectors and  $I_p$  is the  $p \times p$  identity matrix. The population shift between trials is controlled by setting  $\mu_{AC} = \mu_{AB} + \delta \cdot \mathbf{1}_p$ , where  $\delta$  determines the magnitude of the population difference and  $\mathbf{1}_p$  is a vector of ones.

Binary outcomes will be generated from logistic regression models that incorporate both prognostic effects and treatment-by-covariate interactions (effect modification):

$$\text{logit}(P(Y_i = 1)) = \alpha + \sum_{k=1}^p \beta_k X_{ik} + \tau T_i + \sum_{k \in EM} \gamma_k X_{ik} T_i$$

where  $\alpha$  is the intercept,  $\beta_k$  are prognostic effects,  $\tau$  is the main treatment effect, and  $\gamma_k$  are effect modification coefficients for covariates in the set  $EM$  of effect modifiers. Treatment effects will be set to  $\tau_{AB} = -0.5$  (log odds ratio for B vs A) and  $\tau_{AC} = -0.7$  (log odds ratio for C vs A), yielding a true indirect comparison effect of  $\theta_{BC} = 0.2$ .

The aggregate data from the AC trial will be generated by first simulating individual patient data from the AC population, computing the outcome model, and then summarizing the covariate means and the treatment effect estimate with its variance.

## 2.8 Simulation Scenarios

The simulation will vary the following factors in a full factorial design:

The IPD sample size  $n_{AB}$  will take values of 100, 300, and 500, representing small, moderate, and large trials respectively. The number of covariates  $p$  will be either 3 or 6, representing scenarios with few or many adjustment variables. The population overlap, controlled by the shift parameter  $\delta$ , will have three levels: high overlap ( $\delta = 0.2$ ), medium overlap ( $\delta = 0.5$ ), and low overlap ( $\delta = 1.0$ ). The number of effect modifiers will be 0, 1, or 2 covariates.

This factorial design yields  $3 \times 2 \times 3 \times 3 = 54$  primary scenarios. Each scenario will be replicated 1,000 times to ensure adequate precision in the performance metric estimates.

## 2.9 Weighting Methods to be Compared

The following five methods will be evaluated in the primary analysis:

The Bucher method serves as the unadjusted reference, computing the indirect comparison without any population adjustment. This method estimates the treatment effect from the unweighted IPD and subtracts the aggregate data treatment effect.

The SigTotal method implements standard MAIC using the method of moments approach with total population matching, as originally proposed by Signorovitch et al. The weights are estimated by solving the moment equations, and the treatment effect is estimated using weighted logistic regression.

The EbTotal method implements entropy balancing with uniform base weights, matching the total population. Based on the equivalence theorem of Phillippo et al., this method should yield identical results to SigTotal.

The EbArm method implements arm-separate entropy balancing, matching treatment and control arms separately to arm-specific aggregate data targets. This requires that arm-specific covariate means be available or estimable from the AC trial.

The EbEL method implements empirical likelihood loss instead of entropy loss, using the Cressie-Read divergence with  $\gamma = -1$  for total population matching.

## 2.10 Estimands and Treatment Effect Estimation

The primary estimand is the log odds ratio for the indirect comparison of treatment B versus treatment C, denoted  $\theta_{BC}$ . This is estimated as:

$$\hat{\theta}_{BC} = \hat{d}_{AB}^{weighted} - \hat{d}_{AC}$$

where  $\hat{d}_{AB}^{weighted}$  is the weighted treatment effect estimate from the IPD and  $\hat{d}_{AC}$  is the treatment effect reported in the aggregate data.

For each weighting method, the treatment effect  $\hat{d}_{AB}^{weighted}$  will be estimated using weighted logistic regression with robust (sandwich) standard errors. Specifically, a logistic regression model  $\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 T$  will be fitted with observation weights  $w_i$ , and  $\hat{d}_{AB}^{weighted} = \hat{\beta}_1$ . The variance will be estimated using the Huber-White sandwich estimator to account for the estimation of weights.

The variance of the indirect comparison is computed as:

$$\text{Var}(\hat{\theta}_{BC}) = \text{Var}(\hat{d}_{AB}^{weighted}) + \text{Var}(\hat{d}_{AC})$$

assuming independence between the trials. Confidence intervals will be constructed using normal approximation.

## 2.11 Performance Metrics

Four primary performance metrics will be computed for each method in each scenario:

Bias is defined as the difference between the mean estimated effect and the true effect:  $\text{Bias} = E[\hat{\theta}_{BC}] - \theta_{BC}$ . This will be estimated as the mean of the point estimates minus the true value of 0.2 across simulation replications.

Root mean squared error (RMSE) is defined as  $\text{RMSE} = \sqrt{E[(\hat{\theta}_{BC} - \theta_{BC})^2]}$ , capturing both bias and variance in a single metric.

Coverage probability is the proportion of 95% confidence intervals that contain the true parameter value. Nominal coverage is 95%, and departures from nominal coverage indicate problems with the variance estimation or distributional assumptions.

Effective sample size (ESS) quantifies the information loss due to weighting and is computed as  $ESS = (\sum_i w_i)^2 / \sum_i w_i^2$ . When all weights are equal, ESS equals the sample size; extreme weights reduce ESS substantially.

## 2.12 Statistical Analysis Plan

Performance metrics will be computed for each method within each scenario, along with Monte Carlo standard errors. Methods will be compared using the following analyses:

Overall performance will be summarized by averaging metrics across all scenarios for each method. This provides a global assessment of method performance but may obscure important heterogeneity.

Performance by scenario characteristics will be examined by stratifying results by sample size, population overlap, and number of effect modifiers. This analysis will identify whether certain methods are preferable under specific conditions.

The equivalence of SigTotal and EbTotal will be assessed by computing the maximum absolute difference in point estimates across all simulation replications. A difference below numerical precision ( $10^{-8}$ ) will confirm the theoretical equivalence.

Graphical displays will include boxplots of bias, RMSE, and coverage by method; heatmaps showing mean bias across scenario configurations; and scatterplots illustrating the bias-ESS tradeoff.

## 2.13 Software and Reproducibility

All analyses will be conducted using R statistical software (version 4.3 or later). The simulation code, weighting method implementations, and analysis scripts will be made publicly available in the `advmaic` R package, hosted on GitHub at <https://github.com/choxos/advmaic>.

Random number generation will be controlled using `set.seed()` with documented seed values to ensure full reproducibility. The simulation will use parallel processing to reduce computation time, with the number of cores detected automatically.

## 3 Timeline

The study will proceed according to the following timeline: Protocol development and registration in Month 1; R package development and testing in Months 2-3; Pilot simulation and debugging in Month 4; Full simulation execution in Month 5; Results analysis and manuscript preparation in Months 6-7; and Manuscript submission in Month 8.

## 4 Ethics Statement

This simulation study uses only simulated data and does not involve human subjects, patient data, or any identifiable information. No ethics approval is required.

## 5 Funding

This study received no external funding.

## References

Bucher, Heiner C, Gordon H Guyatt, Lauren E Griffith, and Stephen D Walter. 1997. "The Results of Direct and Indirect Treatment Comparisons in Meta-Analysis of Randomized Controlled Trials." *Journal of Clinical Epidemiology* 50 (6): 683–91.

- Cressie, Noel, and Timothy RC Read. 1984. "Multinomial Goodness-of-Fit Tests." *Journal of the Royal Statistical Society: Series B (Methodological)* 46 (3): 440–64.
- Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20 (1): 25–46.
- Owen, Art B. 2001. *Empirical Likelihood*. Chapman; Hall/CRC.
- Petto, Hwee, Zbigniew Kadziola, Alan Brnabic, et al. 2019. "Alternative Weighting Approaches for Anchored Matching-Adjusted Indirect Comparisons via a Common Comparator." *Value in Health* 22 (1): 85–91.
- Phillippo, David M, Sofia Dias, A E Ades, and Nicky J Welton. 2020. "Equivalence of Entropy Balancing and the Method of Moments for Matching-Adjusted Indirect Comparison." *Research Synthesis Methods* 11 (4): 568–72.
- Signorovitch, James E, Vanja Sikirica, M Haim Erder, et al. 2012. "Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research." *Value in Health* 15 (6): 940–47.
- Signorovitch, James E, Eric Q Wu, Andy P Yu, et al. 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 (10): 935–45.
- Williamson, Elizabeth J, Andrew Forbes, and Ian R White. 2014. "Variance Reduction in Randomised Trials by Inverse Probability Weighting Using the Propensity Score." *Statistics in Medicine* 33 (5): 721–37.