ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Methodology

The Effects of Model Misspecification in Unanchored Matching-Adjusted Indirect Comparison: Results of a Simulation Study

Anthony James Hatswell, MSc,[1,2,*] Nick Freemantle, PhD,[3] Gianluca Baio, PhD[1]

[1]Department of Statistical Science, University College London, London, England, UK; [2]Delta Hat, Nottingham, England, UK; [3]Institute of Clinical Trials and Methodology, University College London, London, UK.

ABSTRACT

Objectives: To assess the performance of unanchored matching-adjusted indirect comparison (MAIC) by matching on first moments or higher moments in a cross-study comparisons under a variety of conditions. A secondary objective was to gauge the performance of the method relative to propensity score weighting (PSW).

Methods: A simulation study was designed based on an oncology example, where MAIC was used to account for differences between a contemporary trial in which patients had more favorable characteristics and a historical control. A variety of scenarios were then tested varying the setup of the simulation study, including violating the implicit or explicit assumptions of MAIC.

Results: Under ideal conditions and under a variety of scenarios, MAIC performed well (shown by a low mean absolute error [MAE]) and was unbiased (shown by a mean error [ME] of about zero). The performance of the method deteriorated where the matched characteristics had low explanatory power or there was poor overlap between studies. Only when important characteristics are not included in the matching did the method become biased (nonzero ME). Where the method showed poor performance, this was exaggerated if matching was also performed on the variance (ie, higher moments). Relative to PSW, MAIC provided similar results in most circumstances, although it exhibited slightly higher MAE and a higher chance of exaggerating bias.

Conclusions: MAIC appears well suited to adjust for cross-trial comparisons provided the assumptions underpinning the model are met, with relatively little efficiency loss compared with PSW.

Keywords: historical control, MAIC, propensity score, Signorovitch weighting, single-arm trial.

VALUE HEALTH. 2020; 23(6):751–759

Introduction

For the assessment of comparative efficacy, interventions are ideally studied in a head-to-head clinical trial. Where such trial evidence is not available, techniques such as indirect comparisons[1] or network meta-analysis[2] can provide estimates of the relative efficacy of interventions. However, where trials either have substantially different comparator arms, no available link to connect them (ie, a disconnected network), or lack control arms entirely, options are more limited; meta-regression requires a large number of studies to recover information on differences in patient characteristics between trials, whereas propensity score techniques (ie, propensity score matching or propensity score weighting [PSW]) require access to patient-level data,[3] which are frequently not available for at least 1 of the relevant comparators.

This lack of access to patient-level data for comparator trials is a limiting factor in many health technology appraisal submissions. The reasons for the lack of access can be complex but often involve the data being owned by a competitor manufacturer, confidentiality issues, or the data being inaccessible because of the passage of time. Where this patient-level data are not available, matching adjusted indirect comparison (MAIC) has been proposed.[4] Analogous to PSW, MAIC involves weighting the patient-level data available (usually from a manufacturer's own trial) to match the aggregate characteristics of the target trial for which individual patient data are unavailable.

The result of the MAIC weighting can be used in 2 approaches: first, to account for differences between trials in predictive characteristics, with the results subsequently used to inform network meta-analysis (this is often termed an *anchored comparison*).

Conflict of interest: No funding was received for this work, and the authors declare no conflicts.

* Address correspondence to: Anthony J. Hatswell, MSc, Delta Hat, 212 Tamworth Road, Nottingham, NG10 3GS England, United Kingdom. Email: ahatswell@deltahat.co.uk

https://doi.org/10.1016/j.jval.2020.02.008

Alternatively, MAIC is used to weight one study to match the population of a second trial, allowing for cross-trial comparisons (this is termed an *unanchored* comparison). Each of these approaches aims to minimize bias in estimates of comparative efficacy, for example, accounting for one population being younger or having better performance status.

Although the promise of MAIC is considerable, the method is relatively new, having been first described in 2010. As of December 2019, there were 126 publications listed on PubMed addressing the approach (including the original conceptual papers), with little consistency in the application of the technique. The only formal guidance available is a National Institute for Health and Care Excellence Decision Support Unit Technical Support Document report and associated publication,[5,6] which defines terminology, reviews the theoretical validity of the method, and suggests best practice (ie, in unanchored MAIC, include all prognostic and predictive characteristics). This work, however, does not provide guidance on how the method should be applied in specific circumstances and indeed highlights the need for simulation studies to understand the properties of the method. Although 2 simulation studies have been published, these focus on the use of MAIC in anchored indirect comparisons as opposed to the unanchored form.[7,8]

As most published MAIC applications consider unanchored comparisons (as an alternative to a naïve comparison), we examined the performance of the method under such conditions. To do so, we conducted a simulation study to understand the performance of the method under different data structures and assumptions, with a secondary objective of comparing the efficiency of alternative matching approaches: unanchored MAIC matching on first moments ($MAIC_{FM}$), MAIC matching on first and second moments (ie, matching on the mean and variance; $MAIC_{HM}$), and PSW. PSW is included because although it is not a direct comparator to MAIC (it requires patient data for both trials), it represents the most widely respected approach to weighting. Thus, it is therefore possible to gauge the loss of efficiency by only being able to match to the moments of the data using unanchored MAIC as opposed to matching using patient data from both studies (as in PSW).

Methods

Aims and Design

Our review of published MAICs found that most published applications are in oncology (23 applications), compared with 30 in all other diseases combined; a further 5 papers discussed the method without a specific example. For this reason, we based our simulation exercise on time-to-event data and comparing an intervention to a historical control.[9] In keeping with the literature on historical comparisons, the individuals in the target population for the contemporary trial of the intervention (termed *population A*) were assumed to have more favorable characteristics than the patients who received the historical control (termed *population B*), leading to a bias in favor of the intervention.[10] A simulation study was therefore programmed to mimic such circumstances to understand the effectiveness of MAIC in removing the bias in such naïve comparisons. The study was designed using guidance on best practice for simulation studies in medical statistics.[11,12]

Data-Generating Mechanism

In the study, 6 patient characteristics (X_1-X_6) were simulated; 4 were assumed to be fully observed and available for matching (X_1-X_4), whereas 2 (X_5, X_6) were assumed to be unobserved. In the base case, these were all assumed to be uncorrelated. The 4 observed covariates were simulated from the same distributional form (in the base case, a normal distribution), providing a bias of half a standard deviation for each characteristic in favor of the intervention. Unobserved characteristics were drawn from the same distributions for both populations, implicitly assuming they do not bias the comparison, although these do add variability in outcomes (as is seen in reality). Four characteristics were selected for matching, as this is in line with analysis of cancer data identifying prognostic cancers such as in bladder cancer with 3 prognostic characteristics,[13] and in line with previous MAICs in which Phillippo et al[14] found that a median of 6 characteristics were adjusted for (range, 1-13). Each characteristic was then multiplied by an effect size for that characteristic (β_1- β_6). The sum of these products was added to a constant (intercept) and then used as a linear predictor in a Weibull proportional hazards survival model with a corresponding survival time (Y) sampled for each patient both with and without receiving the intervention. In other words, for each patient, the linear predictor $LP = \sum_{i=1}^{6} \beta_i X_i + c$ with survival outcomes for each patient then sampled from the corresponding distribution $Weibull(\alpha = 1.3, \exp(\mathrm{LP}))$. In each run of the simulation study, patient characteristics (X_1-X_6) were resampled, as were different effect sizes (β_1- β_6, α) as shown graphically in Figure 1. By allowing these parameters to vary, we are able to ensure the result holds for the distribution in general, as opposed to testing only a specific distribution.

As the objective of the study was to understand the performance of MAIC under different assumptions, a large number (n = 1000) of patients was simulated for population A (treated), and for population B (historical control), survival times were assumed to be observed until death, and no data were assumed to be missing. This simulation setup (a large number of patients with fully observed survival times and no missing data) was chosen to ensure the study assessed the matching methods and not the variability of outcomes in individual patients or the approach to extrapolation or missing data (as would have been the case had censoring been assumed). In practice, such data are unlikely to be fully observed or available for all studies, although methods do exist for digitization of survival outcomes,[15] estimation of missing

Figure 1. Data generation mechanism for the simulation study.

Survival ~ Weibull (shape = α, scale = λ)

$\alpha = 1.3$

$\lambda = exp\dfrac{(2 + \sum(\beta_1 X_1 \ldots \beta_6 X_6))}{Treatment\ Hazard\ Ratio}$

$\beta_1 \ldots \beta_4 \sim Normal(mean = 0.5, standard\ deviation = 0.2)$

$\beta_5 \ldots \beta_6 \sim Normal(mean = 0, standard\ deviation = 0.2)$

For the intervention:

$X_1 \ldots X_4 \sim Normal(mean = 0.3, standard\ deviation = 0.1)$

$Treatment\ Hazard\ Ratio = 0.75$

For the control:

$X_1 \ldots X_4 \sim Normal(mean = 0.25, standard\ deviation = 0.1)$

$Treatment\ Hazard\ Ratio = 1$

For intervention & control:

$X_5 \ldots X_6 \sim Normal(mean = 0.25, standard\ deviation = 0.1)$

Table 1. Parameters used in the base case and changed to form each scenario analysis.*

Scenario	Base case	Scenario setting
Changing the setup of the simulation study		
All variables are binary	Covariates 1 to 4: Population A: $X_A \sim N(0.3, 0.1)$ Population B: $X_B \sim N(0.25, 0.1)$	Covariates 1 to 4: Population A: $X_A \sim$ Binomial(probability = 0.3) Population B: $X_B \sim$ Binomial(probability = 0.25)
Exponential distribution used as the survival function	Survival: $Y \sim$ Weibull(shape = 1.3, scale = exp(2+∑Xβ)/ TreatmentHR))	Survival: $Y \sim$ Weibull(shape = 1, scale = exp(2+∑Xβ))
Explanatory variable power is low	Covariates 1:4: $\beta \sim N(0.5, 0.2)$	Covariates 1:4: $\beta \sim N(0.1, 0.05)$
Explanatory variable power is high		Covariates 1:4: $\beta \sim N(1, 0.4)$
Treatment effect is low	TreatmentHR = hazard ratio of 0.75	TreatmentHR = hazard ratio of 0.9
Treatment effect is high		TreatmentHR = hazard ratio of 0.2
Covariate sampling is reversed, ie, population A less favorable	Covariates 1 to 4: Population A: $X_A \sim N(0.3, 0.1)$	Covariates 1 to 4: Population A: $X_A \sim N(0.2, 0.1)$
Exploring the limits of MAIC		
Half the matched parameters are nuisance parameters	Covariates 1:4: $\beta \sim N(0.5, 0.2)$	Covariates 1:2: $\beta \sim N(1.0, 0.2)$ Covariates 3:6: $\beta \sim N(0.0, 0.2)$
All the matched parameters are nuisance parameters		Covariates 1:4: $\beta \sim N(0, 0.2)$
The effect of parameters is nonlinear	Survival: $Y \sim$ Weibull(shape = 1.3, scale = exp(2+∑Xβ)/ TreatmentHR))	Survival: $Y \sim$ Weibull(shape = 1.3, scale = exp(2+∑exp(Xβ))/ TreatmentHR))
Small difference is covariate sampling (0.1SD)	Covariates 1 to 4: Population A: $X_A \sim N(0.3, 0.1)$	Covariates 1 to 4: Population A: $X_A \sim N(0.26, 0.1)$
Large difference in covariate sampling (1SD)		Covariates 1 to 4: Population A: $X_A \sim N(0.35, 0.1)$
Parameters correlated		Underlying health: Population A: $H_A \sim N(0.3, 0.1)$ Population B: $H_B \sim N(0.25, 0.1)$ Covariates 1:4: $X \sim$ N(0, 0.1) + H
Violating assumptions implicit or explicit in MAIC		
Missing parameters correlated with observed parameters	Covariates 5 and 6: $X \sim N(0, 0.2)$	Covariates 5 and 6: $X \sim$ mean of parameters 1:4 + $N(0, 0.1)$
Missing parameters uncorrelated with observed parameters		Covariates 5 and 6: Population A: $X_A \sim N(0.3, 0.1)$ Population B: $X_B \sim N(0.25, 0.1)$ Covariates 1:6: $\beta \sim N(0.35, 0.15)$
Non-normal distributions sampled in population A	Covariates 1 to 4: Population A: $X_A \sim N(0.3, 0.1)$ Population B: $X_B \sim N(0.25, 0.1)$	Covariates 1 to 4: Population A: $X_A \sim$ *Lognormal*(SDlog = 0.5, meanlog = log(0.27))
Non-normal distributions sampled in population B		Covariates 1 to 4: Population A: $X_B \sim$ *Lognormal*(SDlog = 0.5, meanlog = log(0.22))
Trimmed patient characteristics in population A (no poor performers)		Covariates: Population A: $X_A \sim N(0.3, 0.1)$ truncated at min of 0.2
Trimmed patient characteristics in population B (no good performers)		Covariates: Population B: $X_B \sim N(0.25, 0.1)$ truncated at max of 0.35

HR indicates hazard ratio.
*Distribution parameterizations are given as in the statistical package R to allow easy reproducibility; thus, normal distributions are given as normal ~ (mean, standard deviation) and not normal (mean, variance), and the Weibull is specified using the shape (and not rate) parameter.

data,[16,17] and extrapolation of survival times,[18] which can be implemented alongside matching procedures.

Methods Under Investigation

A naïve comparison contrasts the observed outcome in population A of the intervention (Y_{A_INT}) with the outcomes seen in population B of the historical control (Y_{B_HC}). This comparison is subject to bias caused by the more favorable characteristics in population A. Matching methods (both MAIC and PSW) therefore attempt to reweight Y_{A_INT} with the aim of estimating the effect of the intervention in population B (what would be Y_{B_INT}). This can then be compared with the observed historical control outcome (Y_{B_HC}) for a fair comparison; the result that would have been obtained had a controlled study of A versus B been performed in population B. Because it is a simulation study, the data-generation

mechanisms are known, and thus outcomes can be computed with and without the intervention for both groups. By comparing the estimated effect to the (unobserved) true effect, the success of both MAIC and PSW in estimating this true effect can be assessed.

Reweighting was then conducted by matching the observed patient characteristics in population A to the characteristics in population B. This was done using 3 approaches: first, using MAIC matching on the means (first moments) of population B, $MAIC_{FM}$, matching was then conducted using also the standard deviation of the summarized data from population B (ie, matching also on higher moments), $MAIC_{HM}$; this approach was also proposed in the original MAIC paper by Signorovitch et al.[4] "For example, given the baseline mean and standard deviation of age, it is straightforward to compute the mean of squared age, which can then be treated as a separate mean baseline characteristic for matching." Finally, PSW was conducted, in which weights were calculated for all patients (assuming access to individual patient data for both trials). This allowed us to assess the impact of not having access to the individual patient data from the historical control by comparing MAIC methods to the gold standard of PSW. Each set of weights ($MAIC_{FM}$, $MAIC_{HM}$, and PSW) was then used to estimate outcomes on a per-simulation basis.

Outcomes of the Study

To ascertain the effectiveness of matching methods, the Cox proportional hazard was estimated for each method used: a naïve comparison of Y_{B_HC} and Y_{A_INT}, as well as between the reweighted value of Y_{A_INT} seen with MAIC, and PSW. Using this point estimate of the hazard ratio, 3 outcomes were calculated: the mean percentage error (which overall should be zero for an unbiased method), the mean absolute percentage error (a lower value leading to more accurate predictions; both over- and underpredictions are penalized equally), and the coverage probability (whether the 95% interval for each estimated hazard ratio contained the "true" value). In addition, for the weighting methods, whether the point estimate of the hazard was more accurate than the corresponding naïve comparison was calculated; this was used to determine how often a method would be more likely to introduce bias than to remove it.

Scenario Analyses

Scenario analysis was then conducted with 3 broad aims: varying the characteristics of the simulation study, testing the limits of where MAIC can be applied, and violating the assumptions implicit or explicit in the approach.

In changing the setup of the simulation study, several factors were considered, including the survival model used, type of variables used in matching (binary as opposed to continuous), relative importance of covariates, and efficacy of treatment. These were extended in testing the limits of MAIC by matching also on nuisance parameters (ie, variables that were not linked to outcomes or variables were linked in a nonlinear fashion). Further scenarios considered the degree of overlap between the 2 studies, and correlation between parameters.

When considering the violation of assumptions in MAIC, the simulation was altered to test the effects of the unobserved parameters (X_5, X_6) also being important and either correlated or uncorrelated with X_1 through X_4. The effect of outcome distributions of X_1 through X_4 was then explored by deviating from the initial assumption of normality (the lognormal was used), with also trimmed distributions of X used (mimicking trials that have inclusion and exclusion criteria, such as age limits).

A final set of sensitivity analyses involved varying the number of patients available for matching with the base-case settings. In these scenarios, the number of patients available in population A and population B was varied individually and jointly to include n = 30, n = 300, and n = 3000 patients. The aim of these scenarios was to understand the relative importance of the number of patients in each trial and how the level of error was affected by the number of patients in each study. Technical details of all scenario analyses conducted are presented in Table 1.

Implementation

The simulation study was programmed in the statistical software R version 3.6.1,[19] with survival curves simulated using the stats package, and Cox proportional hazards and robust standard errors (using the "sandwich" method) calculated using the survival package. Monte Carlo standard errors were calculated using the mcmcse package. Plots were created using *ggplot* and ggsurvplot. Truncated distributions were sampled using the MSM package. To account for Monte Carlo error, 5000 iterations of each scenario were performed.

Results

Figure 2 shows the modeled survival for 1 iteration of the simulation study, with a naïve comparison of population B historical control data (blue line, with median survival of 9.2 months, mean survival of 11.4 months over all simulations) with the data on the intervention from population A (black line, median = 13.4, mean = 16.8 months). However, had patients had the same distribution of covariates (ie, not had more favorable characteristics), the outcomes that would have been seen are those seen by the green line (median = 12.3, mean = 15.2 months). This bias in survival curves (median bias of 1.1 months and 1.6 months) due to more favorable patient characteristics leads to an underestimate

Figure 2. Example of simulated survival in the base-case analysis showing the longer survival of population A compared with population B with treatment (owing to more favorable patient characteristics) and the resulting bias in a naïve comparison to a historical population B cohort.

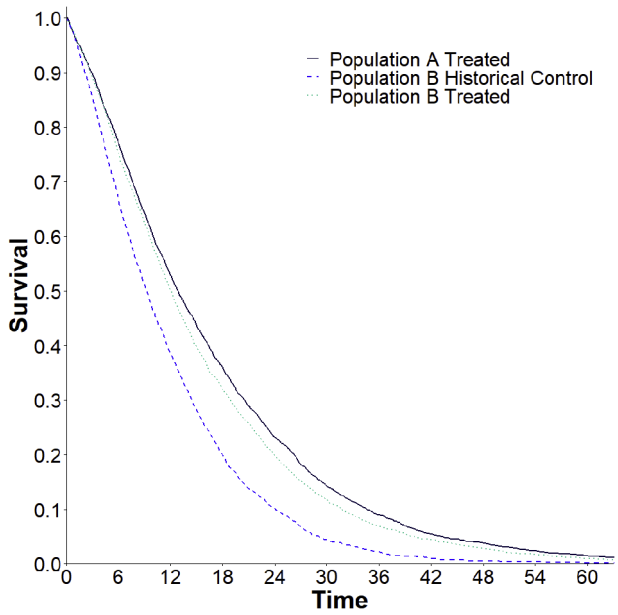


Table 2. Tabulated results of the base-case and scenario analyses.

Method	Mean percentage error (MCSE)	Absolute percentage error (MCSE)	Mean standard error	Coverage probability	Percentage of scenarios worse than a naïve comparison
Base case					
Naïve comparison	11.8% (<0.01)	11.8% (<0.01)	0.03	0%	—
$MAIC_{MM}$	−0.2% (<0.01)	2.6% (<0.01)	0.03	95%	2%
$MAIC_{HM}$	−0.2% (<0.01)	2.6% (<0.01)	0.03	95%	2%
PSW	−0.1% (<0.01)	2.7% (<0.01)	0.03	95%	2%
All variables are binary					
Naïve comparison	5.2% (<0.01)	5.2% (<0.01)	0.03	48%	—
$MAIC_{MM}$	0% (<0.01)	1.8% (<0.01)	0.03	98%	12%
$MAIC_{HM}$	4.1% (<0.01)	4.2% (<0.01)	0.03	63%	4%
PSW	0% (<0.01)	1.8% (<0.01)	0.03	98%	12%
Exponential distribution used as the survival function					
Naïve comparison	9.4% (<0.01)	9.4% (<0.01)	0.03	3%	—
$MAIC_{MM}$	−0.1% (<0.01)	2.6% (<0.01)	0.03	95%	4%
$MAIC_{HM}$	−0.1% (<0.01)	2.6% (<0.01)	0.03	95%	4%
PSW	−0.1% (<0.01)	2.7% (<0.01)	0.03	95%	4%
Lognormal used as the survival function					
Naïve comparison	7.5% (<0.01)	7.5% (<0.01)	0.04	55%	—
$MAIC_{MM}$	−6.3% (<0.01)	6.4% (<0.01)	0.05	81%	42%
$MAIC_{HM}$	−6.3% (<0.01)	6.4% (<0.01)	0.05	81%	42%
PSW	0% (<0.01)	2.9% (<0.01)	0.05	99%	10%
Explanatory variable power is low					
Naïve comparison	2.5% (<0.01)	2.9% (<0.01)	0.03	84%	—
$MAIC_{MM}$	−0.1% (<0.01)	2.8% (<0.01)	0.04	95%	43%
$MAIC_{HM}$	−0.1% (<0.01)	2.8% (<0.01)	0.04	95%	43%
PSW	−0.1% (<0.01)	2.8% (<0.01)	0.04	95%	44%
Explanatory variable power is high					
Naïve comparison	21.7% (<0.01)	21.7% (<0.01)	0.03	0%	—
$MAIC_{MM}$	−0.9% (<0.01)	7.1% (<0.01)	0.08	94%	2%
$MAIC_{HM}$	−1% (<0.01)	7.1% (<0.01)	0.08	94%	2%
PSW	0.2% (<0.01)	7.7% (<0.01)	0.09	94%	4%
Treatment effect is low (0.9 hazard ratio)					
Naïve comparison	11.9% (<0.01)	11.9% (<0.01)	0.03	0%	—
$MAIC_{MM}$	0% (<0.01)	2.5% (<0.01)	0.03	95%	1%
$MAIC_{HM}$	0% (<0.01)	2.5% (<0.01)	0.03	95%	1%
PSW	0% (<0.01)	2.6% (<0.01)	0.03	95%	1%
Treatment effect is high (0.2 hazard ratio)					
Naïve comparison	11.7% (<0.01)	11.7% (<0.01)	0.04	10%	—
$MAIC_{MM}$	−0.8% (<0.01)	4.3% (<0.01)	0.05	94%	10%
$MAIC_{HM}$	−0.8% (<0.01)	4.3% (<0.01)	0.05	94%	10%
PSW	−0.1% (<0.01)	4.4% (<0.01)	0.05	94%	9%
Covariate sampling is reversed, ie, population A is worse by 0.5SD					
Naïve comparison	−13.6% (<0.01)	13.6% (<0.01)	0.03	0%	—
$MAIC_{MM}$	v0.2% (<0.01)	3% (<0.01)	0.04	95%	1%
$MAIC_{HM}$	−0.2% (<0.01)	3% (<0.01)	0.04	95%	1%
PSW	−0.1% (<0.01)	3% (<0.01)	0.04	95%	1%
Half of the matched parameters are nuisance parameters					
Naïve comparison	11.7% (<0.01)	11.7% (<0.01)	0.03	0%	—
$MAIC_{MM}$	−0.1% (<0.01)	2.6% (<0.01)	0.03	96%	2%
$MAIC_{HM}$	−0.1% (<0.01)	2.6% (<0.01)	0.03	96%	2%
PSW	0% (<0.01)	2.6% (<0.01)	0.03	95%	2%
All the matched parameters are nuisance parameters					
Naïve comparison	0% (<0.01)	2.1% (<0.01)	0.03	95%	—
$MAIC_{MM}$	0% (<0.01)	2.9% (<0.01)	0.04	95%	64%
$MAIC_{HM}$	0% (<0.01)	2.9% (<0.01)	0.04	95%	64%
PSW	0% (<0.01)	2.9% (<0.01)	0.04	95%	64%
The effect of parameters is nonlinear					
Naïve comparison	11.9% (<0.01)	11.9% (<0.01)	0.03	0%	—
$MAIC_{MM}$	−0.1% (<0.01)	2.6% (<0.01)	0.03	96%	1%
$MAIC_{HM}$	−0.1% (<0.01)	2.6% (<0.01)	0.03	96%	1%
PSW	0% (<0.01)	2.6% (<0.01)	0.03	96%	1%

continued on next page

Table 2. Continued

Method	Mean percentage error (MCSE)	Absolute percentage error (MCSE)	Mean standard error	Coverage probability	Percentage of scenarios worse than a naïve comparison
Small difference is covariate sampling (0.1SD)					
Naïve comparison	2.5% (<0.01)	3% (<0.01)	0.03	84%	—
$MAIC_{MM}$	0% (<0.01)	2.1% (<0.01)	0.03	95%	31%
$MAIC_{HM}$	0% (<0.01)	2.1% (<0.01)	0.03	95%	31%
PSW	0% (<0.01)	2.1% (<0.01)	0.03	95%	31%
Large difference in covariate sampling (1SD)					
Naïve comparison	22.4% (<0.01)	22.4% (<0.01)	0.03	0%	—
$MAIC_{MM}$	−0.7% (<0.01)	6.9% (<0.01)	0.08	95%	2%
$MAIC_{HM}$	−0.7% (<0.01)	6.9% (<0.01)	0.08	95%	2%
PSW	0.2% (<0.01)	7.7% (<0.01)	0.09	94%	4%
All parameters correlated					
Naïve comparison	10.7% (<0.01)	10.7% (<0.01)	0.03	1%	—
$MAIC_{MM}$	−0.1% (<0.01)	2% (<0.01)	0.03	96%	1%
$MAIC_{HM}$	−0.1% (<0.01)	2% (<0.01)	0.03	96%	1%
PSW	0.1% (<0.01)	2% (<0.01)	0.03	96%	1%
Missing parameters correlated with observed parameters					
Naïve comparison	11.3% (<0.01)	11.3% (<0.01)	0.03	0%	—
$MAIC_{MM}$	−0.2% (<0.01)	2.6% (<0.01)	0.03	96%	2%
$MAIC_{HM}$	−0.2% (<0.01)	2.6% (<0.01)	0.03	96%	2%
PSW	0% (<0.01)	2.6% (<0.01)	0.03	96%	2%
Missing parameters uncorrelated with observed parameters					
Naïve comparison	12.6% (<0.01)	12.6% (<0.01)	0.03	0%	—
$MAIC_{MM}$	4.3% (<0.01)	4.6% (<0.01)	0.03	75%	0%
$MAIC_{HM}$	4.3% (<0.01)	4.6% (<0.01)	0.03	75%	0%
PSW	4.4% (<0.01)	4.7% (<0.01)	0.03	74%	0%
Non-normal distributions sampled in population A					
Naïve comparison	14.6% (<0.01)	14.6% (<0.01)	0.03	0%	—
$MAIC_{MM}$	1% (<0.01)	6.4% (<0.01)	0.03	63%	12%
$MAIC_{HM}$	17.6% (<0.01)	17.7% (<0.01)	0.03	1%	99%
PSW	−3.7% (<0.01)	3.9% (<0.01)	0.03	73%	1%
Non-normal distributions sampled in population B					
Naïve comparison	9% (<0.01)	9% (<0.01)	0.03	6%	—
$MAIC_{MM}$	−2.9% (<0.01)	3.5% (<0.01)	0.03	88%	12%
$MAIC_{HM}$	−2.9% (<0.01)	3.5% (<0.01)	0.03	88%	12%
PSW	5.9% (<0.01)	5.9% (<0.01)	0.03	38%	0%
Trimmed patient characteristics in population A (no poor performers)					
Naïve comparison	18% (<0.01)	18% (<0.01)	0.03	0%	—
$MAIC_{MM}$	−1.8% (<0.01)	9.3% (<0.01)	0.11	92%	14%
$MAIC_{HM}$	−7.6% (<0.01)	17.8% (<0.01)	0.18	90%	38%
PSW	6.8% (<0.01)	7% (<0.01)	0.04	60%	0%
Trimmed patient characteristics in population B (no good performers)					
Naïve comparison	18.3% (<0.01)	18.3% (<0.01)	0.03	0%	—
$MAIC_{MM}$	−0.1% (<0.01)	4.3% (<0.01)	0.05	95%	0%
$MAIC_{HM}$	−0.3% (<0.01)	4.3% (<0.01)	0.05	95%	0%
PSW	−5.5% (<0.01)	9.3% (<0.01)	0.09	87%	13%

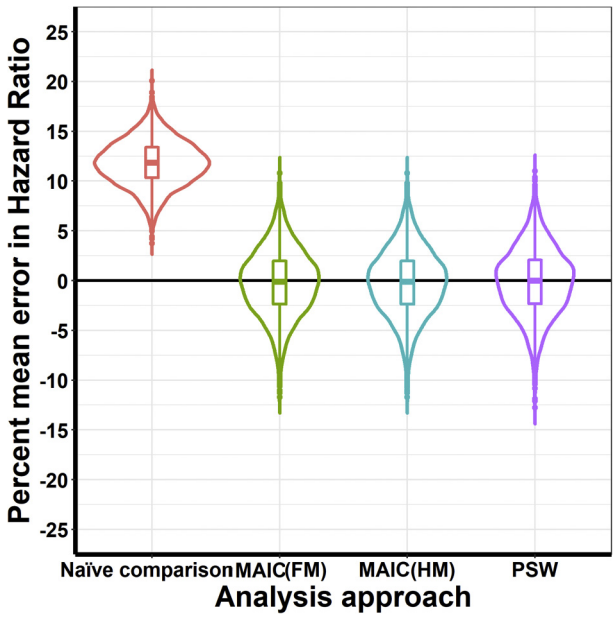
MCSE indicates Monte Carlo standard error; MAIC, matching adjusted indirect comparison; MM, method of moments; HM, includes higher moments; PSW, propensity score weighting.

of the hazard ratio, favoring the intervention; in this case, rather than the "true" value of 0.75, it is estimated to be 0.70.

With more than 5000 simulations, the results of the base-case analysis (Table 2; Figure 3) indicate MAIC (both $MAIC_{FM}$ and $MAIC_{HM}$) to be unbiased, as shown by the mean error being centered around zero, and accurate (absolute percentage error of 3% in estimating the true hazard ratio [HR]). In the vast majority (90%+) of scenarios, the 95% confidence interval contained the true HR, and in only 2% of scenarios was the error greater than in a naïve comparison. Indeed, in the base case, both forms of MAIC performed similarly to propensity weighting. These results are shown graphically in Figure 3 using a violin plot; this presents the density of the percentage error, with a bar chart overlaid to show the quartiles of the error distribution for each method.

These findings held when the setup of the simulation study was changed (Tables 1 and 2). The only areas of concern identified were those in which either the explanatory variable power was low or the treatment effect large (with reweighting then introducing bias in all forms). Of note are the mean absolute error, coverage probabilities, and chance of estimates being worse than a naïve comparison. Again, a similar pattern was seen with $MAIC_{FM}$ performing nearly as well as PSW in terms of mean percentage error and coverage probability, although $MAIC_{HM}$ performed slightly less well than the other 2 methods in having lower

Figure 3. Violin plot of the base-case result showing the density of the percentage mean error in the hazard ratio and the quartiles of error.



coverage probabilities, and more often provided estimates with a higher level of error than a naïve comparison.

Sensitivity analysis introducing complexities to the outcome model caused the performance of all matching methods to worsen but remain broadly adequate. The main concerns identified were the inclusion of variables not linked to outcomes in the matching (which would reduce the precision of estimates) or characteristics that were already well matched between studies. Although the same pattern in performance generally remains ($MAIC_{FM}$ matching or outperforming $MAIC_{HM}$, with both being outperformed by PSW).

Where the assumptions underpinning matching methods were violated, performance was considerably worse (as may be expected). Where variables were not included in the matching but linked to outcomes (and not correlated with other characteristics), there was an increase in both mean error and absolute error in the estimation of the treatment effect. Indeed, this is the only scenario in which the mean error is nonzero for $MAIC_{FM}$, demonstrating that should important variables be omitted that are more prevalent in one population, bias will not be adjusted for appropriately. Where the data are correlated, this bias is mitigated, although the mean absolute error remains higher than in many other scenarios.

Although MAIC as a method was broadly comparable with PSW, it did perform notably less well if the patient data available to use for reweighting (population A) used a different distribution than the historical control, either through a different distribution or with trimmed characteristics limiting the overlap, with both forms of MAIC exhibiting much increased levels of mean absolute error (indicating inaccuracy). In particular, $MAIC_{HM}$ in such instances performed exceptionally poorly (on mean error, mean absolute error, and coverage probability) and frequently exacerbated bias compared with a naïve comparison (Table 2).

The final set of analyses relates to the numbers of patients available in population A and population B and is shown in Figure 4. In analyses with low patient numbers (n = 30) either for matching or in the control, although the matching methods appear unbiased (shown by the median error being about zero), they are highly imprecise because of the low patient numbers (tabulated results are available in Supplementary Table A1 found at https://doi.org/10.1016/j.jval.2020.02.008). As the number of patients increases, the precision of methods improves, although a clear pattern emerges (comparing the north east versus south west off-diagonals in the figure) in which for MAIC, it appears more important to have more patients to use for reweighting (ie, the individual patient data) than greater precision in the moments to be matched (ie, the aggregate data).

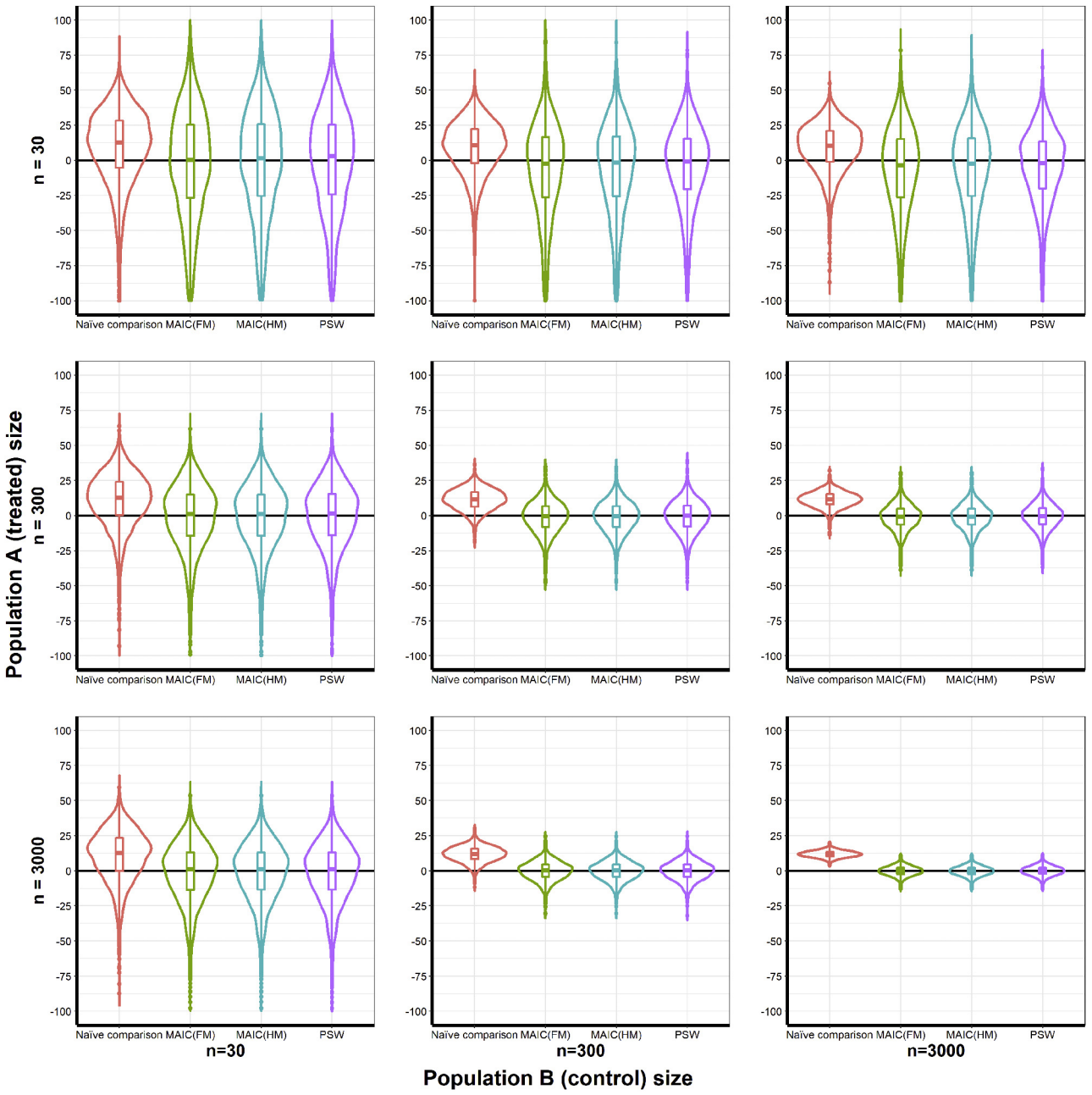
Discussion

Under ideal conditions in which the method is indicated, MAIC appears to be a valid and well-performing method for addressing bias in cross-study comparisons. This finding, however, does not remain constant where certain assumptions are not met, for instance, if important uncorrelated (and imbalanced) variables are omitted from the matching, the sample size is too low, or the variables matched on have only a limited impact on outcomes. Although noting the limitations of the approach, the performance is broadly comparable with those produced by the more established method of PSW (which requires access to the patient-level data from the historical trial). Most reassuringly is that under normal conditions, MAIC rarely exacerbates bias as compared with a naïve comparison. It is likely, however, that because of the study design (ie, data simulated from normal distributions), this is likely to flatter MAIC relative to PSW. In more complex examples including confounding by indication, the additional data available to PSW is likely to lead to improved estimates. Similarly, PSW would not be applied blindly, with data being able to be trimmed to match as necessary, further improving estimates.

Although MAIC matching on the first moments of the patient characteristics ($MAIC_{FM}$) appeared to work well on all endpoints, the same cannot be said for matching on higher moments ($MAIC_{HM}$). Although in many scenarios it performed similarly to $MAIC_{FM}$, in no scenarios did it provide a meaningful advantage, while also showing the potential for large errors (many of which are likely to be seen in practice, for instance, non-normally distributed data). Because of the lack of clear advantage and clear potential for harm based on the results of this study, it is not possible to recommend the use of $MAIC_{HM}$ as standard: careful justification should be given if it is to be used beyond sensitivity analyses. If $MAIC_{HM}$ is to be used, we would also note higher moments of binary variables should not be matched on. As highlighted by a reviewer, "once the mean is matched the variances would also be matched"—a point we had also overlooked. In this scenario, the poor performance of $MAIC_{HM}$ is due to our own effective misspecification of the model.

Although MAIC appears to function well as an approach based on the lack of bias and improved accuracy as compared with a naïve comparison, there are conditions highlighted by this study that should be met in order for MAIC to be used appropriately and circumstances in which we would caution against a reliance on MAIC-derived analysis. In addition to the need for a sufficient sample size, we suggest that there be good overlap between the studies included; explicit assessment of such overlap would therefore seem appropriate where MAIC is to be used. Similarly, the demonstration (where possible) of the link between matched characteristics and outcomes should be performed, for instance, in a third data set and using clinician input. We would also use caution with MAIC where there does not appear to be a large difference bias between studies, either because patient characteristics do not influence outcomes or because the difference between trials is small. Similarly, where an intervention

Figure 4. Violin plots of the percentage mean error in the hazard ratio when changing the number of patients available in population A and population B.



effect is large, MAIC may not be required, for instance, where there is a dramatic improvement in function following the delivery of an intervention.[20] In such instances, matching methods appear to have a substantial chance of overcorrection. The same criteria may be considered appropriate for propensity score analyses, although we acknowledge that the additional data available in propensity score–based analysis allows data to be analyzed to avoid such issues, for instance, aligning inclusion and exclusion criteria on data sets.

The high coverage probability (included as is the convention in simulation studies[12]) demonstrates that in most cases, the 95% confidence interval around the estimated outcome for MAIC does include the true value. It should be noted, however, that the use of MAIC results in a lower effective sample size and thus greater uncertainty in the resulting 95% interval (seen with the larger standard errors in the study). For this reason, we have focused on the more informative mean error and mean absolute error when interpreting results.

We believe that although the study presented here is comprehensive in the areas investigated, further studies are required. In particular, we highlight that we have conducted our analysis on simulated data. Although we are able to establish where the limitations of the methods lie, further work (including simulation studies and data analysis) is needed on how many parameters can feasibly be matched with different sample sizes, given the distribution of data seen in the real world. Similarly, understanding which variables should be included in matching

appears important; for instance, with several candidate variables linked to outcomes, at which point should the link to outcomes be considered too weak to include in matching? Likewise, how characteristics are included is a point for future research: should age be used as a continuous variable or in a grouping? There are numerous commonly used variables (particularly laboratory-measured values) to which this question applies. Until such information is known, the provision of sensitivity analyses with alternative model specifications seems prudent.

In addition to the need for further research, we would also highlight that this study compared 2 approaches (MAIC and PSW), but other approaches are available and could be considered suitable. In particular, we highlight simulated treatment comparison (STC),[21] in which access is not available to the individual patient data from both trials. STC is a regression-based method and requires that an outcome model be constructed. It is thus subject to different assumptions, such as the role of missing data and the need to specify an approach to model construction. Although STC and MAIC have yet to be compared, STC is able to overcome one of the key limitations of MAIC: that the population of interest may not be the one in the historical control but rather may be that of population A or indeed have different characteristics altogether. For this reason, further work comparing MAIC and STC in real-world problems would therefore be advantageous.

Although an imperfect tool, MAIC appears to be a useful method for the estimation of comparative efficacy. Although not without disadvantages, it performs similarly to PSW under most scenarios, but in the real-world, PSW would not be an available comparative method (as individual patient data may not be available for 2 studies). Provided careful consideration is given to the circumstances in which it is used, MAIC has the potential to provide accurate estimates of relative efficacy. We would, however, urge analysts to carefully examine the assumptions inherent in the approach to determine its suitability for a given problem.

Acknowledgments

We would like to thank the peer reviewers of the article, as their comments were extremely insightful and led to a revision of methods and change in how results are reported. Their input substantially improved our work, for which we are grateful.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2020.02.008.

REFERENCES

1. 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;50:683–691.
2. Jansen JP. Network meta-analysis of survival data with fractional polynomials. *BMC Med Res Methodol*. 2011;11:61.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41.
4. Signorovitch JE, Wu EQ, Andrew PY, et al. Comparative effectiveness without head-to-head trials. *Pharmacoeconomics*. 2010;28:935–945.
5. Phillippo D, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. *NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE*; 2016.
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*. 2017;38(2):200–211.
7. Kühnast S, Schiffner-Rohe J, Rahnenführer J, Leverkus F. Evaluation of adjusted and unadjusted indirect comparison methods in benefit assessment: A simulation study for time-to-event endpoints. *Methods Inf Med*. 2017;56:261–267.
8. 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;22:85–91.
9. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chron Dis*. 1976;29:175–188.
10. Moroz V, Wilson JS, Kearns P, Wheatley K. Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in favour of the novel treatment. *Trials*. 2014;15:481.
11. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25:4279–4292.
12. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38:2074–2102.
13. Bellmunt J, Choueiri TK, Fougeray R, et al. Prognostic factors in patients with advanced transitional cell carcinoma of the urothelial tract experiencing treatment failure with platinum-containing regimens. *J Clin Oncol*. 2010;28:1850–1855.
14. Phillippo DM, Dias S, Elsada A, Ades AE, 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.
15. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9.
16. Gabrio A, Mason AJ, Baio G. A full Bayesian model to handle structural ones and missingness in economic evaluations from individual-level data: handling structural ones and missingness in economic evaluations. *Stat Med*. 2019;38:1399–1420.
17. Leurent B, Gomes M, Faria R, Morris S, Grieve R, Carpenter JR. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. *Pharmacoeconomics*. 2018;36:889–901.
18. 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;33:743–754.
19. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017.
20. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ*. 2007;334:349–351.
21. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics*. 2010;28:957–967.