Received: 1 May 2020 | Revised: 1 June 2021 | Accepted: 21 June 2021

DOI: 10.1002/jrsm.1511

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Methods for population adjustment with limited access to individual patient data: A review and simulation study

Antonio Remiro-Azócar[1,2] | **Anna Heath**[1,3,4] | **Gianluca Baio**[1]

[1]Department of Statistical Science, University College London, London, UK

[2]Quantitative Research, Statistical Outcomes Research & Analytics (SORA) Ltd., London, UK

[3]Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, Ontario, Canada

[4]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Correspondence
Antonio Remiro-Azócar, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.
Email: antonio.remiro.16@ucl.ac.uk

Funding information
Canadian Institutes of Health Research, Grant/Award Number: MYG-151207; Engineering and Physical Sciences Research Council; Mapi/ICON

Abstract

Population-adjusted indirect comparisons estimate treatment effects when access to individual patient data is limited and there are cross-trial differences in effect modifiers. Popular methods include matching-adjusted indirect comparison (MAIC) and simulated treatment comparison (STC). There is limited formal evaluation of these methods and whether they can be used to accurately compare treatments. Thus, we undertake a comprehensive simulation study to compare standard unadjusted indirect comparisons, MAIC and STC across 162 scenarios. This simulation study assumes that the trials are investigating survival outcomes and measure continuous covariates, with the log hazard ratio as the measure of effect. MAIC yields unbiased treatment effect estimates under no failures of assumptions. The typical usage of STC produces bias because it targets a conditional treatment effect where the target estimand should be a marginal treatment effect. The incompatibility of estimates in the indirect comparison leads to bias as the measure of effect is non-collapsible. Standard indirect comparisons are systematically biased, particularly under stronger covariate imbalance and interaction effects. Standard errors and coverage rates are often valid in MAIC but the robust sandwich variance estimator underestimates variability where effective sample sizes are small. Interval estimates for the standard indirect comparison are too narrow and STC suffers from bias-induced undercoverage. MAIC provides the most accurate estimates and, with lower degrees of covariate overlap, its bias reduction outweighs the loss in precision under no failures of assumptions. An important future objective is the development of an alternative formulation to STC that targets a marginal treatment effect.

KEYWORDS

clinical trials, comparative effectiveness research, health technology assessment, indirect treatment comparison, oncology, simulation study

1 | INTRODUCTION

Evaluating the comparative effectiveness of alternative health care interventions lies at the heart of health technology assessments (HTAs), such as those commissioned by the National Institute of Health and Care Excellence (NICE), the body responsible for providing guidance on whether health care technologies should be publicly funded in England and Wales.[1] The randomized controlled trial (RCT) is the most reliable design for estimating the relative efficacy of new treatments.[2] However, new treatments are typically compared against placebo or standard of care before the licensing stage, but not necessarily against other active interventions—a comparison that is required for HTAs. In the absence of data from head-to-head RCTs, indirect treatment comparisons (ITCs) are at the top of the hierarchy of evidence when assessing the relative efficacy of interventions and can inform treatment and reimbursement decisions.[3]

Standard ITC techniques, such as network meta-analysis, are useful when there is a common comparator arm between RCTs, or more generally a connected network of studies.[3,4] These methods can be used with individual patient data (IPD) or aggregate-level data (ALD), with IPD considered the gold standard.[5] However, standard ITCs assume that there are no cross-trial differences in the distribution of effect-modifying variables (more specifically, that relative treatment effects are constant) and produce biased estimates when these exist.[6] Popular balancing methods such as propensity score matching[7] can account for these differences but require access to IPD for all the studies being compared.[8]

In many HTA processes, there are: (1) no head-to-head trials comparing the interventions of interest; (2) IPD available for at least one intervention (e.g., from the submitting company's own trial), but only published ALD for the relevant comparator(s); and (3) cross-trial differences in effect modifiers, implying that relative treatment effects are not constant across trial populations. Several methods, labeled *population-adjusted indirect comparisons*, have been introduced to estimate relative treatment effects in this scenario. These include matching-adjusted indirect comparison (MAIC),[9,10,11] based on inverse propensity score weighting,[12] and simulated treatment comparison (STC),[13] based on regression adjustment,[14] and require access to IPD from at least one of the trials.

The NICE Decision Support Unit has published formal submission guidelines for population adjustment with limited access to IPD.[6,15] Various reviews[6,15,16,17] define the relevant terminology and assess the theoretical validity of these methodologies but do not express a preference. Questions remain about the correct application of the methods and their validity in HTA.[6,15,18] Thus, Phillippo et al.[6] state that current guidance can only be provisional, as more thorough understanding of the properties of population-adjusted indirect comparisons is required.

Consequently, several simulation studies have been published since the release of the NICE guidance.[19,20,21,22,23,24] These have primarily assessed the performance of MAIC relative to standard ITCs in a limited number of simulation scenarios. In general, the studies set relatively low effect modifier imbalances and do not vary these, even though MAIC is prone to large reductions in effective sample size and imprecise estimates of the treatment effect when high imbalances lead to poor overlap.[25] Most importantly, existing simulation studies typically consider binary covariates at non-extreme values, not close to zero or one. In these scenarios, MAIC is likely to perform well as covariate overlap is strong. Propensity score weighting methods such as MAIC are known to be highly sensitive to scenarios with poor overlap,[26,27,28] because of their inability to extrapolate beyond the observed covariate space. Hence, evaluating the performance of MAIC in the face of practical scenarios with poor covariate overlap is important.

In this paper, we carry out an up-to-date review of MAIC and STC, and a comprehensive simulation study to benchmark the performance of the methods against the standard ITC. The simulation study provides proof-of-principle for the methods and is based on scenarios with survival outcomes and continuous covariates, with the log hazard ratio as the measure of effect. The methods are evaluated in a wide range of settings; varying the trial sample size, effect-modifying strength of covariates, prognostic effect of covariates, imbalance/overlap of covariates and the level of correlation in the covariates. One hundred sixty-two simulation scenarios are considered, providing the most extensive evaluation of population adjustment methods to date. An objective of the simulation study is to inform the circumstances under which population adjustment should be applied and which specific method is preferable in a given situation.

In Section 2, we establish the context and data requirements for population-adjusted indirect comparisons. In Section 3, we present an updated review of MAIC and STC. Section 4 describes a simulation study, which evaluates the properties of these approaches under a variety of conditions. Section 5 presents the results of the simulation study. An extended discussion of our findings and their implications is provided in Section 6. Finally, we make some concluding remarks in Section 7.

2 | CONTEXT

HTA often takes place late in the drug development process, after a new medical technology has obtained

regulatory approval, typically based on a two-arm RCT that compares the new intervention to placebo or standard of care. At the licensing stage, the question of interest is whether or not the drug is effective. In HTA, the relevant policy question is: "given that there are finite resources available to finance health care, which is the best treatment of all available options in the market?" In order to answer this question, one must evaluate the relative effectiveness of interventions that may not have been trialed against each other.

Indirect treatment comparison methods are used when we wish to compare the relative effect of interventions *A* and *B* for a specific outcome, but no head-to-head trials are currently available. Typically, it is assumed that the comparison is undertaken using additive effects for a given linear predictor, for example, log hazard ratio for time-to-event outcomes or log-odds ratio for binary outcomes. Indirect comparisons are typically performed on this scale.[3,4] In addition, we assume that the comparison is "anchored", that is, a connected treatment network is available through a common comparator *C*, for example, placebo or standard of care. We note that comparisons can be unanchored, for example, using single-arm trials or disconnected treatment networks, but this requires much stronger assumptions.[6] The NICE Decision Support Unit discourages the use of unanchored comparisons when there is connected evidence and labels these as problematic.[6,15] This is because they do not respect within-study randomization and are not protected from imbalances in any covariates that are prognostic of outcome (in essence implying that absolute outcomes can be predicted from the covariates, a heroic assumption). Hence, we do not present the methodology behind these.

A manufacturer submitting evidence for reimbursement to HTA bodies has access to patient-level data from its own trial that compares its product *A* against standard intervention *C*. However, as disclosure of proprietary, confidential patient-level data from industry-sponsored clinical trials is rare, IPD for the competitor's trial, comparing its treatment *B* against *C*, are, almost invariably, unavailable (for both the manufacturer submitting evidence for reimbursement and the national HTA agency evaluating the evidence). We consider, without loss of generality, that IPD are available for a trial comparing intervention *A* to intervention *C* (denoted *AC*) and published ALD are available for a trial comparing *B* to *C* (*BC*).

Standard methods for indirect comparisons such as the Bucher method,[4] a special case of network meta-analysis, allow for the use of ALD and estimate the *A* versus *B* treatment effect as:

$$\hat{\Delta}_{AB} = \hat{\Delta}_{AC} - \hat{\Delta}_{BC}, \quad (1)$$

where $\hat{\Delta}_{AC}$ is the estimated relative treatment effect of *A* versus *C* (in the *AC* population), and $\hat{\Delta}_{BC}$ is the estimated relative effect of *B* versus *C* (in the *BC* population). The estimate $\hat{\Delta}_{AC}$ and an estimate of its variance can be calculated from the available IPD. The estimate $\hat{\Delta}_{BC}$ and an estimate of its variance may be directly published or derived from aggregate outcomes made available in the literature. As the indirect comparison is based on relative treatment effects observed in separate RCTs, the within-trial randomization of the originally assigned patient groups is preserved. The within-trial relative effects are statistically independent of each other; hence, their variances are simply summed to estimate the variance of the *A* versus *B* treatment effect.

Standard indirect comparisons assume that there are no cross-trial differences in the distribution of effect-modifying variables. That is, the relative treatment effect of *A* versus *C* in the *AC* population (indicated as Δ_{AC}) is assumed equivalent to the treatment effect that would occur in the *BC* population* (denoted Δ^*_{AC})—throughout the paper the asterisk superscript represents a quantity that has been mapped to a different population; for example, in our case, the *A* versus *C* treatment effect in the *AC* population is mapped to the population of the *BC* trial.

Often, treatment effects are influenced by variables that interact with treatment on a specific scale (e.g., the linear predictor), altering the effect of treatment on outcomes. If these *effect modifiers* are distributed differently across *AC* and *BC*, relative treatment effects differ in the trial populations and the assumptions of the Bucher method are broken. In this case, a standard ITC between *A* and *B* is liable to bias and may produce overly precise efficacy estimates.[30] From the economic modeling point of view, these features are undesirable, as they impact negatively on the "probabilistic sensitivity analysis,"[31] the (often mandatory) process used to characterize the impact of the uncertainty in the model inputs on decision-making.

As a result, population adjustment methodologies such as MAIC and STC have been introduced. These target the *A* versus *C* treatment effect that would be observed in the *BC* population, thereby performing an adjusted indirect comparison in such population. The adjusted *A* versus *B* treatment effect is estimated as:

$$\hat{\Delta}^*_{AB} = \hat{\Delta}^*_{AC} - \hat{\Delta}_{BC}, \quad (2)$$

where $\hat{\Delta}^*_{AC}$ is the estimated relative treatment effect of *A* versus *C* (in the *BC* population, implicitly assumed to be the relevant target population). Variances are combined in the same way as the Bucher method.

Those studying the generalizability of treatment effects often make a distinction between sample and population treatment effects.[32,33,34,35] Typically, another implicit assumption made by population-adjusted indirect comparisons is that the treatment effects estimated in the *BC* sample, as described by its published covariate moments in the case of $\hat{\Delta}^{*}_{AC}$, coincide with those that would be estimated in the target population of the trial. Namely, either the study sample on which inferences are made is the study target population, or it is a simple random sample (i.e., representative) of such population, ignoring the sampling variability in the descriptive characteristics.

The use of population adjustment in HTA, both in published literature as well as in submissions for reimbursement, and its acceptability by national HTA bodies, for example, in England and Wales, Scotland, Canada and Australia,[18] is increasing across diverse therapeutic areas.[18,25,36,37] As of April 11, 2020, a search among titles, abstracts and keywords for "matching-adjusted indirect comparison" and "simulated treatment comparison" in Scopus, reveals at least 89 peer-reviewed applications of MAIC and STC and conceptual papers about the methods. In addition, at least 30 technology appraisals (TAs) published by NICE use MAIC or STC—of these, 23 have been published since 2017. Figure 1 shows the rapid growth of peer-reviewed publications and NICE TAs featuring MAIC or STC since the introduction of these methods in 2010. MAIC and STC are predominantly applied in the evaluation of cancer drugs, as 26 of the 30 NICE TAs using population adjustment have been in oncology.

3 | METHODOLOGY

We shall assume that the following data are available for the i-th subject ($i = 1,...,N$) in the *AC* trial:

- A covariate vector of K baseline characteristics $\boldsymbol{X}_i = (X_{i,1},...,X_{i,K})$, for example, age, gender, comorbidities.
- A treatment indicator T_i. Without loss of generality, we assume here for simplicity that $T_i \in \{0,1\}$ for the common comparator and active treatment, respectively.
- An observed outcome Y_i, for example, a time-to-event or binary indicator for some clinical measurement.

Given this information, one can compute an unadjusted estimate $\hat{\Delta}_{AC}$ of the A versus C treatment effect, and an estimate of its variance. In the Bucher method, such estimate would be plugged in to Equation (1). On the other hand, MAIC and STC generate a population-adjusted estimate $\hat{\Delta}^{*}_{AC}$ of the A versus C treatment effect that would be plugged in to Equation (2).

For the *BC* trial, data available are:

- A vector $\overline{\boldsymbol{X}}_{BC} = \left(\overline{X}_{BC,1},...,\overline{X}_{BC,K}\right)$ of published summary values for the baseline characteristics. For ease of exposition, we shall assume that these are means and are available for all K covariates (alternatively, one would take the intersection of the available covariates).
- An estimate $\hat{\Delta}_{BC}$ of the B versus C treatment effect in the *BC* population, and an estimate of its variance, either published directly or derived from aggregate outcomes in the literature.

Each baseline characteristic $k = 1,...,K$ can be classed as a prognostic variable (a covariate that affects outcome),

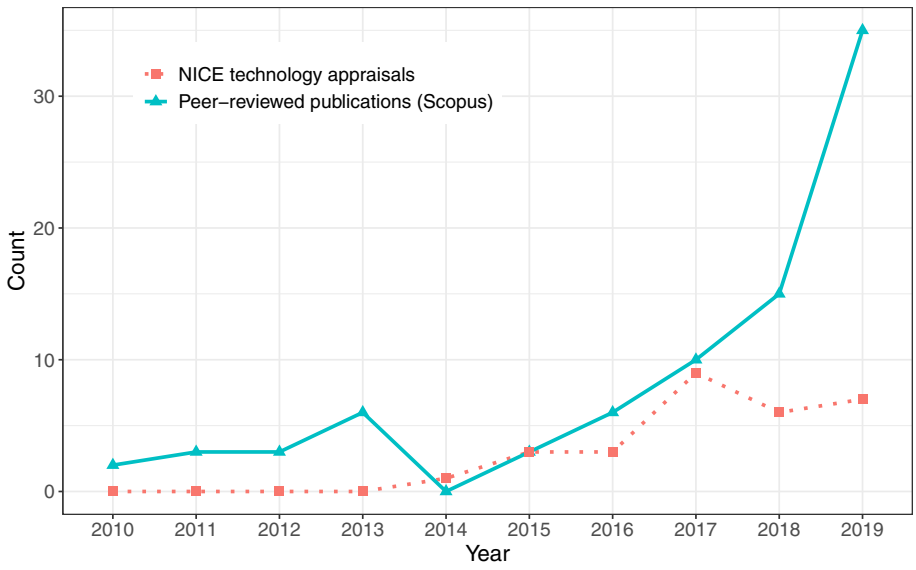

FIGURE 1 Number of peer-reviewed publications and technology appraisals from the National Institute for Health and Care Excellence (NICE) using population-adjusted indirect comparisons per year [Colour figure can be viewed at wileyonlinelibrary.com]

an effect modifier (a covariate that interacts with treatment A to affect outcome), both or none. For simplicity in the notation, it is assumed that all available baseline characteristics are prognostic of the outcome and that a subset of these, $\boldsymbol{X}_i^{(\boldsymbol{EM})} \subseteq \boldsymbol{X}_i$, are selected as effect modifiers (of treatment A) on the linear predictor scale. Similarly, for the published summary values, $\bar{\boldsymbol{X}}_{BC}^{(EM)} \subseteq \bar{\boldsymbol{X}}_{BC}$. Note that we select the effect modifiers of treatment A with respect to C (as opposed to the effect modifiers of treatment B with respect to C), because we have to adjust for these to perform the indirect comparison in the BC population, implicitly assumed to be the target population.†

3.1 | Matching-adjusted indirect comparison

Matching-adjusted indirect comparison (MAIC) is a population adjustment method based on inverse propensity score weighting.[12] IPD from the AC trial are weighted so that the means and, potentially, higher moments of specified covariates match those in the BC trial. The weights are estimated using a propensity score logistic regression model:

$$\ln(w_i) = \alpha_0 + \boldsymbol{X}_i^{(\boldsymbol{EM})}\boldsymbol{\alpha_1},$$

where α_0 and $\boldsymbol{\alpha_1}$ are the regression parameters, and the weight w_i assigned to each individual i represents the "trial selection" odds, that is, the odds of being enrolled in the BC trial as opposed to being enrolled in the AC trial. These are defined as a function of the baseline characteristics modifying the effect of treatment A, $\boldsymbol{X}_i^{(\boldsymbol{EM})}$ for subject i. Note that in standard applications of propensity score weighting, for example, in observational studies, the propensity score logistic regression is for the *treatment group* assigned to the subject. In MAIC, the objective is to balance covariates across studies so the propensity score model is for the *trial* in which the participant is enrolled.

The regression parameters cannot be derived using conventional methods such as maximum-likelihood estimation because IPD are not available for BC. Signorovitch et al.[9] propose using a method of moments to estimate the model parameters by setting the weights so that the mean effect modifiers are exactly balanced across the two trial populations. After centering the AC effect modifiers on the published BC means, such that $\overline{\boldsymbol{X}}_{BC}^{(EM)} = \boldsymbol{0}$, the weights are estimated by minimizing the objective function:

$$Q(\boldsymbol{\alpha}_1) = \sum_{i=1}^{N} \exp\left(\boldsymbol{X}_i^{(\boldsymbol{EM})}\boldsymbol{\alpha_1}\right),$$

where N represents the number of subjects in the AC trial. $Q(\boldsymbol{\alpha}_1)$ is a convex function that can be minimized using standard algorithms, for example, BFGS,[38] to yield a unique finite solution $\hat{\boldsymbol{\alpha}}_{\boldsymbol{1}} = \text{argmin}(Q(\boldsymbol{\alpha}_1))$. Then, the estimated weight for subject i is:

$$\hat{w}_i = \exp\left(\boldsymbol{X}_i^{(\boldsymbol{EM})}\hat{\boldsymbol{\alpha}}_{\boldsymbol{1}}\right).$$

Consequently, the mean outcomes under treatment $t \in \{A, C\}$ in the BC population are predicted as the weighted average:

$$\hat{Y}_t^* = \frac{\sum_{i=1}^{N_t} Y_{i,t}\hat{w}_i}{\sum_{i=1}^{N_t}\hat{w}_i},$$

where N_t represents the number of subjects in arm t of the AC trial, and $Y_{i,t}$ denotes the outcome for patient i receiving treatment t in the patient-level data. Note that we have summary data from the BC trial to estimate absolute outcomes under C. However, in the anchored scenario, we do not focus on the absolute outcomes as the objective is to generate a relative effect for A versus C in the BC population.

Such relative effect is typically estimated by fitting a weighted model, that is, a model where the contribution of each subject to the likelihood is weighted. For instance, if the outcome of interest is a time-to-event outcome, an "inverse odds"-weighted Cox model can be fitted by maximizing its weighted partial likelihood. In this case, a subject i from the AC trial, who has experienced an event at time τ, contributes the following term to the partial likelihood function:

$$\left(\frac{\exp(\beta_T T_i)}{\sum_{j \in R(\tau)} \hat{w}_j \exp\left(\beta_T T_j\right)}\right)^{\hat{w}_i}, \tag{3}$$

where $R(\tau)$ is the set of subjects without the event and uncensored prior to τ, that is, the risk set. Here, the fitted coefficient $\hat{\beta}_T$ of the weighted regression (i.e., the value of the parameter maximizing the partial likelihood in Equation 3) is the estimated relative effect for A versus C, such that $\hat{\Delta}_{AC}^* = \hat{\beta}_T$.

In the original MAIC approach, covariates are balanced for active treatment and control arms combined and standard errors are computed using a robust

sandwich estimator, which allows for heteroskedasticity.[9,39] Typically, implementations of this estimator do not explicitly account for the fitting of the logistic regression model for the weights, assuming these to be fixed.

Terms of higher order than means can also be balanced, for example, by including squared covariates in the method of moments to match variances. However, this decreases the degrees of freedom and may increase finite-sample bias.[40] Matching both means and variances (as opposed to means only) appears to result in more biased and less accurate treatment effect estimates when covariate variances differ across trials.[20,22]

A proposed modification to MAIC uses entropy balancing[41] instead of the method of moments to estimate the weights.[20,23] Entropy balancing has the additional constraint that the weights are as close as possible to unit weights. Potentially, it should penalize extreme weighting schemes and provide greater precision. However, Phillippo et al. recently demonstrated that weight estimation via entropy balancing and the method of moments are mathematically identical.[42] Other proposed modifications to MAIC include balancing the covariates separately for active treatment and common comparator arms,[20,23] and using the bootstrap[43,44] to compute standard errors,[45] which does not rely upon strong assumptions about the estimation of the MAIC weights. Balancing the covariates separately seems to provide greater precision in simulation studies.[20] However, we do not recommend this approach because it may break randomization, distorting the balance between treatment arms *A* and *C* on covariates that are not accounted for in the weighting. If these covariates are prognostic of outcome, this would compromise the internal validity of the within-study treatment effect estimate for *A* versus *C*.

As MAIC is a reweighting procedure, it will reduce the effective sample size (ESS) of the *AC* trial. The approximate ESS of the weighted IPD is estimated as $\left(\sum_i \hat{w}_i\right)^2 / \sum_i \hat{w}_i^2$; the reduction in ESS can be viewed as a rough indicator of the lack of overlap between the *AC* and *BC* covariate distributions. For relative effects to be conditionally constant and eventually produce an unbiased indirect comparison, one needs to include all effect modifiers in the weighting procedure, whether in imbalance or not (see Supplementary Appendix A of the Supporting Information for a non-technical overview of the full set of assumptions made by MAIC, and more generally, by population-adjusted indirect comparisons).[15] The exclusion of balanced covariates does not ensure their balance after the weighting procedure. Including too many covariates or poor overlap in the covariate distributions can induce extreme weights and large reductions in ESS. This is a pervasive problem in NICE TAs, where most of the reported ESSs are small with a large percentage reduction from the original sample size.[25]

Propensity score mechanisms are very sensitive to poor overlap.[26,27,28] In particular, weighting methods are unable to extrapolate—in the case of MAIC, extrapolation beyond the covariate space observed in the *AC* IPD is not possible. Almost invariably, the level of overlap between the covariate distributions will decrease as a greater number of covariates are included. Therefore, no purely prognostic variables should be balanced to avoid loss of effective sample size and consequent inflation of the standard error due to over-balancing.[6] Cross-trial imbalances in purely prognostic variables should not produce bias as relative treatment effects are unaffected in expectation due to within-trial randomization.[15]

3.2 | Simulated treatment comparison

While MAIC is a reweighting method, simulated treatment comparison (STC)[13] is based on regression adjustment.[14] Regression adjustment methods are promising because they may increase precision and statistical power with respect to propensity score-based methodologies.[46,47,48] Contrary to most propensity score methods, regression adjustment mechanisms are able to extrapolate beyond the covariate space where overlap is insufficient, using the linearity assumption or other appropriate assumptions about the input space. However, the validity of the extrapolation depends on the accuracy in capturing the true covariate-outcome relationships.

In the typical version of STC, IPD from the *AC* trial are used to fit a regression of the outcome on the baseline characteristics and treatment. Following the NICE Decision Support Unit Technical Support Document 18,[6,15] the following linear predictor is fitted to the IPD:

$$g\left(\eta_i^*\right) = \beta_0 + \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{BC}\right)\boldsymbol{\beta_1} + \left[\beta_T + \left(\boldsymbol{X}_i^{(\boldsymbol{EM})} - \overline{\boldsymbol{X}}_{BC}^{(\boldsymbol{EM})}\right)\boldsymbol{\beta_2}\right] 1(T_i = 1), \quad (4)$$

where η_i^* is the expected outcome on the natural outcome scale, e.g. the probability scale for binary outcomes, of subject *i*, $g(\cdot)$ is an appropriate link function (e.g., logit for binary outcomes), β_0 is the intercept, $\boldsymbol{\beta_1}$ is a vector of *K* regression coefficients for the prognostic variables, $\boldsymbol{\beta_2}$ is a vector of interaction coefficients for the effect modifiers (modifying the effect of treatment *A* vs. *C*) and β_T is the *A* versus *C* treatment coefficient. The covariates are centered at the published mean values from the *BC* population, $\overline{\boldsymbol{X}}_{BC}$ and $\overline{\boldsymbol{X}}_{BC}^{(\boldsymbol{EM})}$, respectively. Hence, the estimated $\hat{\beta}_T$ is directly interpreted as the *A* versus *C* treatment effect in the *BC* population, such that $\hat{\Delta}_{AC}^{*} = \hat{\beta}_T$.

The variance of said treatment effect is derived directly from the fitted model (see Phillippo et al[6,15] for a breakdown of uncertainty propagation in the estimates resulting from MAIC and STC). In a Cox proportional hazards regression framework, a log link function could be employed in Equation (4) between the hazard function and the linear predictor component of the model.

For relative effects to be conditionally constant across studies, one needs to include in the model the effect modifiers that are imbalanced. In addition, the relationship between the effect modifiers and outcome must be correctly specified; in the case of this article, the effect modifiers must have an additive interaction with treatment on the linear predictor scale. It is optional to include (and to center) imbalanced variables that are purely prognostic. These will not remove bias further but a strong fit of the outcome model may increase precision. NICE guidance[15] suggests adding purely prognostic variables if they increase the precision of the model and account for more of its underlying variance, as reported by model selection criteria (e.g., residual deviance or information criteria). However, such tools should not guide decisions on effect modifier status, which must be defined prior to fitting the outcome model. As effect-modifying covariates are likely to be good predictors of outcome, the inclusion of appropriate effect modifiers should provide an acceptable fit.

Alternative "simulation-based" formulations to STC have been proposed.[16,49] These are outlined as follows. The joint distribution of *BC* covariates is approximated under certain parametric assumptions to characterize the *BC* population, for example, simulating continuous covariates at the individual level from a multivariate normal with the *BC* means and the correlation structure observed in the *AC* IPD. A regression of the outcome on the predictors is fitted to the *AC* patient-level data (this time, the covariates are not centered at the mean *BC* values). Then, the coefficients of this regression are applied to the simulated subject profiles and the linear predictions for patients under *A* and under *C* in the *BC* population are averaged out. The treatment effect for *A* versus *C* is given by subtracting the average linear prediction under *C* from the average linear prediction under *A*. Neither the original conceptual publications nor NICE guidance provide detailed information about variance estimation, which is likely to be complicated and probably requires bootstrapping or similar approaches.

It is worth noting that, in the linear predictor scale, the arithmetic mean of the average linear predictor (the average linear predictor for patients sampled under the centered covariates) and its geometric mean (the linear predictor evaluated at the expectation of the centered covariates) coincide. Therefore, provided that the number of simulated subjects is sufficiently large (i.e., in expectation or ignoring sampling variability), the "covariate simulation" approach generates estimates that are equivalent to those of the "plug-in" methodology adopted in this article.

3.3 | Clarification of estimands

In an indirect treatment comparison, the objective is to emulate the analysis of a head-to-head RCT between *A* and *B*. However, RCTs have two potential target estimands: *marginal* and/or *conditional* treatment effects. In MAIC, as is typically the case for propensity score methods, $\hat{\Delta}^{*}_{AC}$ targets a *marginal* treatment effect.[7,50,51] In biostatistics[52,53,54,55] and epidemiology,[56,57,58] this marginal effect is also known as a *population-average* or *population-level* treatment effect, as it measures the average treatment effect for *A* versus *C* at the population level (conditional on the entire population distribution of covariates, such that the individual-level covariates have been marginalized over). This denotes the average outcome between two identical populations, except that in one population all subjects are under *A*, while in the other population all subjects are under *C*,[59] and where the difference is taken on a suitable transformed scale, for example, the linear predictor scale. MAIC targets a marginal treatment effect because it performs a weighted regression of outcome on treatment assignment alone. Therefore, assuming a reasonably large sample size and proper randomization in the *AC* trial, the fitted coefficient $\hat{\beta}_T$ in Equation (3) estimates a relative effect between subjects that have the same distribution of baseline characteristics (corresponding to the *BC* population).

In HTA and health policy, interest typically lies in the impact of a health technology on the target population for the decision problem, which MAIC and STC implicitly assume to be the *BC* population. Where making decisions at the population level, the effect of interest is a marginal treatment effect: the average effect, at the population level, of moving the target population from treatment *B* to treatment *A*.[7,60] The majority of trials report an estimate $\hat{\Delta}_{BC}$ that targets a marginal treatment effect. It is likely derived from a RCT publication where a simple regression of outcome on a single independent variable, treatment assignment, has been fitted.

In the version of STC outlined by the NICE Decision Support Unit, $\hat{\Delta}^{*}_{AC}$ targets a *conditional* treatment effect. The conditional treatment effect denotes the average effect, at the individual level, of changing a subject's treatment from *C* to *A*.[7,59] STC targets a conditional treatment effect because the estimate is obtained from the regression coefficient of a multivariable regression ($\hat{\beta}_T$ in Equation 4), where the baseline covariates included as

predictors are also adjusted for. Hence, the relative effect is an average at the subject level, fully conditioned on the covariates of the average subject.‡ Conditional measures of effect are clinically relevant as patient-centered evidence in a clinician–patient context, where decision-making relates to the treatment benefit for an individual subject with specific covariate values. Conditional treatment effects are typically not of interest when making decisions at the population level in HTA and health policy, as they are unit-level measures of effect.

A measure of effect is said to be *collapsible* if marginal and conditional effects coincide in the absence of confounding bias.[56,62] The property of collapsibility is closely related to that of linearity,[63,64] for example, mean differences in a linear regression are collapsible.[7,59,56,62] However, most applications of population-adjusted indirect comparisons are in oncology and are typically concerned with time-to-event outcomes, or rate outcomes modeled using logistic regression.[25] These yield non-collapsible measures of treatment effect such as (log) hazard ratios[7,59,56,65] or (log) odds ratios.[7,59,56,62,65,66,55]

With non-collapsible measures of effect, marginal and conditional estimands do not coincide due to non-linearity,[63] even if there is covariate balance and no confounding.[56,62] With both collapsible and non-collapsible measures of effect, estimators targeting distinct estimands will have different standard errors. Therefore, marginal and conditional estimates quantify parametric uncertainty differently, and conflating these will lead to the incorrect propagation of uncertainty to the wider health economic decision model, which will be problematic for probabilistic sensitivity analyses.

Therefore, the relative effect estimate $\hat{\Delta}_{AC}^{(*)}$ in STC is unable to target a marginal treatment effect and the comparison of interest, a comparison of compatible marginal effects, cannot be performed. A comparison of conditional effects is not of interest, and also, cannot be carried out. A compatible conditional effect for *B* versus *C* is unavailable because its estimation requires fitting the non-centered version of Equation (4), adjusting for the same baseline characteristics, to the *BC* patient-level data. Such data are unavailable and it is unlikely that the estimated treatment coefficient from this model is available in the clinical trial publication.

Hence, $\hat{\Delta}_{AC}^{*}$ is incompatible with $\hat{\Delta}_{BC}$ in the indirect comparison (Equation 2) for STC, even if all effect modifiers are accounted for and the outcome model is correctly specified. If we intend to target a marginal estimand for the *A* versus *C* treatment effect (in the *BC* population) and naively assume that STC does so, $\hat{\Delta}_{AB}^{*}$ may produce a biased estimate of the marginal treatment effect for *A* versus *B*, even if all the assumptions in Supplementary Appendix A of the Supporting Information are met. On the other hand, $\hat{\Delta}_{AC}^{*}$ targets a marginal treatment effect in MAIC. There are no compatibility issues in the indirect comparison as $\hat{\Delta}_{AC}^{*}$ and $\hat{\Delta}_{BC}$ target compatible estimands of the same form. In the Bucher method, if the estimate $\hat{\Delta}_{AC}$ is derived from a simple comparison of group means or from an univariable regression of outcome on treatment in the *AC* IPD, this targets a marginal effect and there are no compatibility issues in the indirect treatment comparison either.

4 | SIMULATION STUDY

4.1 | Aims

The objectives of the simulation study are to compare MAIC, STC and the Bucher method across a wide range of scenarios that may be encountered in practice. For each estimator, we assess the following properties[67]: (1) unbiasedness; (2) variance unbiasedness; (3) randomization validity§; and (4) precision. The selected performance measures evaluate these criteria specifically (see Section 4.5). The simulation study is reported following the ADEMP (Aims, Data-generating mechanisms, Estimands, Methods, Performance measures) structure.[67] All simulations¶ and analyses were performed using R software version 3.6.3.[68] Example R code implementing MAIC, STC and the Bucher method on a simulated example is provided in Supplementary Appendix D of the Supporting Information.

4.2 | Data-generating mechanisms

As most applications of MAIC and STC are in oncology, the most prevalent outcome types are survival or time-to-event outcomes (e.g., overall or progression-free survival).[25] Hence we consider these using the log hazard ratio as the measure of effect.

For trials *AC* and *BC*, we follow Bender et al.[69] to simulate Weibull-distributed survival times under a proportional hazards parameterization.** Survival time τ_i (for subject *i*) is generated according to the formula:

$$\tau_i = \left(\frac{-\ln U_i}{\lambda \exp\left[\boldsymbol{X}_i \boldsymbol{\beta_1} + \left(\beta_T + \boldsymbol{X}_i^{(\boldsymbol{EM})} \boldsymbol{\beta_2} \right) 1(T_i = 1) \right]} \right)^{1/\nu}, \quad (5)$$

where U_i is a uniformly distributed random variable, $U_i \sim (0,1)$. We set the inverse scale of the Weibull distribution to $\lambda = 8.5$ and the shape to $\nu = 1.3$ as these parameters produce a functional form reflecting frequently

observed mortality trends in metastatic cancer patients (as illustrated in Figures 2 and 3, which display the survival curves implied by the parameters).[22] Four correlated or uncorrelated continuous covariates $\boldsymbol{X}_i$ are generated per subject using a multivariate Gaussian copula.[71] Two of these are purely prognostic variables; the other two ($\boldsymbol{X}_i^{(\boldsymbol{EM})}$) are effect modifiers, modifying the effect of both treatments *A* and *B* with respect to *C* on the log hazard ratio scale, and prognostic variables.

We introduce random right censoring to simulate loss to follow-up within each trial. Censoring times $\tau_{c,i}$ are generated from the exponential distribution $\tau_{c,i} \sim \text{Exp}(\lambda_c)$, where the rate parameter $\lambda_c = 0.96$ is selected to achieve a censoring rate of 35% under the active treatment at baseline (with the values of the covariates set to zero), considered moderate censoring.[72] We fix the value of λ_c before generating the datasets, by simulating survival times for 1,000,000 subjects with Equation (5) and using the R function optim (Brent's method[73]) to minimize the difference between the observed and targeted censoring proportion.

The number of subjects in the *BC* trial is 600, under a 1:1 active treatment versus control allocation ratio. This sample size corresponds to that of a reasonably large Phase III RCT.[74] Different values are not explored as preliminary results showed that these drive performance less than the number of subjects in the *AC* trial. While the number of subjects in *BC* contributes to sampling variability, the reweighting or regressions are performed in the *AC* patient-level data. For the *BC* trial, the individual-level covariates and outcomes are aggregated to obtain summaries. The continuous covariates are summarized as means—these would typically be available to the analyst in the published study as a table of baseline characteristics. The marginal *B* versus *C* treatment effect and its variance are estimated through a Cox proportional hazards regression of outcome on treatment. These estimates make up the only information on aggregate outcomes available to the analyst.

The simulation study examines five factors in a fully factorial arrangement with $3 \times 3 \times 3 \times 2 \times 3 = 162$ scenarios to explore the interaction between factors. The simulation scenarios are defined by varying the values of the following parameters, which are inspired by applications of MAIC and STC in NICE technology appraisals:

- The number of patients in the *AC* trial, $N \in \{150, 300, 600\}$ under a 1:1 active intervention versus control allocation ratio. The sample sizes correspond to typical values for a Phase III RCT[74] and for trials included in applications of MAIC and STC submitted to HTA authorities.[25]
- The strength of the association between the prognostic variables and the outcome, $\beta_{1,k} \in \{-\ln(0.67), -\ln(0.5), -\ln(0.33)\}$ (moderate, strong and very strong prognostic variable effect), where k indexes a given covariate. These regression coefficients correspond to fixing the conditional hazard ratios for the effect of each prognostic variable at approximately 1.5, 2 and approximately 3, respectively.
- The strength of interaction of the effect modifiers, $\beta_{2,k} \in \{-\ln(0.67), -\ln(0.5), -\ln(0.33)\}$ (moderate, strong and very strong interaction effect), where k indexes a given effect modifier. These parameters have a material impact on the marginal *A* versus *B* treatment effect. Hence, population adjustment is warranted in order to remove the induced bias.
- The level of correlation between covariates, $\text{cor}(X_{i,k}, X_{i,l}) \in \{0, 0.35\}$ (no correlation and moderate correlation), for subject i and covariates $k \neq l$.

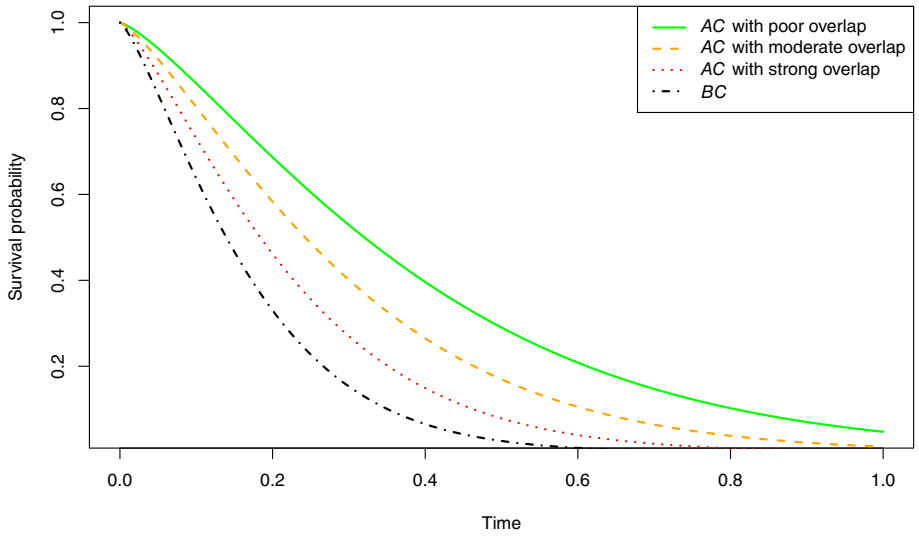

FIGURE 2 Weibull-distributed curves used to simulate survival times for subjects under the active treatment for different trial populations. The covariates are associated with shorter survival and, in the case of the effect modifiers, interact with treatment to render it less effective. As the mean values of the *AC* covariates decrease, overlap decreases [Colour figure can be viewed at wileyonlinelibrary.com]

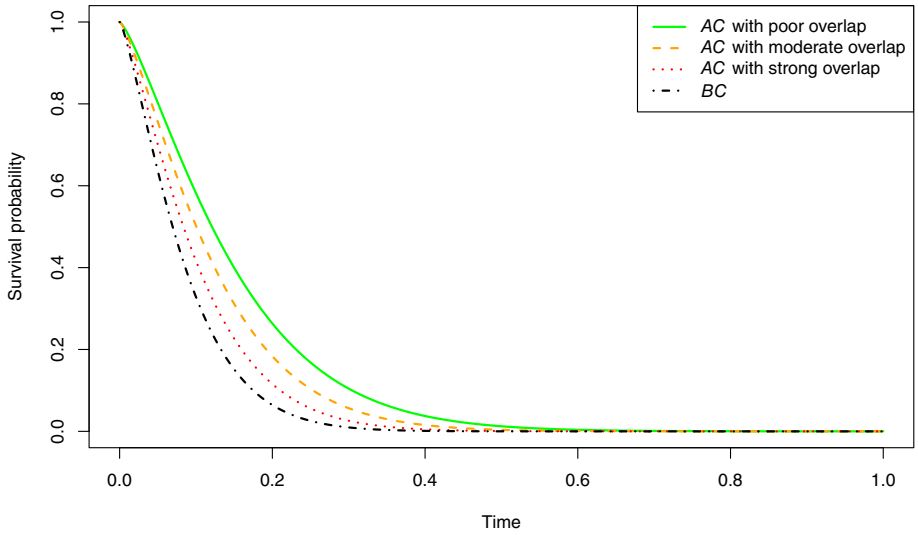

FIGURE 3 Weibull-distributed curves used to simulate survival times for subjects under the common comparator for different trial populations [Colour figure can be viewed at wileyonlinelibrary.com]

- The degree of covariate imbalance.†† For both trials, each covariate k follows a normal marginal distribution. For the BC trial, we fix $X_{i,k} \sim \text{Normal}(0.6, 0.2^2)$, for subject i. For the AC trial, the normal distributions have mean μ_k, such that $X_{i,k} \sim \text{Normal}(\mu_k, 0.2^2)$, varying $\mu_k \in \{0.45, 0.3, 0.15\}$. This yields strong, moderate and poor covariate overlap, respectively, corresponding to average percentage reductions in ESS across scenarios of 19%, 53% and 79%. These percentage reductions in ESS are representative of the range encountered in NICE TAs (see below).

Each active intervention has a very strong conditional treatment effect $\beta_T = \ln(0.25)$ versus the common comparator. The covariates may represent comorbidities, which are associated with shorter survival and, in the case of the effect modifiers, which interact with treatment to render it less effective. Figure 2 shows the Weibull-distributed survival curves for patients under the active treatment (A and B) with varying levels of the covariates. Figure 3 shows the Weibull-distributed survival curves for subjects under the common comparator (C). In Figures 2 and 3, the strength of each prognostic term and each effect-modifying interaction is moderate.

The varying degrees of covariate overlap are inspired by applications of MAIC in technology appraisals submitted to the NICE. Only 13 of the 27 appraisals carrying out a MAIC have effective sample sizes available, albeit some appraisals contain multiple comparisons for different endpoints. In most applications, weighting considerably reduces the effective sample size from the original AC sample size. The median percentage reduction is 58% (range: 7.9%–94.1%; interquartile range: 42.2%–74.2%). The final effective sample sizes are also representative of those in the technology appraisals, which are also small (median: 80; range: 4.8–639; interquartile range: 37–174).

4.3 | Estimands

The estimand of interest is the marginal A versus B treatment effect in the BC population. The treatment coefficient $\beta_T = \ln(0.25)$ is identical for both A versus C and B versus C. Hence, the true conditional effect for A versus B in the BC population is zero (subtracting the treatment coefficient for A vs. C by that for B vs. C). Because the true unit-level treatment effects are zero for all subjects, the true marginal treatment effect in the BC population is zero ($\Delta^*_{AB} = 0$), which implies a "null" simulation setup in terms of the A versus B contrast, and average marginal and conditional effects for A versus B in the BC population coincide by design.

The simulation study meets the shared effect modifier assumption,[15] that is, active treatments A and B have the same set of effect modifiers and the interaction effects $\beta_{2,k}$ of each effect modifier k are identical for both treatments. Hence, the A versus B marginal treatment effect can be generalized to any given target population as effect modifiers are guaranteed to cancel out (the marginal effect for A vs. B is conditionally constant across all populations). If the shared modifier assumption is not met, the true marginal treatment effect for A versus B in the BC population will not be applicable in any target population (one has to assume that the target population is BC), and the average marginal and conditional effects for A versus B will likely not coincide as the measure of effect is non-collapsible.

4.4 | Methods

Each simulated dataset is analyzed using the following methods:

- Matching-adjusted indirect comparison, as originally proposed by Signorovitch et al.,[9] where covariates are balanced for active treatment and control arms combined and weights are estimated using the method of moments. To avoid further reductions in effective sample size and precision, only the effect modifiers are balanced. A weighted Cox proportional hazards model is fitted to the IPD using the R package survival.[75] Standard errors for the *A* versus *C* treatment effect are computed using a robust sandwich estimator[9,39] by setting robust=TRUE in coxph. Given the often arbitrary factors driving selection into different trials, the data-generating mechanism in Section 4.2 does not specify a trial assignment model. Nevertheless, the logistic regression model for estimating the weights is considered approximately correct in that it selects the "right" subset of covariates as effect modifiers. The estimated weights are adequate for bias removal because the balancing property holds,[76,77,78,79] that is, conditional on the weights, the effect modifier means are balanced between the two trials, and one can potentially achieve unbiased estimation of treatment effects in the *BC* population.
- Simulated treatment comparison: a Cox proportional hazards regression on survival time is fitted to the IPD, with the IPD effect modifiers centered at the *BC* mean values. The outcome regression is correctly specified. We include all of the covariates in the regression but only center the effect modifiers.
- The Bucher method[4] gives the standard indirect comparison. We know that this will be biased as it does not adjust for the bias induced by the imbalance in effect modifiers.

In all methods, the variances of the within-trial relative effects are summed to estimate the variance of the *A* versus *B* treatment effect, $\hat{V}\left(\hat{\Delta}^*_{AB}\right)$. Confidence intervals are constructed using normal distributions: $\hat{\Delta}^*_{AB} \pm 1.96\sqrt{\hat{V}\left(\hat{\Delta}^*_{AB}\right)}$, assuming relatively large *N*.

4.5 | Performance measures

We generate and analyze 1000 Monte Carlo replicates of trial data per simulation scenario. Let $\hat{\Delta}^*_{AB,s}$ denote the estimator for the *s*-th Monte Carlo replicate and let $\mathbb{E}\left(\hat{\Delta}^*_{AB}\right)$ denote its mean across the 1000 simulations. Based on a test run of the method and simulation scenario with the highest long-run variability (MAIC under Scenario 109), we assume that $\mathrm{SD}\left(\hat{\Delta}^*_{AB}\right) \leq 0.45$ and that, conservatively, the variance across simulations of the estimated treatment effect is always less than approximately 0.2. Given that the Monte Carlo standard error (MCSE) of the bias is equal to $\sqrt{\mathrm{Var}\left(\hat{\Delta}^*_{AB}\right)/N_{\mathrm{sim}}}$, where N_{sim} is the number of simulations, it is at most 0.014 under 1000 simulations. We consider the degree of precision provided by the MCSE to be acceptable in relation to the size of the effects. If the empirical coverage rate of the methods is 95%, $N_{\mathrm{sim}} = 1000$ implies that the MCSE of the coverage is $\sqrt{(95 \times 5)/1000} = 0.69\%$, with the worst-case MCSE being 1.58% under 50% coverage. We also consider this degree of precision to be acceptable. Hence, the simulation study is conducted under $N_{\mathrm{sim}} = 1000$.

The following criteria are considered jointly to assess the methods' performances. MCSEs are estimated for each performance metric in order to quantify the simulation uncertainty due to using a finite number of simulation replicates.

- To assess aim 1, we compute the **bias** in the estimated treatment effect

$$\mathbb{E}\left(\hat{\Delta}^*_{AB} - \Delta^*_{AB}\right) = \frac{1}{1000}\sum_{s=1}^{1000}\hat{\Delta}^*_{AB,s} - \Delta^*_{AB}.$$

As $\Delta^*_{AB} = 0$, the bias is equal to the average estimated treatment effect across the simulations. The MCSE of the bias is estimated as $\sqrt{\frac{1}{1000\times 999}\sum_{s=1}^{1000}\left(\hat{\Delta}^*_{AB,s} - \mathbb{E}\left(\hat{\Delta}^*_{AB}\right)\right)}$.

- To assess aim 2, we calculate the *variability ratio* of the treatment effect estimate, defined[80] as the ratio of the average model standard error and the observed standard deviation of the treatment effect estimates (empirical standard error):

$$\mathrm{VR}\left(\hat{\Delta}^*_{AB}\right) = \frac{\frac{1}{1000}\sum_{s=1}^{1000}\sqrt{\hat{V}\left(\hat{\Delta}^*_{AB,s}\right)}}{\sqrt{\frac{1}{999}\sum_{s=1}^{1000}\left(\hat{\Delta}^*_{AB,s} - \mathbb{E}\left(\hat{\Delta}^*_{AB}\right)\right)^2}}. \quad (6)$$

VR being greater than (or smaller) than one suggests that, on average, standard errors overestimate (or underestimate) the variability of the treatment effect estimate. It is important to note that this metric assumes that the correct estimand and corresponding variance are being targeted. A variability ratio of one is of little use if this is not the case, for example, if both the model standard

errors and the empirical standard errors are taken over estimates targeting the wrong estimand. The MCSE of the variability ratio is approximated as:

$$\sqrt{\frac{\frac{1}{1000}\sum_{s=1}^{1000}\left(\sqrt{\hat{V}\left(\hat{\Delta}_{AB,s}\right)}-\mathbb{E}\left(\sqrt{\hat{V}\left(\hat{\Delta}_{AB}\right)}\right)\right)^2}{999\times \mathrm{ESE}\left(\hat{\Delta}^*_{AB}\right)^2}+\frac{\left(\frac{1}{1000}\sum_{s=1}^{1000}\sqrt{\hat{V}\left(\hat{\Delta}^*_{AB,s}\right)}\right)^2}{2\times 999\times \mathrm{ESE}\left(\hat{\Delta}^*_{AB}\right)^2}},$$

where $\mathrm{ESE}\left(\hat{\Delta}^*_{AB}\right)$ is the estimated empirical standard error, which is the denominator in Equation (6).

- Aim 3 is assessed using the *coverage* of confidence intervals, estimated as the proportion of times that the true treatment effect is enclosed in the $(100\times(1-\alpha))\%$ confidence interval of the estimated treatment effect, where $\alpha=0.05$ is the nominal significance level. The MCSE of the coverage is computed as $\sqrt{\frac{\mathrm{Cover}\left(\hat{\Delta}^*_{AB}\right)\times\left(1-\mathrm{Cover}\left(\hat{\Delta}^*_{AB}\right)\right)}{1000}}$, where $\mathrm{Cover}\left(\hat{\Delta}^*_{AB}\right)$ is the estimated coverage percentage.
- We use *empirical standard error* (ESE) to assess aim 4 as it measures the precision or long-run variability of the treatment effect estimate. The ESE is defined above, as the denominator in Equation (6). The MCSE of the empirical standard error is estimated as $\frac{\mathrm{ESE}\left(\hat{\Delta}^*_{AB}\right)}{\sqrt{2\times 999}}$.
- The *mean square error* (MSE) of the estimated treatment effect

$$\begin{aligned}\mathrm{MSE}\left(\hat{\Delta}^*_{AB}\right)&=\mathbb{E}\left[\left(\hat{\Delta}^*_{AB}-\Delta^*_{AB}\right)^2\right]\\&=\frac{1}{1000}\sum_{s=1}^{1000}\left(\hat{\Delta}^*_{AB,s}-\Delta^*_{AB}\right)^2,\end{aligned}$$

provides a summary value of overall accuracy (efficiency), integrating elements of bias (aim 1) and variability (aim 4). The Monte Carlo standard error of the MSE is computed as $\sqrt{\frac{\sum_{s=1}^{1000}\left[\left(\hat{\Delta}^*_{AB,s}-\Delta^*_{AB}\right)^2-\mathrm{MSE}\left(\hat{\Delta}^*_{AB}\right)\right]^2}{1000\times 999}}$, where $\mathrm{MSE}\left(\hat{\Delta}^*_{AB}\right)$ is the estimated mean square error.

5 | RESULTS

The performance measures across all 162 simulation scenarios are illustrated in Figures 4–8 using nested loop plots,[81] which arrange all scenarios into a lexicographical order, looping through nested factors. In the nested sequence of loops, we consider first the parameters with the largest perceived influence on the performance metric. Notice that this order is considered on a case-by-case basis for each performance measure. Given the large number of simulation scenarios, depiction of Monte Carlo standard errors, quantifying the simulation uncertainty, is difficult. The Monte Carlo standard errors of each performance metric are reported in Supplementary Appendix C of the Supporting Information. In MAIC, 1 of 162,000 weighted regressions had a separation issue, that is, there is a total lack of covariate overlap (Scenario 115, with $N=150$). Results for this replicate were discarded. The outcome regressions converged for all replicates in STC and the Bucher method.

5.1 | Unbiasedness of treatment effect

The impact of the bias will depend on the uncertainty in the estimated treatment effect,[82,83] measured by the empirical standard error. To assess such impact, we consider standardizing the biases[83] by computing these as a percentage of the empirical standard error. In a review of missing data methods, Schafer and Graham[82] consider bias to be troublesome under 1000 simulations if its absolute size is greater than about one half of the estimate's empirical standard error, that is, the standardized bias has magnitude greater than 50%. Under this rule of thumb, MAIC does not produce problematic biases in any of the simulation scenarios. On the other hand, STC and the Bucher method generate problematic biases in 71 of 162 scenarios, and in 147 of 162 scenarios, respectively. The biases in MAIC do not appear to have any practical significance, as they do not degrade coverage and efficiency.

Figure 4 shows the bias for the methods across all scenarios. MAIC is the least biased method, followed by STC and the Bucher method. In the scenarios considered in this simulation study, STC produces negative bias when the interaction effects are moderate and positive bias when they are very strong. In addition, biases vary more widely when prognostic effects are larger. When interaction effects are weaker, stronger prognostic effects shift the bias negatively. This degree of systematic bias arises from the non-collapsibility of the (log) hazard ratio (see Section 3.3).

In some cases, for example, under very strong prognostic variable effects and moderate effect-modifying interactions, STC even has increased bias compared to the Bucher method. In other scenarios, for example, where there are strong effect-modifying interactions and

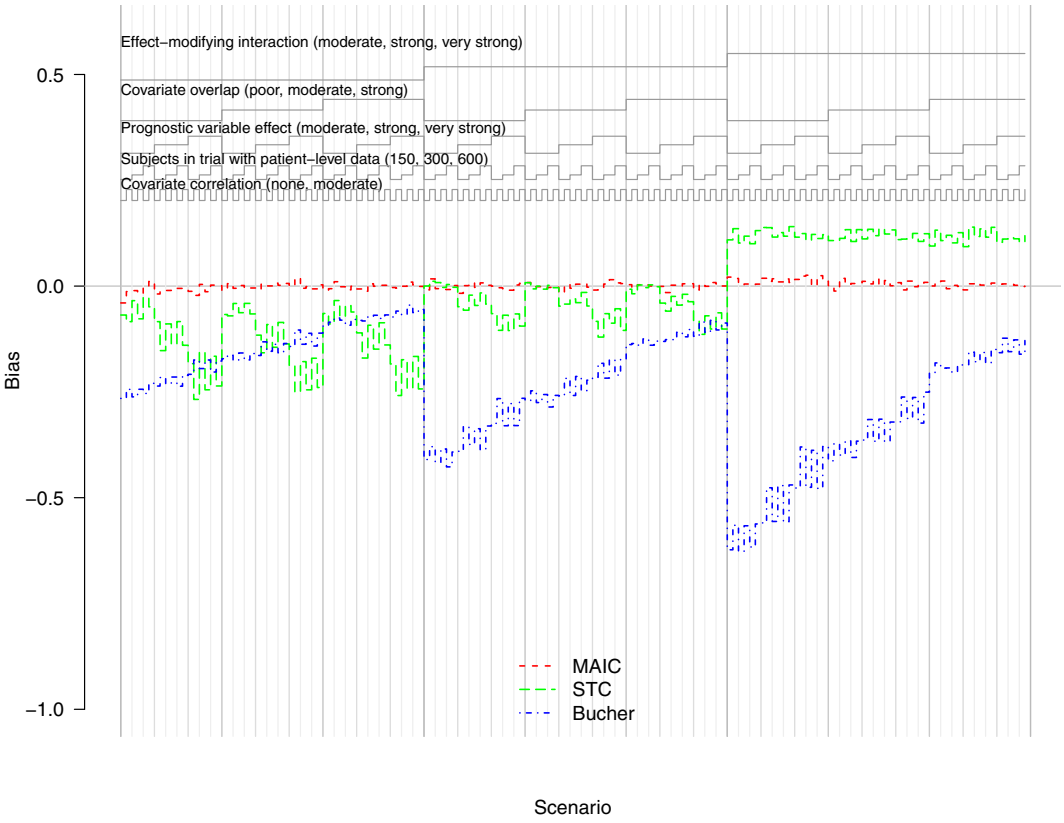

FIGURE 4 Bias across all simulation scenarios. The nested loop plot arranges all 162 scenarios into a lexicographical order, looping through nested factors. In the nested sequence of loops, we consider first the parameters with the largest perceived influence on the performance metric. MAIC, matching-adjusted indirect comparison; STC, simulated treatment comparison [Colour figure can be viewed at wileyonlinelibrary.com]

moderate or strong prognostic variable effects, STC estimates are virtually unbiased. This is because, in these scenarios, the average conditional and marginal treatment effects for A versus C are almost identical and hence the non-collapsibility of the measure of effect is not an issue. It is worth noting that conclusions arising from the interpretation of patterns in Figure 4 for STC are by-products of non-collapsibility. Any generalization should be cautious.

As expected, the strength of interaction effects is an important driver of bias in the Bucher method and the incurred bias increases with greater covariate imbalance. This is because the more substantial the imbalance in effect modifiers and the greater their interaction with treatment, the larger the bias of the unadjusted comparison. The impact of these factors on the bias appears to be slightly reduced when prognostic effects are stronger and contribute more "explanatory power" to the outcome. Varying the number of patients in the AC trial does not seem to have any discernible impact on the bias for any method. Biases in MAIC seem to be unaffected when varying the degree of covariate imbalance/overlap.

5.2 | Unbiasedness of variance of treatment effect

In the Bucher method, the variability ratio is close to one under the vast majority of simulation scenarios (Figure 5). This suggests that standard error estimates for the methods are unbiased, that is, that the model standard errors coincide with the empirical standard errors. In STC, variability ratios are generally close to one under $N = 300$ and $N = 600$, and any bias in the estimated variances appears to be negligible. However, the variability ratios decrease when the AC sample size is small ($N = 150$). In these scenarios, there is some underestimation of variability by the model standard errors. It is important to recall that this metric assumes that the correct estimand and corresponding variance are being targeted. This is not the case in our application of STC, in the sense that both model standard errors and empirical standard errors are taken over an incompatible indirect treatment comparison. MAIC standard errors underestimate variability when $N = 150$, and also when covariate overlap is poor, in which case underestimation under $N = 150$ is exacerbated. Under the smallest sample size

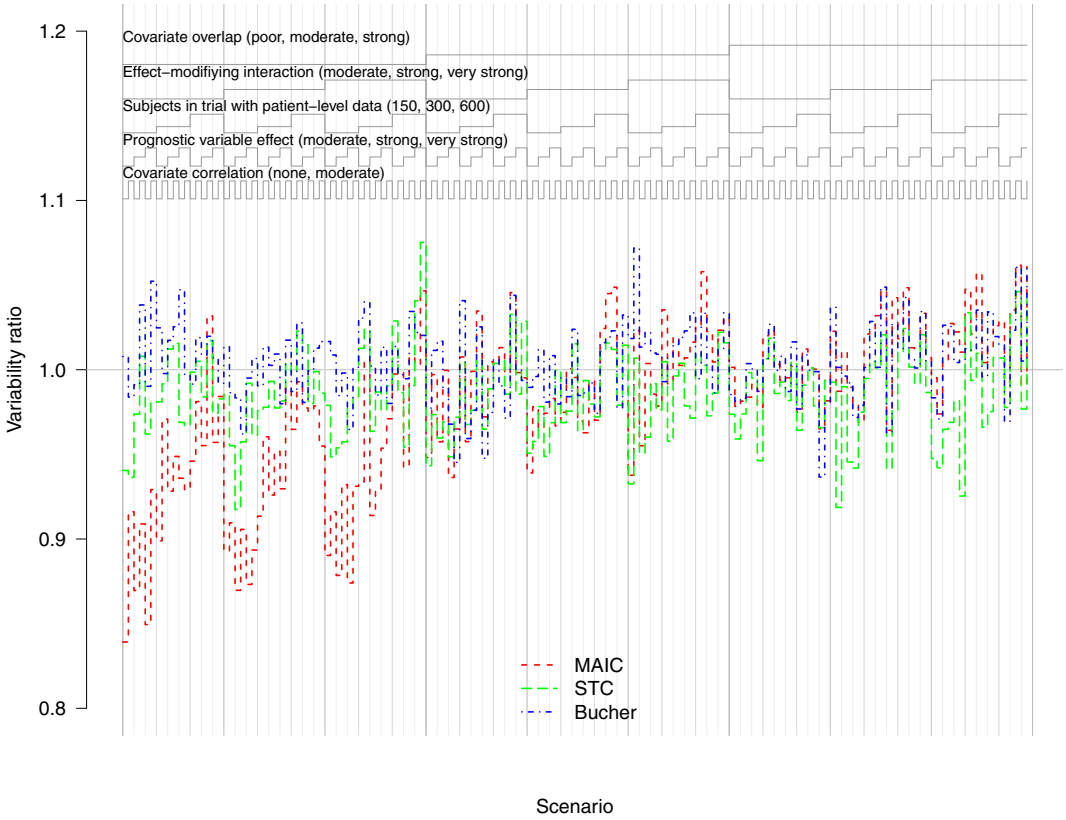

FIGURE 5 Variability ratio across all simulation scenarios. MAIC, matching-adjusted indirect comparison; STC, simulated treatment comparison [Colour figure can be viewed at wileyonlinelibrary.com]

and poor covariate overlap, variability ratios are often below 0.9, with model standard errors underestimating the empirical standard errors. This is likely due to the robust sandwich estimator used to derive the standard errors. In the literature, this has exhibited an underestimation of variability in small samples.[84,85] The understated uncertainty is an issue, as it will be propagated through the cost-effectiveness analysis and may lead to inappropriate decision-making.[86]

5.3 | Randomization validity

From a frequentist viewpoint,[87] 95% confidence intervals are randomization-valid if these are guaranteed to include the true treatment effect at least 95% of the time. This means that the empirical coverage rate should be approximately equal to the nominal coverage rate, in this case 0.95 for 95% confidence intervals, to obtain appropriate type I error rates for testing a "no effect" null hypothesis. Theoretically, the empirical coverage rate is statistically significantly different to 0.95 if, roughly, it is less than 0.9365 or more than 0.9635, assuming 1000 independent simulations per scenario. These values differ by approximately two standard errors from the nominal coverage rate. When randomization validity cannot be attained, one would at least expect the interval estimates to be confidence-valid, that is, the 95% confidence intervals include the true treatment effect *at least* 95% of the time.

In general, empirical coverage rates for MAIC do not overestimate the advertised nominal coverage rate. Only 4 of 162 scenarios have a rate above 0.9635. On the other hand, empirical coverage rates are significantly below the nominal coverage rate when the *AC* sample size is low ($N = 150$) and under poor covariate overlap. With $N = 150$, 24 of 54 coverage rates are below 0.9365. When covariate overlap is poor, 38 of 54 coverage rates are below 0.9365—18 of these under $N = 150$. When there is both poor overlap and a low *AC* sample size, coverage rates for MAIC are inappropriate: these may even fall below 90%, that is, at least double the nominal rate of error. Poor coverage rates are a decomposition of both the bias and the standard error used to compute the width of the confidence intervals. It is not bias that degrades the coverage rates for this method but the standard error underestimation mentioned in Section 5.2. Poor coverage is induced by the standard errors used in the construction of the confidence intervals.

Confidence intervals from the Bucher method are not confidence-valid for virtually all scenarios. Coverage rates

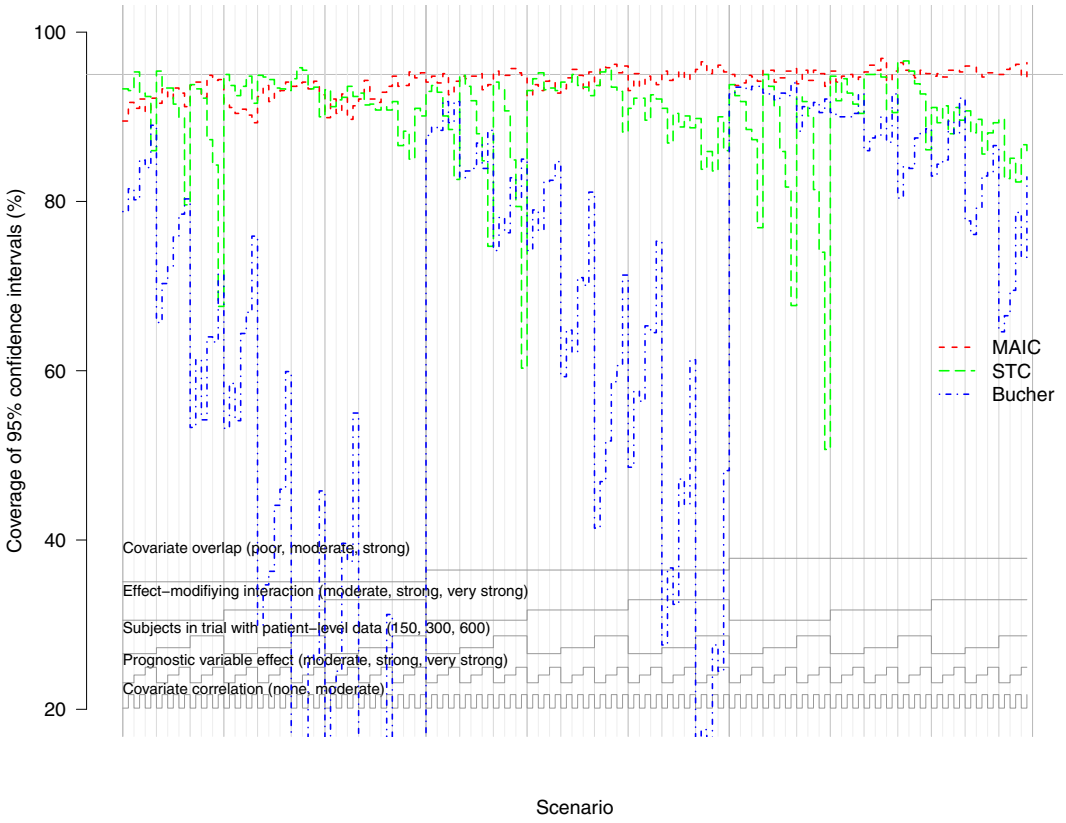

FIGURE 6 Empirical coverage percentage of 95% confidence intervals across all simulation scenarios. MAIC, matching-adjusted indirect comparison; STC, simulated treatment comparison [Colour figure can be viewed at wileyonlinelibrary.com]

deteriorate markedly under the most important determinants of bias. When there is greater imbalance between the covariates and when interaction effects are stronger, the induced bias is larger and coverage rates are degraded. Under very strong interactions with treatment, empirical coverage may drop below 50%. Therefore, the Bucher method will incorrectly detect significant results a large proportion of times in these scenarios. Such overconfidence will lead to very high type I error rates for testing a "no effect" null hypothesis.

5.4 | Precision and efficiency

Several trends are revealed upon visual inspection of the empirical standard error across scenarios (Figure 7). As expected, the ESE decreases for all methods (i.e., the estimate is more precise) as the number of subjects in the *AC* trial increases. The strengths of interaction effects and of prognostic variable effects appear to have a negligible impact on the precision of population adjustment methods.

The degree of covariate overlap has an important influence on the ESE and population adjustment methods incur losses of precision when covariate overlap is poor. When overlap is poor, there exists a subpopulation in *BC* that does not overlap with the *AC* population. Therefore, inferences in this subpopulation rely largely on extrapolation. Regression adjustment methods such as STC require greater extrapolation when the covariate overlap is poorer.[15] In reweighting methods such as MAIC, extrapolation is not even possible. When covariate overlap is poor, observations in the *AC* patient-level data (those that are not covered by the range of the effect modifiers in the *BC* population) are assigned very low weights (low odds of enrollment in *BC* vs. *AC*). On the other hand, the relatively small number of units in the overlapping region of the covariate space are assigned very large weights, dominating the reweighted sample. These extreme weights lead to large reductions in ESS and to the deterioration of precision and efficiency.

In MAIC, the presence of correlation mitigates the effect of decreasing covariate overlap on a consistent basis. This is due to the correlation increasing the overlap between the joint covariate distributions of *AC* and *BC*, lessening the reduction in effective sample size and providing greater stability to the estimates. ESE for the Bucher method does not vary across different degrees of covariate imbalance, as these are not considered by the method, and overprecise estimates are produced.

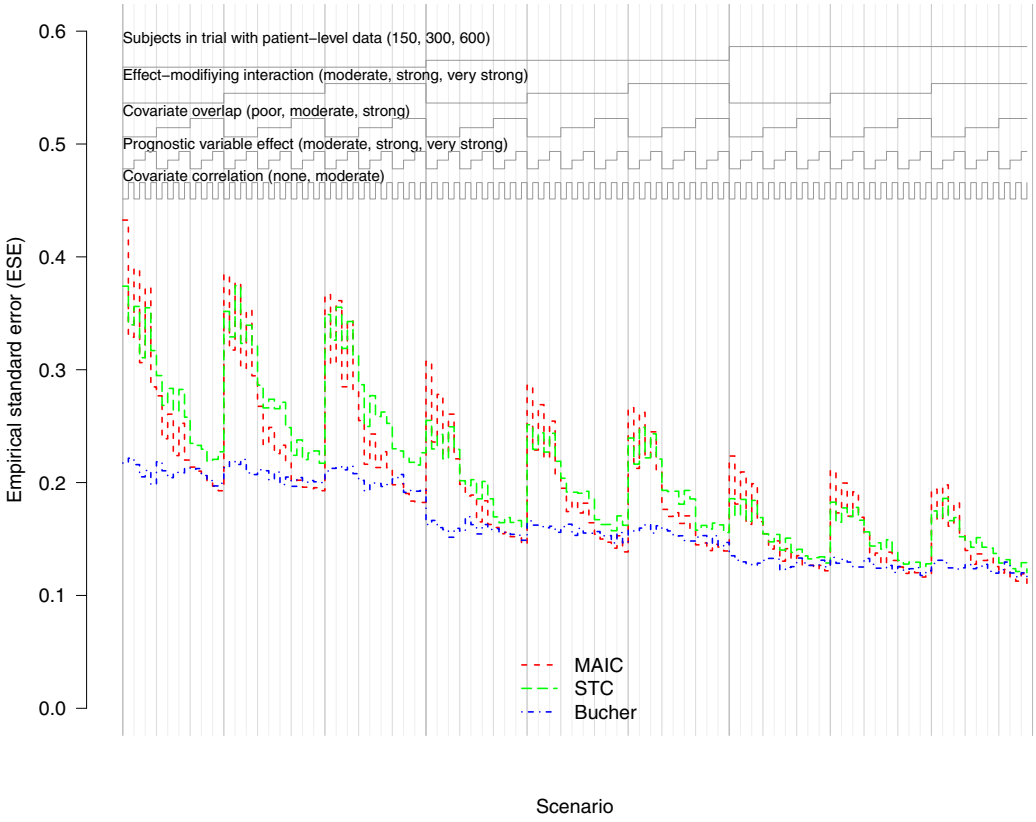

FIGURE 7 Empirical standard error across all simulation scenarios. MAIC, matching-adjusted indirect comparison; STC, simulated treatment comparison [Colour figure can be viewed at wileyonlinelibrary.com]

Contrary to ESE, MSE also takes into account the true value of the estimand as it incorporates the bias. Hence, main drivers of bias and ESE are generally key properties for MSE. Figure 8 is inspected in order to explore patterns in the mean square error. Estimates are less accurate for MAIC when prognostic variable effects are stronger, *AC* sample sizes are smaller and covariate overlap is poorer. As bias is negligible for MAIC, precision is the driver of accuracy. On the contrary, as the Bucher method is systematically biased and overprecise, the driver of accuracy is bias. Poor accuracy in STC is also driven by bias, particularly under low sample sizes and strong prognostic variable effects. STC was consistently less accurate than MAIC, with larger mean square errors in all simulation scenarios. In some cases where the STC bias was strong, for example, very strong prognostic variable effects and moderate effect-modifying interactions, STC even increased the MSE compared to the Bucher method.

In accordance with the trends observed for the ESE, the MSE is also very sensitive to the value of N and decreases for all methods as N increases. We highlight that the number of subjects in the *BC* trial (not varied in this simulation study) is a less important performance driver than the number of subjects in *AC*; while it contributes to sampling variability, the reweighting or regressions are performed in the *AC* patient-level data.

6 | DISCUSSION

In this section, we discuss the implications of, and recommendations for, performing population adjustment, based on the simulation study. Finally, we highlight potential limitations of the simulation study, primarily relating to the extrapolation of its results to practical guidance. We have seen in Section 5 that the typical use of STC produces systematic bias as a result of the non-collapsibility of the log hazard ratio. The estimate $\hat{\Delta}^{*}_{AC}$ targets a conditional treatment effect that is incompatible with the estimate $\hat{\Delta}_{BC}$. This leads to bias in estimating the marginal treatment effect for A versus B, despite all assumptions for population adjustment being met. Given the clear inadequacy of STC in this setting, we focus on MAIC as a population adjustment method.

An important future objective would be the development of an alternative formulation to STC that estimates a marginal treatment effect for A versus C. A crucial additional step, missing from the current implementation, is to integrate or average the conditional effect

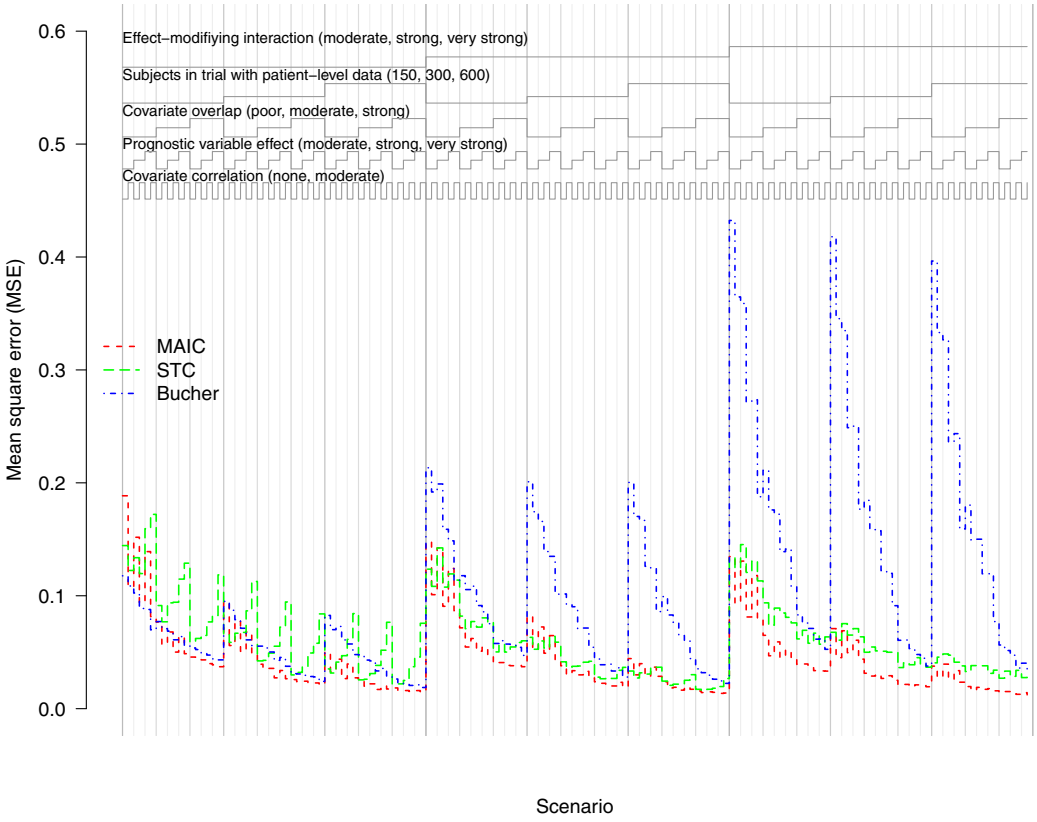

FIGURE 8 Mean square error across all simulation scenarios. MAIC, matching-adjusted indirect comparison; STC, simulated treatment comparison [Colour figure can be viewed at wileyonlinelibrary.com]

estimates over the *BC* covariates. Then, STC could potentially obtain a marginal treatment effect estimate that is comparable to the marginal *B* versus *C* estimate published in the *BC* study. This would avoid the bias caused by incompatibility in the indirect comparison and provide inference for the marginal treatment effect for *A* versus *B* in the *BC* population. During the preparation of this manuscript, a novel regression adjustment method named multilevel network meta-regression (ML-NMR) has been introduced.[88,89] ML-NMR targets a conditional treatment effect but directly avoids the compatibility issues of STC and is also applicable in treatment networks of any size with the two-study scenario as a special case. Using an averaging or integration step, ML-NMR could also be adapted to target a marginal treatment effect.[90]

6.1 | Bias-variance trade-offs

Before performing population adjustment, it is important to assess the magnitude of the bias induced by effect modifier imbalances. Such bias depends on the degree of covariate imbalance and on the strength of interaction effects, that is, the effect modifier status of the covariates. The combination of these two factors determines the level of bias reduction that would be achieved with population adjustment.

Inevitably, due to bias-variance trade-offs, the increase in variability that we are willing to accept with population adjustment depends on the magnitude of the bias that would be corrected. Such variability is largely driven by the degree of covariate overlap and by the *AC* sample size. Hence, while the potential extent of bias correction increases with greater covariate imbalance, so does the potential imprecision of the treatment effect estimate (assuming that the imbalance is accompanied by poor overlap).

In our simulation study, under no failures of assumptions, this trade-off always favors the bias correction offered by MAIC over the precision of the Bucher method, implying that the reductions in ESS based on unstable weights are worth it, even under stronger covariate overlap. Across scenarios, the relative accuracy of MAIC with respect to that of the Bucher method improves under greater degrees of covariate imbalance and poorer overlap. It is worth noting that, even in scenarios where the Bucher method is relatively accurate, it is still flawed in the context of decision-making due to overprecision and undercoverage.

The magnitude of the bias that would be corrected with population adjustment also depends on the strength of interaction effects, that is, the effect modifier status of the covariates. In the simulation study, the lowest effect-modifying interaction coefficient was $-\ln(0.67) = 0.4$. Despite the relatively low magnitude of bias induced in this setting, MAIC was consistently more efficient than the Bucher method. Larger interaction effects warrant greater bias reduction but do not degrade the precision of the population-adjusted estimate. Hence, the relative accuracy of MAIC with respect to the Bucher method improves further as the effect-modifying coefficients increase.

6.2 | Justification of effect modifier status

In the simulation study, we know that population adjustment is required as we set the cross-trial imbalances between covariates and have specified some of these as effect modifiers. Most applications of population adjustment present evidence of the former, for example, through tables of baseline characteristics with covariate means and proportions ("Table 1" in a RCT publication). However, quantitative evidence justifying the effect modifier status of the selected covariates is rarely brought forward. Presenting this type of supporting evidence is very important when justifying the use of population adjustment.

Typically, the selection of effect modifiers is supported by clinical expert opinion. However, clinical expert judgment and subject-matter knowledge are fallible when determining effect modifier status because: (1) the therapies being evaluated are often novel; and (2) effect modifier status is scale-specific—clinical experts may not have the mathematical intuition to assess whether covariates are effect modifiers on the linear predictor scale (as opposed to the natural outcome scale).

Therefore, applications of population adjustment often balance all available covariates on the grounds of expert opinion. This is probably because the clinical experts cannot rule out bias-inducing interactions with treatment for any of the baseline characteristics. Almost invariably, the level of covariate overlap and precision will decrease as a larger number of covariates are accounted for in the analysis. Presenting quantitative evidence along with clinical expert opinion would help establish whether adjustment is necessary for each covariate.[91]

As proposed by Phillippo et al.,[6] we encourage the analyst to fit regression models with interaction terms to the IPD for an exploratory assessment of effect modifier status. One possible strategy is to consider each potential effect modifier one-at-a-time by adding the corresponding interaction term to the main (treatment) effect model.[47] Then, the interaction coefficient can be multiplied by the difference in effect modifier means to gauge the level of induced bias.[15] This analysis should be purely exploratory, since individual trials are typically underpowered for interaction testing.[92,93] The dichotomization or categorization of continuous variables, the poor representation of a variable, for example, a limited age range, and incorrectly assuming linearity may dilute interactions further.

Meta-analyses of multiple trials, involving the same outcome and similar treatments and conditions, provide greater power to detect interactions, particularly using IPD.[93,94] With unavailable IPD, it may still be possible to conduct an IPD meta-analysis if the owners of the data are willing to provide the interaction effects,[95] or one may conduct an ALD meta-analysis if covariate-treatment interactions are included in the clinical trial reports.[92] In any case, the identification of effect modifiers is in essence observational,[96,97] and requires much more evidence than demonstrating a main treatment effect.[98] Therefore, it may be reasonable to balance a variable if there is a strong biological rationale for effect modification, even if the interaction is statistically weak, for example, the p-value is large and the null hypothesis of interaction is not rejected.[98]

6.3 | Nuances in the interpretation of results

It is worth noting that the conclusions of this simulation study are dependent on the outcome and model type. We have considered survival outcomes and the Cox proportional hazards model, as these are the most prevalent outcome type and modeling framework in MAIC and STC applications. However, further simulation studies are required with alternative outcome types and models. For example, exploratory simulations with binary outcomes and logistic regression have found that the performance of MAIC is more affected by low sample sizes and poor covariate overlap than seen for survival outcomes. This is likely due to logistic regression being less efficient[99] and more prone to small-sample bias[100] than Cox regression.

Furthermore, we have only considered and adjusted for two effect modifiers that induce bias in the same direction, that is, the effect modifiers in a given study have the same means, the cross-trial differences in means are in the same direction, and the interaction effects are in the same direction. In real applications of population adjustment, it is not uncommon to see more than

10 covariates being balanced.[25] As this simulation study considered percentage reductions in effective sample size for MAIC that are representative of scenarios encountered in NICE TAs, real applications will likely have imbalances for each individual covariate that are smaller than those considered in this study. In addition, the means for the effect modifiers within a given study will differ, with the mean differences across studies and/or the effect-modifying interactions potentially being in opposite directions. Therefore, the induced biases could cancel out but, then again, this is not directly testable in a practical scenario.

6.4 | Potential failures in assumptions

Most importantly, all the assumptions required for indirect treatment comparisons and valid population adjustment hold, by design, in the simulation study. While the simulation study provides proof-of-principle for the methods, it does not inform how robust these are to failures in assumptions. Population-adjusted analyses create additional complexity since they require a larger number of assumptions than standard indirect comparisons. The additional assumptions are hard to meet and most of them are not directly testable. It is important that researchers are aware of these, as their violation may lead to biased estimates of the treatment effect. In practice, we will never come across an idealistic scenario in which all assumptions perfectly hold. Therefore, researchers should exercise caution when interpreting the results of population-adjusted analyses. These should not be taken directly at face value, but only as tools to simplify a complex reality.

Firstly, MAIC, STC and the Bucher method rely on trials *AC* and *BC* being internally valid, implying appropriate designs, the absence of non-compliance, proper randomization and reasonably large sample sizes. Secondly, all indirect treatment comparisons (standard or population-adjusted) rely on consistency under parallel studies, that is, potential outcomes are homogeneous for a given treatment regardless of the study assigned to a subject. For instance, treatment *C* should be administered in the same setting in both trials, or differences in the nature of treatment should not change its effect. This means that MAIC and STC cannot account for cross-trial differences that are perfectly confounded with the nature of treatments, for example, treatment administration or dosing formulation. MAIC and STC can only account for differences in the characteristics of the trial populations.

In practice, the additional assumptions made by MAIC and STC may be problematic. Firstly, it assumed that all effect modifiers for treatment *A* versus *C* are adjusted for.‡‡ By design, the simulation study assumes that complete information is available for both trials and that all effect modifiers have been accounted for. In practice, this assumption is hard to meet—it is difficult to ascertain the effect modifier status of covariates, particularly for new treatments with limited prior empirical evidence and clinical domain knowledge. Hence, the analyst may select the effect modifiers incorrectly. In addition, information on some effect modifiers could be unmeasured or unpublished for one of the trials. The incorrect omission of effect modifiers leads to the wrong specification of the trial assignment logistic regression model in MAIC, and of the outcome regression in STC. Relative effects will no longer be conditionally constant across trials and this will lead MAIC and STC to produce biased estimates.

In the simulation study, we know the correct data-generating mechanism, and are aware of which covariates are purely prognostic variables and which covariates are effect modifiers. This is something that one cannot typically ascertain in practice. Exploratory simulations show that the relative precision and accuracy of MAIC deteriorate, with respect to STC and the Bucher method, if we treat all four covariates as effect modifiers. This is due to the loss of effective sample size and inflation of the standard error due to the over-specification of effect modifiers.

Alternatively, it is more burdensome to specify the outcome regression model for STC than the propensity score model for MAIC; the outcome regression requires specifying both prognostic and interaction terms, while the trial assignment model in MAIC only requires the specification of effect modifiers. The relative precision and accuracy of STC deteriorate if the terms corresponding to the purely prognostic covariates are not included in the outcome regression. Nevertheless, this does not alter the conclusions of the simulation study: the other terms in the outcome regression already account for a considerable portion of the variability of the outcome and relative effects have very similar accuracy in any case.

Another assumption made by MAIC and STC, that holds in this simulation study, is that there is some overlap between the ranges of the selected covariates in *AC* and *BC*. In population adjustment methods, the indirect comparison is performed in the *BC* population. This implies that the ranges of the covariates in the *BC* population should be covered by their respective ranges in the *AC* trial. In practice, this assumption may break down if the inclusion/exclusion criteria of *AC* and *BC* are inconsistent. When there is no overlap, weighting methods like MAIC are unable to extrapolate beyond the *AC* population, and may not even produce an estimate. However,

STC can extrapolate beyond the covariate space observed in the *AC* patient-level data, using the linearity assumption or other appropriate assumptions about the input space. Note that the validity of the extrapolation depends on accurately capturing the true covariate-outcome relationships. We view extrapolation as a desirable property because poor overlap, with small effective sample sizes and large percentage reductions in effective sample size, is a pervasive issue in health technology appraisals.[25]

MAIC and STC make certain assumptions about the joint distribution of covariates in *BC*. Where no correlation information is available for the *BC* study, both methods seem to assume that the joint *BC* covariate distribution is the product of the published marginal distributions. The implicit assumptions are, in fact, more nuanced. In MAIC, as stated in the NICE Decision Support Unit Technical Support Document,[15] "when covariate correlations are not available from the (*BC*) population, and therefore cannot be balanced by inclusion in the weighting model, they are assumed to be equal to the correlations among covariates in the pseudo-population formed by weighting the (*AC*) population." In the typical usage of STC, the correlations between the *BC* covariates are assumed to be equal to the correlations between covariates in the *AC* study. In the "covariate simulation" approach to STC, discussed in Section 3.2, this assumption is also made, albeit more explicitly, if the correlation structure observed in the *AC* IPD is used to simulate the covariates. In an anchored comparison, only effect-modifying covariates need balancing, so the assumptions can be relaxed to only include effect modifiers. This set of assumptions will only induce bias if higher-order interactions (involving two or more covariates) are unaccounted for or misspecified. If these interactions are not included in the weighting model for MAIC or in the outcome regression for STC, the specification of pairwise correlations will not make a difference in terms of bias, as observed in a recent simulation study that investigates this set of assumptions.[88]

All indirect treatment comparisons should be performed and are typically conducted on the linear predictor scale,[15] upon which the effect of treatment is assumed to be additive. We have assumed that the effect modifiers have been defined on the linear predictor scale and are additive on this scale. In the simulation study, it is known that effect modification is linear on the log hazard ratio scale. A central component of population-adjusted indirect comparisons is the specification of a model that is typically parametric. That is the propensity score model for the weights in MAIC or the outcome regression in STC. Parametric modeling assumptions may not be appropriate in real applications, where there is a danger of model misspecification. This is more evident in a regression adjustment method like STC, where an explicit outcome regression is formulated. The parametric model depends on functional form assumptions that will be violated if the covariate-outcome relationships are not correctly captured.

Even though the logistic regression model for the weights in MAIC does not make reference to the outcome, MAIC is also susceptible to model misspecification bias, albeit in a more implicit form. The model for estimating the weights is approximately correct in the simulation study because the right subset of covariates has been selected as effect modifiers and the balancing property holds for the weights, as mentioned in Section 4.4. In practice, the model will be incorrectly specified if this is not the case, potentially leading to a biased estimate. Scale conflicts may also arise if effect modification status, which is scale-specific, has been justified on the wrong scale, for example, when treatment effect modification is specified as linear but is non-linear or multiplicative, for example, age in cardiovascular disease treatments. Note that, in practice, we find that it may be more difficult to specify a correct parametric model for the outcome than an approximately correct parametric model for the trial assignment weights.

Finally, population-adjusted indirect comparisons only produce an estimate $\hat{\Delta}^{*}_{AB}$ that is valid in the *BC* population, which may not match the target population for the decision unless an additional assumption is made. This is the shared effect modifier assumption,[15] described in Section 4.3. This assumption is met by the simulation study and is required to transport the treatment effect estimate to any given target population. However, it is untestable for MAIC and STC with the data available in practice. Shared effect modification is hard to meet if the competing interventions do not belong to the same class, and have dissimilar mechanisms of action or clinical properties. In that case, there is little reason to believe that treatments *A* and *B* have the same set of effect-modifying covariates and that these interact with active treatment in the same way in *AC* and *BC*. It is worth noting that the target population may not match the *AC* and *BC* trial populations and may be more akin to the joint covariate distribution observed in a registry/cohort study or some other observational dataset. Policy-makers could use such data to define a target population for a specific outcome and disease area into which all manufacturers could conduct their indirect comparisons. This would help relax the shared effect modifier assumption.

Given the large number of assumptions made by population-adjusted indirect comparisons, future simulation studies should assess the robustness of the methods to failures in assumptions under different degrees of data availability and model misspecification.

6.5 | Variance estimation in MAIC

MAIC was generally randomization-valid, except in situations with poor covariate overlap and small sample sizes, where robust sandwich standard errors underestimated empirical estimates of the standard error and, consequently, there was undercoverage. MAIC exhibited variability ratios below 0.9 in scenarios with the smallest sample size and poor covariate overlap. In these scenarios, confidence intervals were narrow, achieving coverage rates which were statistically significantly below 95% and sometimes dropping below 90%. As mentioned in Section 5.2, this is probably due to the robust sandwich estimator used to derive the standard errors, which has previously underestimated variability in small samples in simulation studies.[84,85] It is worth noting that this estimator is based on large-sample (asymptotic) arguments and infinite populations. Therefore, it is not surprising that performance is poor under the smallest effective sample sizes, which occur where the *AC* trial sample size is small and covariate overlap is poor. Where effective sample sizes are small, confidence intervals derived from robust sandwich variance estimators should be interpreted cautiously, as these may understate uncertainty and this underestimation will be propagated through the cost-effectiveness analysis, potentially leading to inappropriate decision-making.

This robust variance estimator is easy to use by analysts performing MAIC (and propensity score weighting, in general) because it is computationally efficient and is typically implemented in standard routines in statistical computing software such as R. For instance, in R, by setting robust = TRUE in the coxph function, built in the survival package[75] for survival analysis, or using the sandwich package[102] for the treatment coefficient of a weighted generalized linear model. It is worth noting that these readily available implementations assume that the weights are fixed or known and do not account for the uncertainty in the estimation of the weights.

In principle, one could circumvent this issue by using the bootstrap to obtain the variance and confidence intervals of the *A* versus *C* treatment effect, as in the simulation study by Petto et al.[20] or in the article by Sikirica et al.[45] Bootstrap methods are beneficial because they can account for the variability of the estimated weights and are straightforward to implement, potentially providing unbiased variance estimators with a large number of resamples. However, bootstrapping is orders of magnitude more expensive computationally than applying the closed-form sandwich variance estimator. In addition, bootstrap resampling procedures are inherently random and exhibit some seed-dependence, which is only mitigated by increasing the number of resamples and computational demand. Future simulation studies should compare different approaches to variance estimation and assess whether implementations of the bootstrap can compete with the robust sandwich estimator.

Another alternative to variance estimation is the development of closed-form robust sandwich estimators that properly account for the uncertainty in estimating the propensity score logistic regression for the weights. These have been explicitly derived for accurate variance estimation in the causal inference literature,[103,104,105,106] but not for MAIC. This is a priority for future research.

6.6 | Unanchored comparisons

Finally, it is worth noting that, while this article focuses on anchored indirect comparisons, most applications of population adjustment in HTA are in the unanchored setting,[25] both in published studies and in health technology appraisals. We stress that RCTs deliver the gold standard for evidence on efficacy and that unanchored comparisons make very strong assumptions which are largely considered impossible to meet (absolute effects are conditionally constant as opposed to relative effects being conditionally constant).[6,15] Unanchored comparisons effectively assume that absolute outcomes can be predicted from the covariates, which requires accounting for all variables that are prognostic of outcome.

However, the number of unanchored comparisons is likely to continue growing as regulators such as the United States Food and Drug Administration and the European Medicines Agency are, increasingly, and particularly in oncology, approving new treatments on the basis of observational or single-armed evidence, or disconnected networks with no common comparator.[107,108] As pharmaceutical companies use this type of evidence to an increasing extent to obtain accelerated or conditional regulatory approval, reimbursement agencies will, in turn, be increasingly asked to evaluate interventions where only this type of evidence is available. Therefore, further examinations of the performance of population adjustment methods must be performed in the unanchored setting.

6.7 | Areas of debate

Population-adjusted indirect comparisons make up a major area of methodological developments in evidence synthesis, with applications in HTA worldwide. We acknowledge that there is still debate in some areas, which may require further study. It is claimed that, for

the (log) hazard ratio, marginal treatment effects may vary across different distributions of the purely prognostic covariates. Hence, these covariates can still modify marginal treatment effects, even in the absence of interaction effects, and cross-trial differences in these can potentially induce bias. In our simulation study, MAIC does not account for imbalances in purely prognostic variables (these are the covariates with only main effects, not interaction effects, in the Cox model) and remains unbiased. Nevertheless, this remains a topic for further investigation.

Another area of debate is whether marginal or conditional effects are more appropriate target estimands for population-level decision-making in HTA.[90,109,110] We endorse the use of marginal effects as population-level estimates, required for reimbursement decisions at the population level.[109] Nevertheless, conditional treatment effect estimates, adjusted for prognostic factors, have been termed "population-average" effects,[90,110] and recommended on the grounds of: (1) providing more statistically precise and efficient decision-making; and (2) the clinical trials literature preferring "covariate-adjusted" over "unadjusted" analyses in order to account for the distribution of covariates.[90] We note that these conclusions are based on covariate adjustment using linear regression and continuous outcomes, and on conflating the terms "marginal" and "unadjusted". Firstly, when working with non-collapsible effect measures such as odds ratios in logistic regression with binary outcomes, or hazard ratios in Cox regression with survival outcomes, conditional covariate-adjusted treatment effect estimates actually reduce precision and efficiency with respect to unadjusted marginal estimates, in the "ideal RCT" analysis.[111,112,113,114] Secondly, it is worth noting that marginal need not mean unadjusted.[111] Marginal effects can also be covariate-adjusted and, in fact, population-adjusted indirect comparisons should ultimately produce covariate-adjusted marginal effect estimates, that account for the relevant covariate distribution and, for non-collapsible effect measures, can potentially increase precision and efficiency with respect to both conditional and unadjusted marginal effect estimates.[111,115,116,117,118]

7 | CONCLUDING REMARKS

In the performance measures we considered, MAIC was the least biased and most accurate method under no failures of assumptions. We therefore recommend its use for survival outcomes, provided that its assumptions are reasonable. MAIC was generally randomization-valid, except in situations with poor covariate overlap and small sample sizes (small effective sample sizes), where robust sandwich standard errors underestimated variability and there was undercoverage.

The typical usage of STC produced systematic bias because it targeted a conditional treatment effect for *A* versus *C*, where the target estimand should be different, a marginal treatment effect. Note that STC is not intrinsically biased; it simply targets the wrong estimand in this setting. If we intend to target a marginal treatment effect for *A* versus *C* and naively assume that this version of STC does so, there will be bias because this effect is incompatible in the indirect comparison due to the non-collapsibility of the log hazard ratio. This bias could have considerable impact on decision making and policy, and could lead to perverse decisions and subsequent misuse of resources. Therefore, the typical use of STC should be avoided, particularly in settings with a non-collapsible measure of effect. An important future objective would be the development of an alternative formulation to STC that estimates a marginal treatment effect for *A* versus *C*. A crucial additional step, missing from the current implementation, is to integrate or average the conditional effect estimates over the *BC* covariates. Then, STC could potentially obtain a marginal treatment effect estimate that is compatible with the marginal *B* versus *C* estimate published in the *BC* study.

The Bucher method is systematically biased and overprecise when there are imbalances in effect modifiers and interaction effects that induce bias in the treatment effect. Future simulation studies should evaluate population adjustment methods with different outcome types and when assumptions fail.

ACKNOWLEDGMENTS

The authors thank Anthony Hatswell for discussions that contributed to the quality of the manuscript and acknowledge Andreas Karabis for his advice and expertise in MAIC. In addition, the authors thank the editor and peer reviewers of the article. Their comments were hugely insightful and substantially improved the article, for which the authors are grateful. Finally, the authors thank Tim Morris, who provided very helpful comments after evaluating Antonio Remiro Azócar's PhD proposal defense. This article is based on research supported by Antonio Remiro-Azócar's PhD scholarship from the Engineering and Physical Sciences Research Council of the United Kingdom. G.B. is partially funded by a research grant sponsored by Mapi/ICON at University College London. A.H. was funded through an Innovative Clinical Trials Multi-year Grant from the Canadian Institutes of Health Research (funding reference number MYG-151207; 2017–2020).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The files required to generate the data, run the simulations, and reproduce the results are available at http://github.com/remiroazocar/population_adjustment_simstudy.

ORCID

Antonio Remiro-Azócar https://orcid.org/0000-0002-2877-2315

ENDNOTES

* In fact, standard ITC methods do not typically specify their target population explicitly (whether this is *AC*, *BC* or otherwise), regardless of whether the analysis is based on ALD or on IPD from each study.[29]

† If we had IPD for the *BC* study and ALD for the *AC* study, we would have to adjust for the covariates that modify the effect of treatment *B* versus *C*, in order to perform the comparison in the *AC* population.

‡ While the treatment coefficient $\hat{\beta}_T$ is an *average* treatment effect, it is not a population-level measure, contrary to the *marginal* or *population-average* treatment effect, which is the effect of moving the entire population from one treatment to the other. Firstly, while there is only *one* marginal effect for a specific population (as described by its covariate distribution), there may be many average conditional effects for a given population, one for every possible combination of covariates and model specification considered for adjustment. Secondly, conditioning on covariates changes the nature of the effect, so that it is no longer a population-level effect but an effect with a subject-level interpretation. Note that, with non-collapsible effect measures, marginal effects may not be equal to a weighted average of the conditional effects under any weighting scheme, even under no confounding bias.[61]

§ In a sufficiently large number of repetitions, $(100 \times (1-\alpha))\%$ confidence intervals based on normal distributions should contain the true value $(100 \times (1-\alpha))\%$ of the time, for a nominal significance level α.

¶ The files required to run the simulations are available at http://github.com/remiroazocar/population_adjustment_simstudy. Supplementary Appendix B of the Supporting Information lists the specific settings of each simulation scenario.

** At baseline, this formulation has a hazard function $h_0(\tau) = \lambda\nu\tau^{\nu-1}$, a cumulative hazard function $H_0(\tau) = \lambda\tau^{\nu}$, a density function $f_0(\tau) = \lambda\nu\tau^{\nu-1}\exp(-\lambda\tau^{\nu})$ and a survival function $S_0(\tau) = \exp(-\lambda\tau^{\nu})$ at time $0 \leq \tau < \infty$, where $\lambda > 0$ is a positive inverse scale (rate) parameter, and $\nu > 0$ is a positive shape parameter. This follows the proportional hazards parameterization of the Weibull distribution in NICE guidelines, where λ is referred to as a scale parameter.[70]

†† Due to the simulation study design, where the covariate distributions are symmetric, covariate *balance* is a proxy for covariate *overlap* in this parameter setting. Imbalance refers to the difference in covariate distributions across studies, as measured by the difference in (standardized) average covariate values. Overlap describes the degree of similarity in the covariate ranges across studies—there is complete overlap if the ranges are the same. In real scenarios, lack of complete overlap does not necessarily imply imbalance (and vice versa). Imbalances in effect modifiers across studies bias the standard indirect comparison, motivating the use of population adjustment. Lack of complete overlap hinders the use of population adjustment, as the covariate data may be too limited to make any conclusions in the regions of non-overlap.

‡‡ In the anchored scenario, we are interested in a comparison of *relative* outcomes or effects, not *absolute* outcomes. Hence, an anchored comparison only requires conditioning on the effect modifiers, the covariates that explain the heterogeneity of the *A* versus *C* treatment effect. This assumption is denoted the *conditional constancy of relative effects* by Phillippo et al.,[6,15] that is, given the selected effect-modifying covariates, the marginal *A* versus *C* treatment effect is constant across the *AC* and *BC* populations. There are analogous formulations of this assumption,[32,33,34,101,35] such as the conditional ignorability, unconfoundedness or exchangeability of trial assignment for such treatment effect, that is, trial selection is conditionally independent of the treatment effect, given the selected effect modifiers. One can consider that being in population *AC* or population *BC* does not carry any information about the marginal *A* versus *C* treatment effect, once we condition on the treatment effect modifiers. This means that after adjusting for these effect modifiers, treatment effect heterogeneity and trial assignment are conditionally independent.

REFERENCES

1. Sutton A, Ades A, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics*. 2008;26(9):753-767.
2. Glenny A, Altman D, Song F, et al. Indirect Compareisons of Competing Interventions; 2005.
3. Dias S, Sutton AJ, Ades A, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making*. 2013;33(5):607-617.
4. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50(6):683-691.
5. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval Health Prof*. 2002;25(1):76-97.
6. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*. 2018;38(2):200-211.
7. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46(3):399-424.
8. Faria R, Hernandez Alava M, Manca A, Wailoo A. *NICE DSU Technical Support Document 17: the Use of Observational Data to Inform Estimates of Treatment Effectiveness for Technology Appraisal: Methods for Comparative Individual Patient Data*. Sheffield: NICE Decision Support Unit; 2015.

9. Signorovitch JE, Wu EQ, Andrew PY, et al. Comparative effectiveness without head-to-head trials. *Pharmacoeconomics*. 2010;28(10):935-945.
10. Signorovitch J, Erder MH, Xie J, et al. Comparative effectiveness research using matching-adjusted indirect comparison: an application to treatment with guanfacine extended release or atomoxetine in children with attention-deficit/hyperactivity disorder and comorbid oppositional defiant disorder. *Pharmacoepidemiol Drug Safe*. 2012;21:130-137.
11. Signorovitch JE, Sikirica V, Erder MH, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value Health*. 2012;15(6):940-947.
12. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82(398):387-394.
13. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics*. 2010;28(10):957-967.
14. Zhang Z. Covariate-adjusted putative placebo analysis in active-controlled clinical trials. *Stat Biopharmaceut Res*. 2009;1(3):279-290.
15. Phillippo D, Ades T, Dias S, Palmer S, Abrams KR, Welton N. NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submissions to NICE; 2016.
16. Ishak KJ, Proskorovsky I, Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics*. 2015;33(6):537-549.
17. Stevens JW, Fletcher C, Downey G, Sutton A. A review of methods for comparing treatments evaluated in studies that form disconnected networks of evidence. *Res Synth Methods*. 2018;9(2):148-162.
18. Thom H, Jugl S, Palaka E, Jawla S. Matching adjusted indirect comparisons to assess comparative effectiveness of therapies: usage in scientific literature and health technology appraisals. *Value Health*. 2016;19(3):A100-A101.
19. Kühnast S, Schiffner-Rohe J, Rahnenführer J, Leverkus F. Evaluation of adjusted and unadjusted indirect comparison methods in benefit assessment. *Methods Inf Med*. 2017;56(03):261-267.
20. Petto H, Kadziola Z, Brnabic A, Saure D, Belger M. Alternative weighting approaches for anchored matching-adjusted indirect comparisons via a common comparator. *Value Health*. 2019;22(1):85-91.
21. Cheng D, Ayyagari R, Signorovitch J. The statistical performance of matching-adjusted indirect comparisons. arXiv preprint arXiv:1910.06449; 2019.
22. Hatswell AJ, Freemantle N, Baio G. The effects of model misspecification in unanchored matching-adjusted indirect comparison (MAIC): results of a simulation study. *Value Health*. 2020;23:751-759.
23. Belger M, Brnabic A, Kadziola Z, Petto H, Faries D. Inclusion of multiple studies in matching adjusted indirect comparisons (MAIC). *Value Health*. 2015;18(3):A33.
24. Leahy J, Walsh C. Assessing the impact of a matching-adjusted indirect comparison in a Bayesian network meta-analysis. *Res Synth Methods*. 2019;10:546-568.
25. Phillippo DM, Dias S, Elsada A, Ades A, Welton NJ. Population adjustment methods for indirect comparisons: a review of National Institute for Health and Care Excellence Technology Appraisals. *Int J Technol Assess Health Care*. 2019;35:221-228.
26. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21.
27. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One*. 2011;6(3):e18174.
28. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcome Res Methodol*. 2001;2(3–4):259-278.
29. Manski CF. Meta-Analysis for Medical Decisions; 2019.
30. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*. 2003;326(7387):472.
31. Baio G, Dawid AP. Probabilistic sensitivity analysis in health economics. *Stat Methods Med Res*. 2015;24(6):615-634.
32. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc A Stat Soc*. 2011;174(2):369-386.
33. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am J Epidemiol*. 2010;172(1):107-115.
34. Kern HL, Stuart EA, Hill J, Green DP. Assessing methods for generalizing experimental impact estimates to target populations. *J Res Educ Effect*. 2016;9(1):103-127.
35. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *J R Stat Soc A Stat Soc*. 2015;178(3):757-778.
36. Veroniki AA, Straus SE, Soobiah C, Elliott MJ, Tricco AC. A scoping review of indirect comparison methods and applications using individual patient data. *BMC Med Res Methodol*. 2016;16(1):47.
37. Ndirangu K, Tongbram V, Shah D. Trends in the use of matching-adjusted indirect comparisons in published literature and Nice technology assessments: a systematic review. *Value Health*. 2016;19(3):A99-A100.
38. Nocedal J, Wright S. *Numerical optimization*. New York: Springer Science & Business Media; 2006.
39. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48:817-838.
40. Windmeijer F. A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*. 2005;126(1):25-51.
41. Hainmueller J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Polit Anal*. 2012;20(1):25-46.
42. Phillippo DM, Dias S, Ades A, Welton NJ. Equivalence of entropy balancing and the method of moments for matching-adjusted indirect comparison. *Res Synt Meth*. 2020;11(4):568-572.
43. Efron B. *Bootstrap Methods: another Look at the Jackknife*. New York: Springer; 1992:569-593.
44. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, United States: CRC Press; 1994.
45. Sikirica V, Findling RL, Signorovitch J, et al. Comparative efficacy of guanfacine extended release versus atomoxetine for the treatment of attention-deficit/hyperactivity disorder in children and adolescents: applying matching-adjusted indirect comparison methodology. *CNS Drugs*. 2013;27(11):943-953.

46. Steyerberg EW. *Clinical Prediction Models*. Springer; 2019. New York.
47. Harrell FE, Slaughter JC. Biostatistics for Biomedical Research; 2016.
48. Senn S, Graf E, Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Stat Med*. 2007;26(30): 5529-5544.
49. Ishak K, Rael M, Phatak H, Masseria C, Lanitis T. Simulated treatment comparison of time-to-event (and other non-linear) outcomes. *Value Health*. 2015;18(7):A719.
50. Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models: review and new applications. *Am Stat*. 2004;58(4):272-279.
51. Rosenbaum P, Colton T, Armitage P. Encyclopedia of Biostatistics; 1998.
52. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32(16):2837-2849.
53. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med*. 2010;29(20):2137-2148.
54. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *J Clin Epidemiol*. 2010;63(2):142-153.
55. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev*. 1991;59(1):25-35.
56. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125(5):761-768.
57. Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*. 2010;21:467-474.
58. Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Commun Health*. 2004;58(4):265-271.
59. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med*. 2014; 33(7):1242-1258.
60. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86(1):4-29.
61. Huitfeldt A, Stensrud MJ, Suzuki E. On the collapsibility of measures of effect in the counterfactual causal framework. *Emerg Themes Epidemiol*. 2019;16(1):1-5.
62. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci*. 1999;14(1):29-46.
63. Janes H, Dominici F, Zeger S. On quantifying the magnitude of confounding. *Biostatistics*. 2010;11(3):572-582.
64. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in cox and Aalen regression models. *Lifetime Data Anal*. 2013;19(3):279-296.
65. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*. 1984;71(3):431-444.
66. Miettinen OS, Cook EF. Confounding: essence and detection. *Am J Epidemiol*. 1981;114(4):593-603.
67. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102.
68. R Core Team. R: A language and environment for statistical computing; 2013.
69. Bender R, Augustin T, Blettner M. Generating survival times to simulate cox proportional hazards models. *Stat Med*. 2005; 24(11):1713-1723.
70. Latimer NR. Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Med Decis Making*. 2013;33(6):743-754.
71. Nelsen RB. *An Introduction to Copulas*. Springer Science & Business Media; 2007. New York.
72. Abrahamowicz M, Berger DR, Krewski D, et al. Bias due to aggregation of individual covariates in the cox regression model. *Am J Epidemiol*. 2004;160(7):696-706.
73. Brent RP. An algorithm with guaranteed convergence for finding a zero of a function. *Comput J*. 1971;14(4):422-425.
74. Stanley K. Design of randomized controlled trials. *Circulation*. 2007;115(9):1164-1169.
75. Therneau TM, Grambsch PM. *The Cox Model*. Springer; 2000: 39-77. New York.
76. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J Am Stat Assoc*. 1999;94(448):1053-1062.
77. Waernbaum I. Propensity score model specification for estimation of average treatment effects. *J Stat Plan Infer*. 2010; 140(7):1948-1956.
78. Zhao Z. Sensitivity of propensity score methods to the specifications. *Econ Lett*. 2008;98(3):309-319.
79. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc*. 2000;95(450):573-585.
80. Leyrat C, Caille A, Donner A, Giraudeau B. Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. *Stat Med*. 2014;33(20):3556-3575.
81. Rücker G, Schwarzer G. Presenting simulation results in a nested loop plot. *BMC Med Res Methodol*. 2014;14(1):129.
82. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147-177.
83. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25(24): 4279-4292.
84. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc*. 2001;96 (456):1387-1396.
85. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*. 2001;57(4):1198-1206.
86. Claxton K, Sculpher M, McCabe C, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health Econ*. 2005;14(4):339-347.
87. Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J Roy Stat Soc*. 1934;97(4):558-625.
88. Phillippo DM, Dias S, Ades A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. *J Roy Stat Soc*. 2020;183(3):1189-1210.
89. Phillippo DM. Calibration of treatment effects in network meta-analysis using individual patient data. PhD thesis. University of Bristol, Bristol, UK; 2019.

90. Phillippo DM, Dias S, Ades A, Welton NJ. Target estimands for efficient decision making: response to comments on "assessing the performance of population adjustment methods for anchored indirect comparisons: a simulation study". *Stat Med*. 2021;40:2759-2763.
91. Ricciardi F, Liverani S, Baio G. Dirichlet process mixture models for regression discontinuity designs. arXiv preprint arXiv:2003.11862; 2020.
92. Fisher D, Copas A, Tierney J, Parmar M. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *J Clin Epidemiol*. 2011;64(9):949-967.
93. Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach. *BMJ*. 2017;356:j573.
94. Tierney JF, Vale C, Riley R, et al. Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. *PLoS Med*. 2015;12(7):e1001855.
95. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. *Network Meta-Analysis for Decision-Making*. John Wiley & Sons; 2018. Chichester, United Kingdom.
96. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to Meta-Analysis*. John Wiley & Sons; 2011. Chichester, United Kingdom.
97. Dias S, Sutton AJ, Welton NJ, Ades A. *NICE DSU Technical Support Document 3: Heterogeneity: Subgroups, Meta-Regression, Bias and Bias-Adjustment*; 2011. National Institute for Health and Care Excellence. United Kingdom.
98. Remiro-Azócar A, Heath A, Baio G. Principled selection of effect modifiers: comments on 'Matching-adjusted indirect comparisons: application to time-to-event data'. arXiv preprint arXiv:2012.05127; 2020.
99. Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Stat Med*. 1989;8(12):1515-1521.
100. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*. 2007;165(6):710-718.
101. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. *Stat Sci*. 2014;29(4):579-595.
102. Zeileis A. Econometric Computing with HC and HAC Covariance Matrix Estimators; 2004.
103. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937-2960.
104. Buchanan AL, Hudgens MG, Cole SR, et al. Generalizing evidence from randomized trials using inverse probability of sampling weights. *J Roy Stat Soc*. 2018;181:1193.
105. Li F. Propensity score weighting for causal inference with multiple treatments. *Ann Appl Stat*. 2019;13(4):2389-2415.
106. Mao H, Li L, Greene T. Propensity score weighting analysis and treatment effect discovery. *Stat Methods Med Res*. 2019;28(8):2439-2454.
107. Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. *BMJ Open*. 2016;6(6):1-8.
108. Beaver JA, Howie LJ, Pelosof L, et al. A 25-year experience of US Food and Drug Administration accelerated approval of malignant hematology and oncology drugs and biologics: a review. *JAMA Oncol*. 2018;4(6):849-856.
109. Remiro-Azócar A, Heath A, Baio G. Conflating marginal and conditional treatment effects: comments on 'Assessing the performance of population adjustment methods for anchored indirect comparisons: a simulation study'. arXiv Preprint arXiv:2011.06334; 2020.
110. Phillippo DM, Dias S, Ades A, Welton NJ. Assessing the performance of population adjustment methods for anchored indirect comparisons: a simulation study. *Stat Med*. 2020;39:4885-4911.
111. Daniel R, Zhang J, Farewell D. Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom J*. 2020;63:528-557.
112. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev*. 1991;58:227-240.
113. Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. *Stat Med*. 1995;14(8):735-746.
114. Hernández AV, Steyerberg EW, Habbema JDF. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol*. 2004;57(5):454-460.
115. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Stat Med*. 2008;27(23):4658-4677.
116. Benkeser D, Díaz I, Luedtke A, Segal J, Scharfstein D, Rosenblum M. Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*. 2020.
117. Colantuoni E, Rosenblum M. Leveraging prognostic baseline variables to gain precision in randomized trials. *Stat Med*. 2015;34(18):2602-2617.
118. Díaz I, Colantuoni E, Rosenblum M. Enhanced precision in the analysis of randomized trials with ordinal outcomes. *Biometrics*. 2016;72(2):422-431.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Remiro-Azócar A, Heath A, Baio G. Methods for population adjustment with limited access to individual patient data: A review and simulation study. *Res Syn Meth*. 2021;12(6):750-775. https://doi.org/10.1002/jrsm.1511