How robust are doubly-robust estimators in limited data indirect comparisons

Journal Title
XX(X):1–21

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Chengyang Gao

Abstract

Keywords

Introduction

There has been growing interest in adopting doubly robust (DR) estimators in population-adjusted indirect comparisons (PAIC) with constrained access to individual participants data (IPD) in HTA[1,2]. This trend is primarily driven by an increasing awareness of potential model misspecification issues associated with standard parametric models. DR estimators[3–5], well-established in causal inference literature, offer an appealing solution by providing analysts with 'two chances to get the model correct' - population-adjusted estimates remain consistent when either the outcome or trial assignment model is correctly specified.

A perhaps more important property of the DR estimator lies in its convergence rate[3,6,7] - the DR estimator converges at the product rate of the two nuisance functions (the outcome model and the trial assignment model in this context). The product rate property can be especially appealing, as it allows for nonparametric estimation of nuisance functions—reducing the risk of model mis-specification — while maintaining the parametric $\sqrt{n}$ - convergence rate of the

Corresponding author:

Email: chengyang.gao.15@ucl.ac.uk

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

target parameter.However, the contextual differences between PAICs in HTA and the settings commonly discussed in causal inference literature can limit how the theoretical doubly-robustness properties translate into empirical performance.

One key difference lies in the data accessibility constraints typical in HTA. More often than not, analysts have access to the IPD from their own sponsor's trial, but only aggregate-level data (ALD) for their competitor's trial. In the absence of patient-level data from one trial arm, the trial assignment mechanism cannot be modeled directly through a likelihood function. While matching-adjusted indirect comparisons (MAIC)[8,9] circumvents this challenge by recasting it as an optimization problem, it also limits the class of trial assignment models to parametric logistic regressions, thus remaining susceptible to model misspecification. More importantly, when faced with non-collapsible outcomes,Beyond model mis-specification, PAICs in HTA can suffer from 'population mis-specification bias': since the ALD population must be reconstructed from published summary statistics, any inaccuracy in this reconstruction means the comparison is effectively conducted in the wrong target population. While an obvious problem when stated in plain language, previous explorations have found that mis-specifying the target population covariate structure has minmial impacts on bias[10,11].

In this manuscript, we explore the applications of DR estimator to an anchored indirect comparison where IPD is only available in one of the trials. By comparing their performances against existing population adjustment approaches, we attempt to provide a preliminary examination of the robustness of DR estimators in the context of limited data PAICs.

We begin by reviewing existing population adjustment methods for limited data PAICs, before demonstrating how to adapt two popular DR estimators - augmented inverse propensity weighting (AIPW) estimator[3] and target maximum likelihood (TMLE) estimators[12,13] to the limited data context. We then turn our focus to the problem of population mis-specification, starting from the larger, but subtle issue of estimand in pairwise PAICs. Building on existing experimental findings regarding population mis-specification in the literature we conduct a comprehensive simulation study comparing the performances of DR estimators against the standard MAIC and outcome modelling (OM) based approaches.

Through this comprehensive simulation study, we show that DR estimators behave more like OM based estimators under limited data constraints. The problem of population mis-specification, can sometimes have a bigger impact on the performance of population adjustment. When the 'source populations' are sufficiently distinct, even mis-specifying the correlation structure of the target population, a relatively mild form of mis-specification, can result in large bias and under-coverage.

Doubly robust estimators in indirect comparisons with data accessibility constraints

Nuisance functions and outcome modelling

We begin by introducing nuisance models: statistical constructs that, while not of direct inferential interest, serve as helpful components in estimating the target quantity [14]. In PAICs, the typically causal estimand is the marginal treatment effects in the competitor's trial. The outcome model, which describes the outcome-covariate relationship, and the trial assignment model, which describes individual's probability of inclusion in the IPD population, function as nuisance models. These models, though not primary inferential targets, are crucial in constructing the estimand. DR approaches, as mentioned earlier, encompass a class of methods that maintain consistency when either of these nuisance models is correctly specified.

Before moving onto DR approaches, we take a look at OM approaches, also known as parametric G-computation [15] or simulated treatment comparisons (STC) [11,16], which focus solely on the outcome models. This class of approaches first model the outcome-covariate relationship using the IPD. The fitted model is then used to predict the individual-level outcomes under either treatment conditions, had the treatment being applied in the ALD population. These predicted outcomes are subsequently aggregated and transformed to estimate marginal treatment effects on the scale of interest. Conceptually, the application of OM in data-constrained scenarios mirrors its use when full IPD are available. However, a key distinction arises in the marginalization step for non-collapsible outcomes: due to the lack of IPD for the competitor's trial, marginalization of the conditional outcome model must be performed over a simulated ALD population, necessitating certain parametric assumptions.

The OM approach has gained traction in PAIC applications due to its extrapolation capabilities, flexibility, and potential for efficiency under correct model specifications. However, it is crucial to note that the validity of OM methods hinges entirely on the correct specification of the outcome model. This reliance on a single nuisance model introduces a trade-off between potential efficiency and robustness, a consideration that becomes particularly salient when comparing OM to doubly robust approaches.

Exisiting implementations of DR estimators in PAICs

The use of doubly robust estimators in data-limited indirect comparisons is, to some extent, well-established. The first and still most popular population adjustment method, MAIC [1,8,9]

is doubly robust when the linearity of the outcome model and the log-linearity of the trial assignment model holds at the same time.

Initially introduced as a method of moments (MoM) estimator to the logistic regression model for the trial assignment, the DR property has gradually gained attention as the mathematical equivalence between MAIC and entropy balancing gets established[17,18]. Nevertheless, the conditions required for the DR property are usually violated in the context of HTA: for binary or time to event outcomes that are commonly of interest, it is the transformed rather than the natural outcomes can be assumed to be linear in covariates. Consequently, the validity of MAIC still rests on the logistic regression model for assignment, and should just be considered as a weighting method in these context.

Recent interest has focused on more direct implementation of DR approaches - by augmenting MAIC with a regression model. A recent example is from Park et al.[2], where a weighted regression model is fitted to the IPD by combining MAIC weights with a standard regression model. The fitted model is further used to predict the outcomes had the treatment being trialed in the ALD population. The comparison stage then just follows as in parameteric G-computation. As demonstrated in the ISPOR 2024 workshop, this DR approach is especially useful when the optimisation in MAIC fails - when the aggregated-level summaries in the ALD population fall outside the convex hull formed by the IPD[19].

This 'weighted parameteric G-computation' implementation of DR estimator is the third version of the AIPW estimator explained in Dahabreh et al.[20] adapted to the limited data context. More specifically, the normalized MAIC weights serve as the MoM estimation of the stabilized trial assignment odds. Indeed, when the true trial assignment model follows a logistic regression model, the MAIC weights converge to the true trial assignment odds up to a constant[1,9]. Based on this insights, as will be discussed below, two other popular DR methods - the standard AIPW approach with normalized weights, as well as the TMLE approach can be adapted to the data constrained context.

AIPW and TMLE in limited data PAICs

Settings and notations

The problem of PAIC in HTA is no different to transporting causal effects from experiments to target population except for the limited data constraints. Here we set up PAIC under the standard target population causal inference framework. To start, we adopt a 'super-population' causal inference perspective - the IPD and the ALD trials can be viewed as simple random

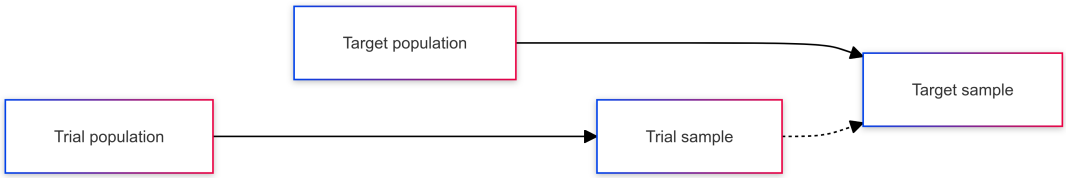

Figure 1. Population diagram for transporting experimental results from trial to target. Under the standard super-population inference framework, the trial population and the target population refer to the underlying super-population that generate the trial and the target sample. The dashed arrow from trial sample to target sample denotes the goal of transporting experimental results to the target sample

samples from distinct super-population (source population). A rough depiction of this setup is shown in Figure 1.

For N^{IPD} units in the IPD trial and N^{ALD} units in the ALD trial, we use the triplet $\{Y_i, X_i, Z_i\}$ to denote individual-level outcomes, covariates and treatment indicator. The covariates can be further divided into covariates X into purely prognostic variables (PV) and effect modifiers (EMs). We introduce a sampling indicator S, where the IPD population are index via $\{i : S_i = 1\}$, and the ALD population are indexed using $\{i : S_i = 0\}$.

Under the conditional constancy of relative effects assumption, the treatment effects in the IPD trial can be transported to the ALD trial after adjusting for the distribution of EMs. With both IPDs available, we could transport treatment effects by reweighting Y_i^{IPD} by their inverse odds of sampling weight (IOSW):

$$w_i^s = \begin{cases} \frac{1-\hat{\rho}(X_i)}{\hat{\rho}(X_i)\hat{\pi}}, & \text{when } S_i = 1, Z_i = 1, \\ \frac{1-\hat{\rho}(X_i)}{\hat{\rho}(X_i)(1-\hat{\pi})}, & \text{when } S_i = 1, Z_i = 0, \\ 0, & \text{when } S_i = 0, \end{cases} \tag{1}$$

where $\rho(X_i) = P(S = 1 \mid X_i)$ denotes probability of sampling into the IPD trial (sampling model thereafter) and $\hat{\pi}$ is the estimated propensity score (PS). For an RCT, the PS can be optional. Under limited data constraint, while the sampling model cannot be estimated using standard maximum likelihood method, but as mentioned before, we could recast the estimation as an optimisation problem and still estimate the sampling odds (trial assignment odds) up to a normalizing constant using MoM as in MAIC. This serves as the basis for constructing DR estimators in the data constrained context.

AIPW formulation

Under standard implementation of AIPW, the fitted outcomes on the natural scale under either treatment conditions, $\hat{Y}(1), \hat{Y}(0)$, can be written as:

$$\begin{aligned}\hat{Y}(1) &= \sum_{\{i:S_i=1\}} \frac{w_i^s Z_i[Y_i - \hat{\mu}_1(X_i)]}{\sum_{\{i:S_i=1\}} w_i^s Z_i} + \frac{1}{n_0} \sum_{\{i:S_i=0\}} \hat{\mu}_1(X_i) \\ \hat{Y}(0) &= \sum_{\{i:S_i=1\}} \frac{w_i^s (1-Z_i)[Y_i - \hat{\mu}_0(X_i)]}{\sum_{\{i:S_i=1\}} w_i^s (1-Z_i)} + \frac{1}{n_0} \sum_{\{i:S_i=0\}} \hat{\mu}_0(X_i),\end{aligned} \tag{2}$$

where $\hat{\mu}_1(X_i), \hat{\mu}_0(X_i)$ denote the fitted conditional outcome means under OM approach alone.

In the above two part sum, the first part of the equation only depends on the data in the IPD trial, while the second part is the same as in standard G-computation. Therefore, we could simply plug-in the MoM estimation of w_i^S. After the standard normalization step, we could exactly recover the normalized sampling weights have we had the full IPD.

Notice the above AIPW formulation can be applied for outcomes on any scale. Nevertheless, the estimation can be unstable for bounded, especially binary outcomes, as the two part sum above may fall outside natural range.

TMLE formulation

To address issues with bounded outcomes, the Targeted Maximum Likelihood Estimation (TMLE) approach[12,13] offers a good alternative for constructing DR estimators. Like outcome regression (G-computation) based approaches, TMLE is a 'plug-in' estimator, where we 'plug in' the desired target population distribution into the final regression function to calculate the target quantity.

The TMLE approach directly targets a 'plug-in' based on the efficient influence function. Operationally, this is achieved by updating a first-stage regression-based estimates of potential outcome, $\hat{\mu}_0^*(S_i = 1, \mathbf{X}_i)$ and $\hat{\mu}_1^*(S_i = 1, \mathbf{X}_i)$ on the logit scale based on the equation below:

$$\begin{aligned}\hat{\eta}_0^*(S_i = 1, \mathbf{X}_i) &= g^{-1}\left\{\hat{\mu}_0^*(S_i = 1, \mathbf{X}_i) + \hat{\epsilon}_0 \frac{[1 - \hat{p}(\mathbf{X}_i)]}{\hat{\rho}(\mathbf{X}_i)[1 - \hat{\pi}_1(\mathbf{X}_i)]}\right\}, \\ \hat{\eta}_1^*(S_i = 1, \mathbf{X}_i) &= g^{-1}\left\{\hat{\mu}_1^*(S_i = 1, \mathbf{X}_i) + \hat{\epsilon}_1 \frac{[1 - \hat{p}(\mathbf{X}_i)]}{\hat{p}(\mathbf{X}_i)\hat{\pi}_1(\mathbf{X}_i)}\right\}.\end{aligned} \tag{3}$$

Fluctuation parameters $\hat{\epsilon_0}$ and $\hat{\epsilon_1}$ are estimated by regressing the transformed outcomes Y_i against the 'clever covariates' $\hat{h}_0(\mathbf{X}_i)$ and $\hat{h}_1(\mathbf{X}_i)$ shown below with the initial estimates as offset.

$$\hat{h}_0(\mathbf{X}_i) = \frac{(1-Z_i)[1-\hat{p}(\mathbf{X}_i)]}{\hat{p}(\mathbf{X}_i)[1-\hat{\pi}_1(\mathbf{X}_i)]}, \quad \hat{h}_1(\mathbf{X}_i) = \frac{Z_i[1-\hat{p}(\mathbf{X}_i)]}{\hat{p}(\mathbf{X}_i)\hat{\pi}_1(\mathbf{X}_i)}.$$

The final causal effect estimates are based on updated estimates of individual-level potential outcomes. The formula for causal contrasts on the natural scale is:

$$\hat{\tau}_{\text{TMLE}} = \frac{1}{n_0} \sum_{\{i:S_i=0\}} \left[\hat{\eta}_1^*(S_i=1,\mathbf{X}_i) - \hat{\eta}_0^*(S_i=1,\mathbf{X}_i)\right].$$

This formulation can be easily adapted for effect measures on other scales.

In the limited data context, 'clever covariates' can be calculated up to a normalizing constant, similar to AIPW. Nevertheless, two key differences emerge: First, the updating step requires estimates of 'raw' trial assignment odds. Scaling by any constant, while providing consistent estimates, could affect convergence rates and introduce instabilities. Second, unlike AIPW, updated estimates require trial assignment odds in the target population. While this seemingly introduces additional assumptions, we argue that these assumptions are no stronger than those in standard parametric G-computation. Once the joint distribution is specified, constructing trial assignment odds simply involves plugging the joint distribution into the implied trial assignment model.

Additional discussions

While the discussions above may have made the construction of DR estimators in the limited data context seems like a 'free lunch', some subtle changes have made our DR estimators less 'robust' than the standard DR estimators. To start with, we no longer have the flexibility to estimate the trial assignment model non-parametrically, the model class has been restricted to a logisitic regression model, and therefore the potential for model mis-specification is much higher. Furthermore, since the unscaled trial assignment odds are not estimable under limited data, we cannot estimate the variances of the estimators based on the empirical influence function. We discuss this in more details in the appendix.

The alternative here would be to use non-parametric bootstrap, which would not only be computationally intensive, but would also introduce additional instabilities when the sample size is small. Finally, an obvious but somehow under-discussed problem of DR approaches in the limited data context is the potential for 'population mis-specification' when estimating the marginal estimands for non-collapsible outcomes. We discuss this part in details from the estimand angle in the section below.

Target estimand and population mis-specification

For non-collapsible outcome measures such as odds ratios or hazard ratios, targeting marginal estimands using regression-based methods requires marginalizing the conditional outcome model over the target population[15,21]. However, this marginalization raises subtle but important questions about the nature of the target population—specifically, whether we are targeting effects in the observed sample (target sample average treatment effects, TSATE) or in the broader population (target population average treatment effects, TATE).

This distinction can be formalized within the potential outcomes framework: TSATE is a function of the realized potential outcomes in the sample, while TATE is a function of the parameters governing both the conditional outcomes and the joint distribution of covariates in the target population. This subtle distinction is explored by Josey et al.[18], with the conclusion being both their proposed calibration estimation methods and MAIC can consistently target TSATE or TATE, but valid inferences can only be drawn for TSATE not TATE. A plausible explanation for this is that using target sample mean in place of the target population mean ignores the sampling variability and therefore under estimate variance of the estimator.

In the context of PAICs, TSATE is typically of primary interest as it aligns with the published estimates from ALD trials. The standard practice in parametric G-computations is to marginalize the conditional outcome model over a 'sufficiently large' parametrically specified population[11,15,17]. While this might appear to target TATE, the specified population should be understood not as the original source population of the ALD trial, but rather as a "pseudo" source population centered on the published ALD summaries. The multiple imputation approach proposed by[22] targets a sample estimand drawn from this pseudo population. Thus, the marginalization procedures in PAIC literature are best viewed as a conservative approach to targeting TSATE. For a deeper discussion of finite population versus super-population estimands, we refer readers to[23].

However, reliance on parameterically specified population based on aggregate-level summaries introduces potential 'population mis-specification bias'. While the joint covariate distribution could be flexibly specified by combining marginal distribution for each EM and normal copula, population mis-specification can still arise from structural differences in dependency structure between ALD and IPD populations - often due to distinct inclusion or exclusion criteria or the non-concurrency of IPD and ALD trials. While previous exploratory work has found that dependency structure to have minimal impacts[11,17], these findings may be limited to scenarios with structurally similar data-generating processes (e.g., multivariate Normal distributions with different locations) or modest effect heterogeneity.

Besides, a more severe form of 'population mis-specification' when some of the marginal distributions, particularly in the tails, are mis-specified due to finite sample sizes. In that case, the potential for bias would obviously be higer. We focus on the 'better scenario' where only dependency structure is mis-specified and conduct a comprehensive simulation study to compare the performances of the DR as well as standard OM estimators in the sections below.

Simulation studies

Following the guidance from Morris et al. (2019) we lay out our simulation set up using the ADEMP framewok (Aim, Data-generating mechanisms, Estimand, Method and Performance measures)

Aim

This simulation study aims to explore the robustness of PAIC methods under mis-specifications of the target population dependency structure. In particular, we focus on performances of the class of DR estimators, and examine whether they provide additional robustness gains against standard regression or weighting based population adjustment methods under varying sample size and population overlap.

Data-generating mechanisms

We consider a binary outcome, generated under a logit model:

$$\mathrm{E}(Y \mid \mathbf{X}, \mathbf{T}) = g^{-1}\left(\beta_0 + \beta_{\mathrm{PV}}\mathbf{X}^{\mathrm{PV}} + \left(\beta_{trt} + \beta_{\mathrm{EM}}\mathbf{X}^{\mathrm{EM}}\right)\mathbf{T}\right), \tag{4}$$

where $g(\cdot)$ is the logit link function. The two trials are AC and BC, involving three treatment A, B, C with C as the common comparator.

We first generate IPD for both trials, with the AC trial designated as the IPD trial. The IPD consist of individual-level outcomes $\mathbf{Y}$, the treatment assignment status $\mathbf{T}$ and a matrix of five covariates $\mathbf{X}$. All covariates are assumed to be both prognostic variables and effect-modifiers, i.e. $\mathbf{X}^{EM} = \mathbf{X}$ in the logit model. The BC trial is the ALD trial so the simulated IPD are aggregated to obtain covariate summaries; N_B, N_C denote the sample sizes and n_B, n_C represent the total event counts in arms B and C, respectively. We fix the size of the ALD trial to be the same as the IPD trial with a $1:1$ treatment allocation ratio.

The model for the probability of experiencing the outcome is:

$$\text{logit}(\theta_i) = \beta_0 + \beta_{\mathbf{1}}\mathbf{X}_{\mathbf{i}}^{\text{EM}} + (\beta_{trt} + \beta_{\text{EM}}\mathbf{X}_{\mathbf{i}}^{\text{EM}})T_i, \tag{5}$$

where β_0 is the fixed control group event rate. The strength of effect-modification $\beta_{\mathbf{EM}}$ is assumed to be the same under the assumption of shared effect-modifiers, and we set $\beta_{EM} = -\log(0.6)$, corresponding to strong effect-modification. The prognostic strength β_1 is also assumed to be the same and set to $\beta_1 = -\log(0.8)$. The treatment effects are assumed to be equally large in both trials with $\beta_{AC} = \beta_{BC} = \beta_{trt} = \log(0.25)$. The treatment effects correspond a decrease in baesline odds by 75% but with strong effect-modification, the sign of the treatment can still be flipped.

To assess the impact of population mis-specification, we generate the covariates for AC and the BC trial under structurally different processes. We consider three continuous and two binary covariates that make up the covariate matrix $\mathbf{X}$. To create the structural difference, we generate the covariates using different parametric family and transformation, such that the dependency structure can be arbitrarily different between the two trials. We detail this generation process in the appendix. Under this covariate generating process, we further set the sample sizes of both trials at 100, 200, or 600. We generated moderate and poor population overlap scenarios by changing the location and scale parameters of covariate distributions, such that they roughly correspond to average ESS reductions of 55% and 75%Ḟactorial combinations of sample sizes and population overlap settings yield 6 scenarios. Detailed parameter configurations are given in the appendix.

Estimand

The estimand of interest is the A vs B marginal treatment effect in the BC population. Based on the current setup, the true A vs B effect is zero.

Methods

The following methods will be compared:

- MAIC
- Bayesian parametric G-computation from Remiro et al[15].
- standard AIPW
- AIPW using weighted regressions
- TMLE

It is worth pointing out that all methods are mis-specifying the target population, either implicitly or explicitly. Even for MAIC, the re-weighting of IPD data, viewed from the perspective of density estimation, approximate ALD trial population with an exponentially tilted population of the IPD. Under the structurally different DGPs considered in this study, this is asymptotically different. For the other four regression-based methods, we also consider marginalizing the conditional outcome models over the true BC population to generate the **'oracle'** estimates.

Performance Measures

We simulate N_{sim} datasets for each scenario and estimate the treatment effect $\hat{d}_i$ of A vs B in the BC population. The true value of this effect d is zero. The bias of the population-adjusted estimator is the expected difference between the estimated value and the truth. Variability is measured using empirical standard errors, (the standard deviation of $\hat{d}_i$ across all repetitions), model average standard errors (the average of the standard error associated providing by the estimating procedure $\hat{\sigma}_{i,mod}$ If these two measures are close, this reflects that the variance estimator is stable. Finally, coverage is the proportion of 95% confidence intervals that contain the true difference using Wald-type confidence intervals with $\hat{d}_{upper(lower),i} = 1.96 \times \hat{\sigma}_{i,mod} \pm \hat{d}_i$. These measures can be calculated as:

- Bias $= \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} (\hat{d}_i - d)$,
- Empirical standard error $= \sqrt{\frac{1}{N_{sim}-1} \sum_{i=1}^{N_{sim}} (\hat{d}_i - d)^2}$,
- Model average standard error $= \sqrt{\frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \hat{\text{var}}(\hat{d}_i)}$,
- Coverage $= \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \mathbb{1}(\hat{d}_{lower,i} \leq \hat{d}_i \leq \hat{d}_{upper,i})$.

To determine N_{sim}, the Monte Carlo standard errors (MCSE) of the performance measures should be low relative to the estimates. Since the bias and precision are of primary interests, we base the number of repetitions on their MCSE:

$$\text{MCSE}_{bias} = \sqrt{\frac{\hat{\text{var}}(\hat{d}_i)}{N_{sim}}} \qquad \text{MCSE}_{EmpSE} = \sqrt{\frac{\hat{\text{var}}(\hat{d}_i)}{2(N_{sim}-1)}}$$

We consider 2000 Monte Carlo replications for our analysis.

Results

All methods exhibit stable performances under a sample size of 600 and 200 with moderate population overlap. For the rest of scenarios, however, the AIPW and TMLE approach exhibit

estimation failure in some of the cases. This is mostly expected for the AIPW approach, as the predicted outcomes may fall outside the natural covariate range under extreme weights, which is inevitable under small sample size and poor population overlap. The estimation failures in TMLE are slightly surprising at first glance as it does not suffer from the problem of unbounded outcomes. After a closer examination, however, it becomes clear that TMLE also suffer from extreme weights, particularly at the updating step: in some of the bootstrapped dataset, extreme weights combined with resampling can cause near perfect separation of the outcomes, leading to estimation failures when regressing the outcomes against the 'clever covariates'.

All the problematic results were considered as missing during the analysis. We think similar problems are unlikely to occur in reality - the smallest sample with an average ESS reduction of 75% is mostly aimed as a stress test.

For a comprehensive bias assessment, we calculate the bias of all methods for both standard and **'oracle'** sceanrios and present the results under a single plot. By doing this, we make sure the resulting bias is truly due to population mis-specification, rather than small sample bias or even improper implementations of different methods. We present the bias plot in figure 2:

As shown in the plot, there seems to be substantial bias even by mis-specifying correlation structure. This bias increases as sample sizes become smaller. Interestingly, at least in terms of the point estimates, the magnitude of bias is larger under moderate overlap scenarios than that under poor population overlap for regression-based methods. Whereas for MAIC, the opposite can be true under a sample size of 100. This is likely due to the fact that weighting and regression-based approaches are approximating the target distribution in different ways. Under the smallest sample size, MAIC suffer from the unstable weights more, and therefore incur more small sample bias.

By comparing the results against the 'oracle' estimates, we indeed confirm that most of the bias is attributable to population mis-specification. There are some non-negligible small sample bias, but it pales in comparison to the overall bias when mis-specifying the correlation structure.

For DR estimators, they behave similarly to OM approaches (Bayesian G-computation in this case). This is mostly due to the fact all outcome models are correctly specified in our simulation study, and both OM and DR approaches suffer from population mis-specification. DR estimators might offer superior performances when the outcome models are mis-specified, but neither can we rule out the case of bias amplification - the mis-specification of outcome and trial assignment models might exacerbate the bias in each, particularly under extreme weights. This needs to be further explored in future studies. Nevertheless, in addition to the potential benefits in robustness, the DR estimators, especially TMLE and AIPW, does seem to offer a slight edge in bias in small samples when marginalizing over the correct target population. This does

make sense. With the trial assignment model being mis-specified, the one-step adjustment factor in AIPW or the 'clever covariate' in TMLE mostly serves as a dimensional reduced summary of the covariate that might offer slight help in adjusting residual variability.

		Method				
Sample size	BC location	MAIC	AIPW	OM	TMLE	Weighted DR
Moderate population overlap						
100	0.2	0.952	0.953	0.940	0.951	0.953
200	0.2	0.940	0.942	0.944	0.943	0.940
600	0.2	0.929	0.923	0.941	0.926	0.925
Poor population overlap						
100	0.5	0.982	0.957	0.941	0.952	0.956
200	0.5	0.934	0.933	0.932	0.940	0.924
600	0.5	0.930	0.936	0.949	0.942	0.931

Coverage values are reported for each method and scenario combination.

[1] MAIC: Matching-Adjusted Indirect Comparison

[2] AIPW: Augmented Inverse Probability Weighting

[3] OM: Outcome modelling using Bayesian G-computation

[4] TMLE: Targeted Maximum Likelihood Estimation

[5] Weighted DR: Doubly Robust estimation with weighted regressions

Table 1. Coverage probabilities of different population adjustment methods when marginalizing over the ALD population with mis-specified dependency structure

Nevertheless, when marginalising over the wrong target population, the performance gain in point estimation is minimal compared to the overall magnitude of bias. Worse still, as shown in table 1, the DR methods sometimes exhibit worse coverage compared to OM based methods, especially for the two different implementations of AIPW estimators. To further investigate this under-coverage problem, we calculate the bias-eliminated coverage shown in table 2. It seems that the under-coverage in TMLE estimator is mostly driven by bias, as nominal coverage is almost restored after correcting for the bias in point estimate. Meanwhile, the under-coverage in the two AIPW estimators persists, especially under poor population overlap. A further check in standard error estimation shows that the non-parameteric bootstrap tend to under-estimate the true variance in both cases. We attach the table in the appendix.

Overall, among all DR estimators, we find that the TMLE implementation proposed in this manuscript, can be superior to the previous weighted regression implementation, both in terms of coverage and efficiency. Regardless marginalizing over the true or wrong population, the TMLE achieves smaller empirical standard errors, with performances staying rather close to those based

on OM alone. Considering this small efficiency loss, one might find the additional improvements in robustness worthwhile.

		Method				
N_sample	BC_location	MAIC	AIPW	OM	TMLE	Weighted DR
Moderate population overlap						
100	0.2	0.956	0.957	0.948	0.956	0.959
200	0.2	0.950	0.948	0.952	0.954	0.952
600	0.2	0.949	0.944	0.957	0.948	0.942
Poor population overlap						
100	0.5	0.979	0.959	0.944	0.949	0.961
200	0.5	0.934	0.933	0.947	0.949	0.927
600	0.5	0.939	0.943	0.961	0.954	0.938

Bias-eliminated coverage values are reported for each method and scenario combination.

[1] MAIC: Matching-Adjusted Indirect Comparison

[2] AIPW: Augmented Inverse Probability Weighting

[3] OM: Outcome modelling using Bayesian G-computation

[4] TMLE: Targeted Maximum Likelihood Estimation

[5] Weighted DR: Doubly Robust estimation with weighted regressions

Table 2. Bias-eliminated coverage probabilities of different population adjustment methods when marginalizing over the ALD population with wrong dependency structure

Discussion

This chapter presents an exploratory investigation into the robustness of DR estimators in PAICs under data accessibility constraints. We proposed two novel implementations of common DR estimators and evaluated their performance against standard weighting and OM methods under population mis-specification. Our results demonstrate that DR approaches are not immune to population mis-specification bias, as with all other methods for PAIC. Contrary to previous findings, substantial bias was observed even when mis-specifying only the correlation structure in the ALD population. This discrepancy from prior research can be attributed to the structurally distinct data-generating process, the increased number of effect modifiers as well as the strength of effect modification in our simulation setup.

From a practical point of view, some may argue that this combination of effect-modifiers and effect-modification is rare on the log-odds ratio scale. We think this is a valid critic. But in practice, it is not uncommon to see outcome models that include numerous prognostic covariates, and for a non-collapsible effect measure, differences in the distribution of these variables across trials will alter the marginal effect, even when the conditional effect is constant [24].

From an estimand perspective, this finding aligns with theoretical expectations: under population mis-specification, any population-adjustment method inherently targets an incorrect marginal estimand. When effect modification is strong, bias in population-adjusted estimates becomes inevitable.

Conceptually, this bias parallels the well-documented issue in MAIC, where varying the availability of IPD and ALD can lead to divergent population-adjusted estimates or even conflicting conclusions in pairwise comparisons. However, the magnitude of bias from population mis-specification depends on the discrepancy between the true and the parametrically-specified covariate distribution of the target population. This bias is expected to be more modest than that seen in MAIC, because the error from mis-specifying the correlation structure can be smaller than the structural differences that can exist between two distinct clinical trial populations

Nevertheless, Despite their susceptibility to population mis-specification, DR properties remain helpful in the context of data constrained PAICs. When marginalizing over the true ALD population, DR approaches yield unbiased estimates even when the implied trial assignment model from MAIC is mis-specified. Between the two newly proposed adaptations of DR estimators, AIPW exhibited greater variability and reduced efficiency, particularly under small ESS. This is primarily due to the instability introduced by inverse weighting in its formulation as shown in (2). Conversely, the proposed TMLE implementation emerged as the most efficient DR estimator in this simulation study. As a substitution estimator, its efficiency nearly matched that of the OM approach. Importantly, the TMLE implementation does not impose additional assumptions beyond those required by other DR approaches, with its 'clever covariates' fully determined by the MAIC weights and specified ALD population.

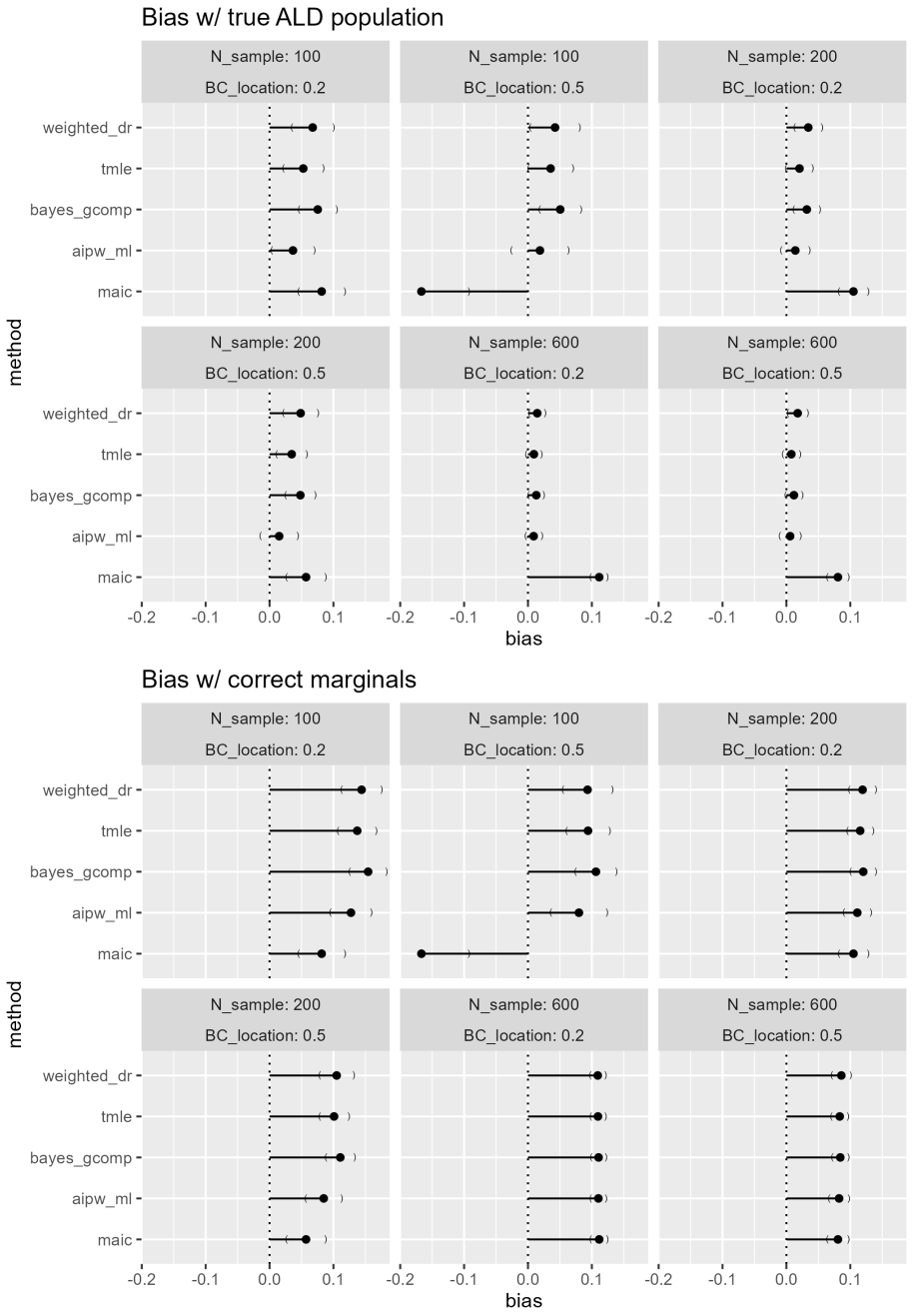

Figure 2. Bias across scenarios when marginalizing true ALD population vs marginalizing over simulated ALD with correct marginal distributions

The implications of correlation structure mis-specification extend well beyond pairwise comparisons into more complex evidence synthesis frameworks. Multi-level network meta-regression[10,25], the current gold standard for synthesizing evidence from mixed ALD and IPD trials in larger networks, fundamentally relies on correlation structures derived from IPD trials. Particularly in ALD trials, the derived correlation structure, along with published summary statistics and assumed marginal covariate distributions is used to reconstruct the ALD population for numerically integrating the individual-level likelihood. Our preliminary findings on the substantial bias induced by mis-specified dependency structures raise important methodological considerations for evidence synthesis in networks containing both IPD and ALD trials - selecting appropriate correlation structure for reconstructing the ALD population would be crucial to minimize bias given the included trial populations could be structurally different due to temporal and geographical differences.

Finally, It is crucial to note that our comparison of DR approaches to standard weighting and OM methods was limited to mis-specification of the trial assignment model, assuming correct specification of the outcome model. Under these conditions, the similar performance of DR and OM approaches is unsurprising. It is highly likely that DR approaches outperform OM under mis-specified outcome models. Meanwhile, as we briefly alluded to before, the potential for 'bias amplification' can be just as likely under outcome model mis-specifications, given that MAIC's assumed logistic regression model for trial assignment would be mis-specified to some degree when covariate structures in IPD and ALD are complex and structurally different. For a more comprehensive examination of how DR approaches compare to OM under outcome model mis-specification, future studies should consider a range of DGPs for the IPD and ALD population to encompass different degrees of violations in the implied trial assignment model.

In conclusion, this manuscript introduces two novel implementations of DR estimators in the data-constrained PAIC context, with the proposed TMLE implementation demonstrating superior efficiency among DR approaches. Our findings underscore that DR approaches, like standard OM methods, are not immune to population mis-specification. Ultimately, the limited accessibility to IPD in the competitor's trial creates a missing data problem, where the method of 'imputing' the missing IPD may be equally, if not more, crucial than the models for population-adjusted outcomes.

References

1. Cheng D, Tchetgen ET and Signorovitch J. On the double-robustness and semiparametric efficiency of matching-adjusted indirect comparisons. *Research Synthesis Methods* 2023; 14(3): 438–442.
2. Park JE, Campbell H, Towle K et al. Unanchored population-adjusted indirect comparison methods for time-to-event outcomes using inverse odds weighting, regression adjustment, and doubly robust methods with either individual patient or aggregate data. *Value in Health* 2024; 27(3): 278–286.
3. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 1994; 89(427): 846–866.
4. Bang H and Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; 61(4): 962–973.
5. Scharfstein DO, Rotnitzky A and Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 1999; 94(448): 1096–1146.
6. Kennedy EH. Semiparametric theory and empirical processes in causal inference. *Statistical causal inferences and their applications in public health research* 2016; : 141–167.
7. Chernozhukov V, Chetverikov D, Demirer M et al. Double debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 2018; 21.
8. Signorovitch JE, Sikirica V, Erder MH et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value in Health* 2012; 15(6): 940–947.
9. Cheng D, Ayyagari R and Signorovitch J. The statistical performance of matching-adjusted indirect comparisons: Estimating treatment effects with aggregate external control data. *The Annals of Applied Statistics* 2020; 14(4): 1806–1833.
10. Phillippo DM. *Calibration of treatment effects in network meta-analysis using individual patient data*. PhD Thesis, University of Bristol, 2019.
11. Ren S, Ren S, Welton NJ et al. Advancing unanchored simulated treatment comparisons: A novel implementation and simulation study. *Research Synthesis Methods* 2024; 15(4): 657–670. DOI: 10.1002/jrsm.1718.
12. Van Der Laan MJ and Rubin D. Targeted maximum likelihood learning. *The international journal of biostatistics* 2006; 2(1).
13. Gruber S and van der Laan MJ. A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome. *The International Journal of Biostatistics* 2010; 6(1). DOI: 10.2202/1557-4679.1260.
14. Daniel RM. Double robustness. *Wiley StatsRef: Statistics Reference Online* 2014; : 1–14.

15. Remiro-Azócar A, Heath A and Baio G. Parametric g-computation for compatible indirect treatment comparisons with limited individual patient data. *Research synthesis methods* 2022; 13(6): 716–744.
16. Ishak KJ, Proskorovsky I and Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics* 2015-06; 33(6): 537–549. DOI:10.1007/s40273-015-0271-1.
17. Phillippo DM, Dias S, Ades AE et al. Equivalence of entropy balancing and the method of moments for matching-adjusted indirect comparison. *Res Synth Methods* 2020-07; 11(4): 568–572. DOI:10.1002/jrsm.1416.
18. Josey KP, Berkowitz SA, Ghosh D et al. Transporting experimental results with entropy balancing. *Statistics in Medicine* 2021; 40(19): 4310–4326. DOI:10.1002/sim.9031.
19. Glimm E and Yau L. Geometric approaches to assessing the numerical feasibility for conducting matching-adjusted indirect comparisons. *Pharmaceutical Statistics* 2022; 21(5): 974–987. DOI:10.1002/pst.2210.
20. Dahabreh IJ, Robertson SE, Steingrimsson JA et al. Extending inferences from a randomized trial to a new target population. *Statistics in Medicine* 2020; 39(14): 1999–2014. DOI:10.1002/sim.8426.
21. Phillippo DM, Dias S, Ades AE et al. Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study. *Stat Med* 2020-12-30; 39(30): 4885–4911. DOI:10.1002/sim.8759.
22. Remiro-Azócar A, Heath A and Baio G. Model-based standardization using multiple imputation. *BMC Medical Research Methodology* 2024; 24(1): 32.
23. Ding P, Li X and Miratrix LW. Bridging finite and super population causal inference. *Journal of Causal Inference* 2017; 5(2): 20160027.
24. Remiro-Azócar A. Transportability of model-based estimands in evidence synthesis. *Statistics in Medicine* 2024; 43(22): 4217–4249.
25. Phillippo DM, Dias S, Ades A et al. Multilevel network meta-regression for population-adjusted treatment comparisons. *Journal of the Royal Statistical Society Series A: Statistics in Society* 2020; 183(3): 1189–1210.

Appendix

The influence function in the transportability context is:

$$D^z_\Psi(P)(O) = \frac{I(S=1,Z=z)}{P(S=1,Z=z|W)}\frac{P(S=0|W)}{P(S=0)}\{Y-\bar{Q}(1,W,z)\} + \frac{I(S=0)}{P(S=0)}[\bar{Q}(1,W,z) - E\{\bar{Q}(1,W,z)|S=0\}].$$

		Method				
N_sample	BC_location	MAIC	AIPW	OM	TMLE	Weighted DR
Moderate population overlap						
100	0.2	0.825	0.762	0.674	0.702	0.739
200	0.2	0.523	0.503	0.462	0.474	0.491
600	0.2	0.296	0.289	0.272	0.276	0.286
Poor population overlap						
100	0.5	1.675	1.015	0.741	0.780	0.876
200	0.5	0.700	0.659	0.526	0.541	0.616
600	0.5	0.378	0.365	0.303	0.307	0.352

Empirical standard errors are reported for each method and scenario combination.

[1] MAIC: Matching-Adjusted Indirect Comparison

[2] AIPW: Augmented Inverse Probability Weighting

[3] OM: Outcome modelling using Bayesian G-computation

[4] TMLE: Targeted Maximum Likelihood Estimation

[5] Weighted DR: Doubly Robust estimation with weighted regressions

Table 3. Empirical standard errors for different population adjustment methods when marginalizing over the true ALD population

		Method				
N_sample	BC_location	MAIC	AIPW	OM	TMLE	Weighted DR
Moderate population overlap						
100	0.2	0.825	0.727	0.657	0.675	0.714
200	0.2	0.523	0.478	0.445	0.453	0.471
600	0.2	0.296	0.276	0.260	0.262	0.272
Poor population overlap						
100	0.5	1.675	0.993	0.731	0.758	0.875
200	0.5	0.700	0.638	0.509	0.522	0.604
600	0.5	0.378	0.352	0.290	0.293	0.337

Empirical standard errors are reported for each method and scenario combination.

[1] MAIC: Matching-Adjusted Indirect Comparison

[2] AIPW: Augmented Inverse Probability Weighting

[3] OM: Outcome modelling using Bayesian G-computation

[4] TMLE: Targeted Maximum Likelihood Estimation

[5] Weighted DR: Doubly Robust estimation with weighted regressions

Table 4. Empirical standard errors for different population adjustment methods when marginalizing over the ALD populations with wrong dependency structure

		Method				
N_sample	BC_location	MAIC	AIPW	OM	TMLE	Weighted DR
Moderate population overlap						
100	0.2	17.804	3.935	-5.031	-0.696	1.146
200	0.2	-0.317	1.214	0.912	1.950	0.591
600	0.2	-1.408	-1.836	3.564	-0.691	-1.943
Poor population overlap						
100	0.5	182.148	7.706	-5.190	-0.322	29.144
200	0.5	5.666	-2.563	-1.521	-0.306	-5.498
600	0.5	-2.385	-3.203	3.202	0.350	-3.501

Relative errors in standard error estimates (%) are reported for each method and scenario combination.

[1] MAIC: Matching-Adjusted Indirect Comparison

[2] AIPW: Augmented Inverse Probability Weighting

[3] OM: Outcome modelling using Bayesian G-computation

[4] TMLE: Targeted Maximum Likelihood Estimation

[5] Weighted DR: Doubly Robust estimation with weighted regressions

Table 5. Relative errors in standard error estimation (%) for different population adjustment methods when marginalizing over the true ALD population