Title:

The effects of model misspecification in unanchored Matching Adjusted Indirect Comparison (MAIC); Results of a simulation study

Running title:

Model misspecification in unanchored MAIC

Article information

Number of text pages	14 (including title page, text, references, legends)
Number of tables	2
Number of figures	4
Words	4016
Online Appendix	1 Table, 3 Figures

Data availability

All data used in this study are simulated, with the distributions they are drawn from described in appendices.

Authors

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

1. Department of Statistical Science, University College London, London, UK, WC1E 6BT
2. Delta Hat, 212 Tamworth Road, Nottingham, UK, NG10 3GS
3. Department of Primary Care & Population Health, University College London, London, UK

Corresponding author

Anthony J Hatswell
Delta Hat
212 Tamworth Road
Nottingham
NG10 3GS
+44 (0)7759 243157
ahatswell@deltahat.co.uk

Conflicts of interest

No funding was received for this work, and the authors declare no conflicts.

Acknowledgements

We would like to thank the peer reviewers of the paper; 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.

Abstract

Objectives

To assess the performance of MAIC matching on first moments or higher moments in unanchored comparison 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 where patients had more favourable 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] around 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 i.e. higher moments. Relative to PSW, MAIC gave similar results in the majority of circumstances, though 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 to PSW.

Key words

MAIC; Signorovitch weighting; Propensity score; single arm trial; historical control

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 (Bucher et al., 1997), or network meta-analysis (NMA) (Jansen, 2011) can provide estimates of the relative efficacy of interventions. However where trials either have substantially different comparator arms, no available link to connect them (i.e. 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, whilst propensity score techniques (i.e. propensity score matching or propensity score weighting [PSW]) require access to patient level data (Rosenbaum and Rubin, 1983), which are frequently not available for at least one 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 reasons, or the data being inaccessible due to the passage of time. Where this patient level data are not available, Matching Adjusted Indirect Comparison (MAIC) has been proposed (Signorovitch et al., 2010). 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 two approaches – firstly to account for differences between trials in predictive characteristics, with the results subsequently used to inform NMA (this is often termed an 'anchored comparison'). 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 minimise bias in estimates of comparative efficacy – for example accounting for one population being younger or having better performance status.

While 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 NICE Decision Support Unit Technical Support Document report and associated publication (Phillippo et al., 2016, 2017), which defines terminology, reviews the theoretical validity of the method, and suggests best practice (i.e. in unanchored MAIC include all prognostic and predictive characteristics). This work however does not give 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; whilst two simulation studies simulation studies have been published, these focus on the use of MAIC in 'anchored' indirect comparisons, as opposed to the unanchored form (Kühnast et al., 2017; Petto et al., 2019).

As the majority 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 (i.e. matching on the mean and variance; $MAIC_{HM}$), and PSW. PSW is included as 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 & design

Our review of published MAICs found that the majority of published applications are in oncology (23 applications), compared to 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, comparing an intervention to a 'historical control' (Pocock, 1976). 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 favourable characteristics than the patients who received the historical control (termed Population B), leading to a bias in favour of the intervention (Moroz et al., 2014). 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 (Burton et al., 2006; Morris et al., 2019).

Data generating mechanism

In the study, six patient characteristics ($X_1, \ldots, X_6$) were simulated; four assumed to be fully observed and available for matching ($X_1, \ldots, X_4$), whilst two (X_5, X_6) were assumed to be unobserved. In the base case, these were all assumed to be uncorrelated. The four 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 favour 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 three prognostic characteristics (Bellmunt et al., 2010), and in line with previous MAICs where Phillippo et al. (2019) found a median of six characteristics were adjusted for (range 1-13). Each characteristic was then multiplied by an effect size for that characteristic ($\beta_1, \ldots, \beta_6$). The sum of these products were added to a constant (intercept) and then used as a linear predictor (LP) 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 $LP = \sum_{i=1}^{6} \beta_i \mathrm{X}_i + c$ with survival outcomes for each patient then sampled from the corresponding distribution $Weibull(\alpha = 1.3, \exp(LP))$. In each run of the simulation study patient characteristics ($\mathrm{X}_1, \ldots, \mathrm{X}_6$) were resampled, as were different effect sizes ($\beta_1, \ldots, \beta_6, \alpha$) 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 only testing a specific 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 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 approach to extrapolation and/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, though methods do exist for digitisation of survival outcomes (Guyot et al., 2012), estimation of missing data (Gabrio et al., 2019; Leurent et al., 2018), and extrapolation of survival times (Latimer, 2013) 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 favourable characteristics in Population A. Matching methods (both MAIC and PSW), attempt therefore 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 which would have been obtained had a controlled study of A versus B been performed in Population B. Due to being 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 three approaches; firstly using MAIC matching on the means (first moments) of Population B, $MAIC_{FM}$, matching was then conducted using also the standard deviation of the summarised data from Population B i.e. matching also on higher moments, $MAIC_{HM}$ – this approach was also proposed in the original MAIC paper by Signorovitch et al.(Signorovitch et al., 2010)*: "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, where 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) were 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, three 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; over and under predictions are both penalised 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 remove it.

Scenario analyses

Scenario analysis were then conducted with three broad aims – varying the characteristics of the simulation study, testing the limits of where MAIC can be applied, and then 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 i.e. variables that were not linked to outcomes, or variables were linked in a non-linear fashion. Further scenarios considered the degree of overlap between the two 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 ($X_1, \ldots, X_4$). The effect of outcome distributions of ($X_1, \ldots, X_6$) was then explored deviating from the initial assumption of normality (the lognormal was used), with also trimmed distributions of X used (mimicking trials which 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 were 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 (R Core Team, 2017), 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 modelled survival for one iteration of the simulation study, with a naïve comparison comparing the Population B historical control data (blue line, with median survival of 9.2 months, mean survival of 11.4 months over all simulations) to 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 i.e. not had more favourable 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 (a median bias of 1.1 months, median bias of 1.6 months) due to more favourable patient characteristics leads to an underestimate of the hazard ratio, favouring the intervention; in the case rather than the 'true' value of 0.75, it is estimated to be 0.70.

Over 5,000 simulations, the results of the base case analysis (Table 2, Figure 3) indicate MAIC (both ($MAIC_{FM}$ and $MAIC_{HM}$) to be unbiased - shown by the mean error being centred around zero, and accurate (absolute percentage error of 3% in estimating the true 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 form 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 (Table 1, Table 2). The only areas of concern identified were those where either the explanatory variable power was low or the treatment effect large (with reweighting then introducing bias in all forms) – of note are the MAE, 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, though $MAIC_{HM}$ performed slightly less well than the other two methods in having lower coverage probabilities, and more often giving 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 if characteristics are already well matched between studies. Whilst 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 are not included in the matching but linked to outcomes (and not correlated with other characteristics), there is an increase in both mean error and absolute error in the estimation of the treatment effect. Indeed, this is the only scenario where the mean error is non-zero 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 MAE remains higher than in many other scenarios.

Whilst MAIC as a method was broadly comparable to PSW, it did perform notably less well if the patient data available to use for reweighting (Population A) used a different distribution to the historical control – either through a different distribution, or 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 both mean error, mean absolute error and coverage probability), and frequently exacerbated bias compared to a naïve comparison (Table 2).

The final set of analyses relate 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 matching methods appear unbiased (shown by the median error being around zero), they are highly imprecise due to the low patient numbers (tabulated results are available in Supplementary Table A1). As the number of patients increases the precision of methods improves, though a clear pattern emerges (comparing the North East versus South West off-diagonals in the figure) that for MAIC it appears more important to have more patients to use for reweighting (i.e. the individual patient data), than greater precision on the moments to be matched (i.e. the aggregate data).

Discussion

Under ideal conditions where the method is indicated, MAIC appears to be a valid and well performing method to address 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, sample size is too low, or the variables matched on have only a limited impact on outcomes. Whilst noting the limitations of the approach, the performance is broadly comparable to 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 compared to a naïve comparison. It is likely however that due to the study design i.e. 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.

Whilst 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 ($MAIC_{HM}$). Whilst in many scenarios it performed similarly to $MAIC_{FM}$, in no scenarios did it provide a meaningful advantage, whilst also showing the potential for large errors (many of which are likely to be seen in practice – for instance non-normally distributed data). Due to 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 compared to a naïve comparison, there are conditions highlighted by this study that should be met in order for MAIC to be used appropriately, and circumstances where we would caution against a reliance on MAIC-derived analysis. In addition to the need for 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 dataset 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 effect is large, MAIC may not be required – for instance where there is a dramatic improvement in function following the delivery of an intervention(Glasziou et al., 2007) -in such instances matching methods appear to have a substantial chance of overcorrection. The same criteria may be considered appropriate for propensity score analyses, though we acknowledge that the additional data available in propensity score based analysis allows data to be analysed to avoid such issues; for instance aligning inclusion and exclusion criteria on datasets.

The high coverage probability (included as is convention in simulation studies (Morris et al., 2019)) demonstrates that in the majority of 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); how useful this is therefore relative to the point estimate of the effect. For this reason we have focussed interpretation 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. Whilst 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? Similarly 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 two approaches (MAIC and PSW), however others approaches are available and could be considered suitable. In particular we would highlight Simulated Treatment Comparison (STC) (Caro and Ishak, 2010) where access is not available to the individual patient data from both trials. STC is a regression based method and requires an outcome model be constructed and thus is subject to different assumptions such as the role of missing data and the need to specify an approach to model construction. Whilst 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.

Whilst 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 the majority of scenarios even though in the real world PSW would not be an available comparative method (as individual patient data may not be available for two studies). Provided careful consideration is given to the circumstances in which it is used, MAIC has the potential to provide accurate and 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.

Bellmunt, J., Choueiri, T.K., Fougeray, R., Schutz, F.A.B., Salhi, Y., Winquist, E., Culine, S., von der Maase, H., Vaughn, D.J., Rosenberg, J.E., 2010. Prognostic Factors in Patients With Advanced Transitional Cell Carcinoma of the Urothelial Tract Experiencing Treatment Failure With Platinum-Containing Regimens. Journal of Clinical Oncology 28, 1850–1855. https://doi.org/10.1200/JCO.2009.25.4599

Bucher, H.C., Guyatt, G.H., Griffith, L.E., Walter, S.D., 1997. The Results of Direct and Indirect Treatment Comparisons in Meta-Analysis of Randomized Controlled Trial. J Clin Epidemiol 50, 683–691.

Burton, A., Altman, D.G., Royston, P., Holder, R.L., 2006. The design of simulation studies in medical statistics. Statist. Med. 25, 4279–4292. https://doi.org/10.1002/sim.2673

Caro, J.J., Ishak, K.J., 2010. No head-to-head trial? Simulate the missing arms. Pharmacoeconomics 28, 957–967.

Gabrio, A., Mason, A.J., Baio, G., 2019. 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. Statistics in Medicine 38, 1399–1420. https://doi.org/10.1002/sim.8045

Glasziou, P., Chalmers, I., Rawlins, M., McCulloch, P., 2007. When Are Randomised Trials Unnecessary? Picking Signal from Noise. BMJ: British Medical Journal 334, 349–351.

Guyot, P., Ades, A., Ouwens, M.J., Welton, N.J., 2012. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Medical Research Methodology 12, 9. https://doi.org/10.1186/1471-2288-12-9

Jansen, J.P., 2011. Network meta-analysis of survival data with fractional polynomials. BMC Medical Research Methodology 11, 61. https://doi.org/10.1186/1471-2288-11-61

Kühnast, S., Schiffner-Rohe, J., Rahnenführer, J., Leverkus, F., 2017. Evaluation of Adjusted and Unadjusted Indirect Comparison Methods in Benefit Assessment. A Simulation Study for Time-to-event Endpoints. Methods Inf Med 56, 261–267. https://doi.org/10.3414/ME15-02-0016

Latimer, N.R., 2013. Survival Analysis for Economic Evaluations Alongside Clinical Trials—Extrapolation with Patient-Level Data: Inconsistencies, Limitations, and a Practical Guide. Medical Decision Making 33, 743–754. https://doi.org/10.1177/0272989X12472398

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., Carpenter, J.R., 2018. Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial. PharmacoEconomics 36, 889–901. https://doi.org/10.1007/s40273-018-0650-5

Moroz, V., Wilson, J.S., Kearns, P., Wheatley, K., 2014. 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 15, 481.

Morris, T.P., White, I.R., Crowther, M.J., 2019. Using simulation studies to evaluate statistical methods. Statistics in Medicine 0. https://doi.org/10.1002/sim.8086

Petto, H., Kadziola, Z., Brnabic, A., Saure, D., Belger, M., 2019. Alternative Weighting Approaches for Anchored Matching-Adjusted Indirect Comparisons via a Common Comparator. Value in Health 22, 85–91. https://doi.org/10.1016/j.jval.2018.06.018

Phillippo, D., Ades, A.E., Dias, S., Palmer, S., Abrams, K.R., Welton, N.J., 2016. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE.

Phillippo, D.M., Ades, A.E., Dias, S., Palmer, S., Abrams, K.R., Welton, N.J., 2017. Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. Med Decis Making 0272989X17725740. https://doi.org/10.1177/0272989X17725740

Phillippo, D.M., Dias, S., Elsada, A., Ades, A.E., Welton, N.J., 2019. Population Adjustment Methods for Indirect Comparisons: A Review of National Institute for Health and Care Excellence Technology Appraisals. International Journal of Technology Assessment in Health Care 35, 221–228. https://doi.org/10.1017/S0266462319000333

Pocock, S.J., 1976. The combination of randomized and historical controls in clinical trials. Journal of chronic diseases 29, 175–188.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rosenbaum, P.R., Rubin, D.B., 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70, 41. https://doi.org/10.2307/2335942
Signorovitch, J.E., Wu, E.Q., Andrew, P.Y., Gerrits, C.M., Kantor, E., Bao, Y., Gupta, S.R., Mulani, P.M., 2010. Comparative Effectiveness Without Head-to-Head Trials. Pharmacoeconomics 28, 935–945.

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 = $\exp(2+\sum X\beta)/$ TreatmentHR))	Survival: $Y \sim$ Weibull(shape = 1, scale = $\exp(2+\sum X\beta)$)
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 i.e. Population A less favourable	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 non-linear	Survival: $Y \sim$ Weibull(shape = 1.3, scale = $\exp(2+\sum X\beta)/$ TreatmentHR))	Survival: $Y \sim$ Weibull(shape = 1.3, scale = $\exp(2+\sum \exp(X\beta))/$ 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

<table>
<tr><td colspan="3">Violating assumptions implicit or explicit in MAIC</td></tr>
<tr><td>Missing parameters correlated with observed parameters</td><td rowspan="2">Covariates 5 & 6:
$X \sim N(0, 0.2)$</td><td>Covariates 5 & 6:
$X \sim$ mean of parameters 1:4 + $N(0, 0.1)$</td></tr>
<tr><td>Missing parameters uncorrelated with observed parameters</td><td>Covariates 5 & 6:
Population A: $X_A \sim N(0.3, 0.1)$
Population B: $X_B \sim N(0.25, 0.1)$
Covariates 1:6: $\beta \sim N(0.35, 0.15)$</td></tr>
<tr><td>Non-normal distributions sampled in Population A</td><td rowspan="4">Covariates 1 to 4:
Population A: $X_A \sim N(0.3, 0.1)$
Population B: $X_B \sim N(0.25, 0.1)$</td><td>Covariates 1 to 4:
Population A: $X_A \sim$ Lognormal(SDlog = 0.5, meanlog = log(0.27))</td></tr>
<tr><td>Non-normal distributions sampled in Population B</td><td>Covariates 1 to 4:
Population A: $X_B \sim$ Lognormal(SDlog = 0.5, meanlog = log(0.22))</td></tr>
<tr><td>Trimmed patient characteristics in Population A (no poor performers)</td><td>Covariates:
Population A: $X_A \sim N(0.3, 0.1)$ truncated at min of 0.2</td></tr>
<tr><td>Trimmed patient characteristics in Population B (no good performers)</td><td>Covariates:
Population B: $X_B \sim N(0.25, 0.1)$ truncated at max of 0.35</td></tr>
<tr><td colspan="3">NB: Distribution parameterisations 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 specified using the shape (and not rate) parameter</td></tr>
</table>

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	Percent 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 i.e. Population A are worse by 0.5SD					
Naïve comparison	-13.6% (<0.01)	13.6% (<0.01)	0.03	0%	-
$MAIC_{MM}$	-0.2% (<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 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 non-linear					
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%
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 = Monte Carlo Standard Error, MAIC = Matching Adjusted Indirect Comparison, MM = Method of moments, HM = includes Higher moments, PSW = Propensity Score Weighting

Figure 1: Data generation mechanism for the simulation study

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

$\alpha = 1.3$

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

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

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

For the intervention:

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

$Treatment\ Hazard\ Ratio = 0.75$

For the control:

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

$Treatment\ Hazard\ Ratio = 1$

For intervention & control:

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

Figure 2: Example of simulated survival in the base case analysis showing the longer survival of Population A compared to Population B with treatment (due to more favourable patient characteristics), and the resulting bias in a naïve comparison to a historical Population B cohort

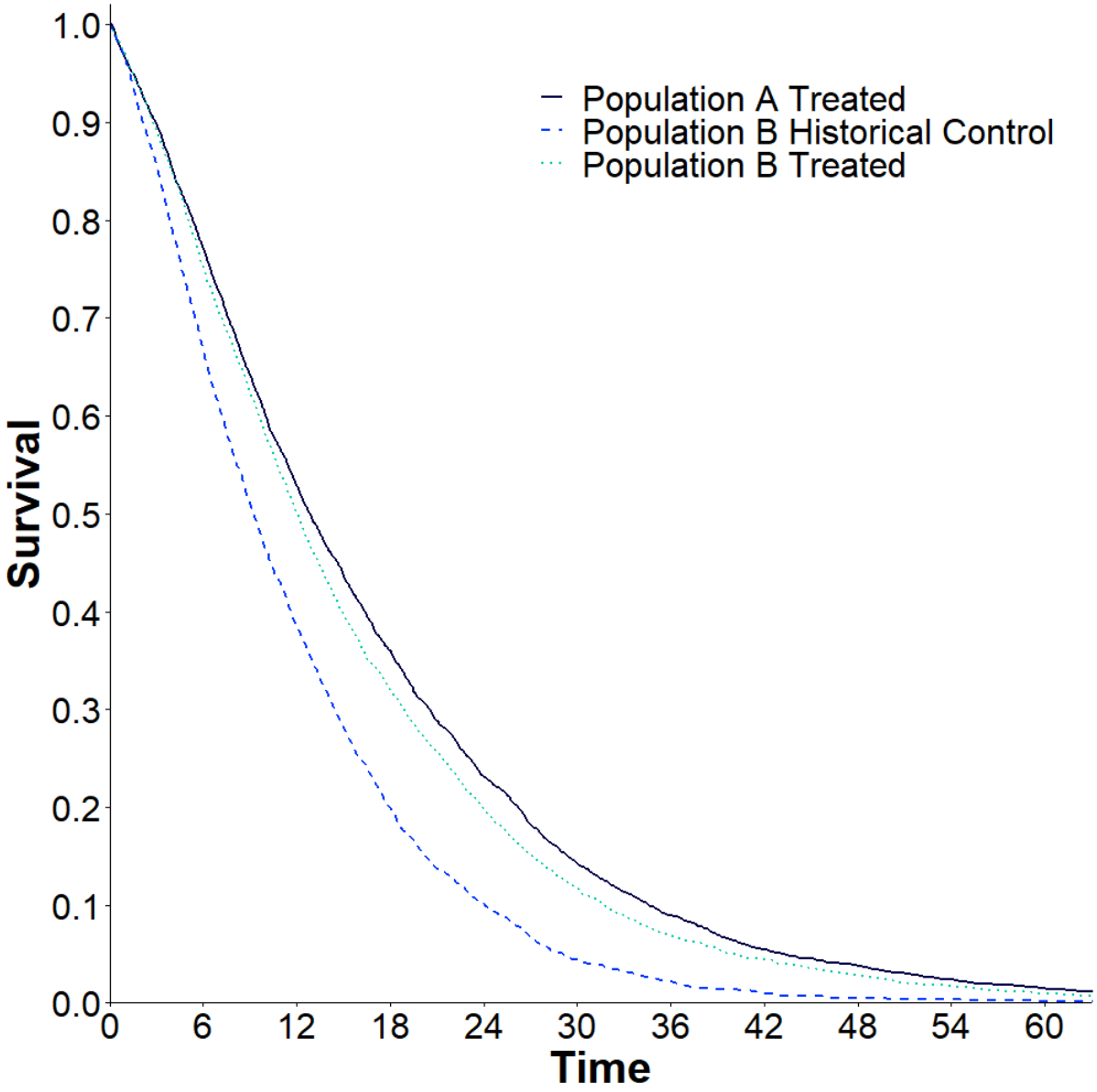


Figure 3: Violin plot of the base case result; showing the density of the percent mean error in the hazard ratio as well as the quartiles of error

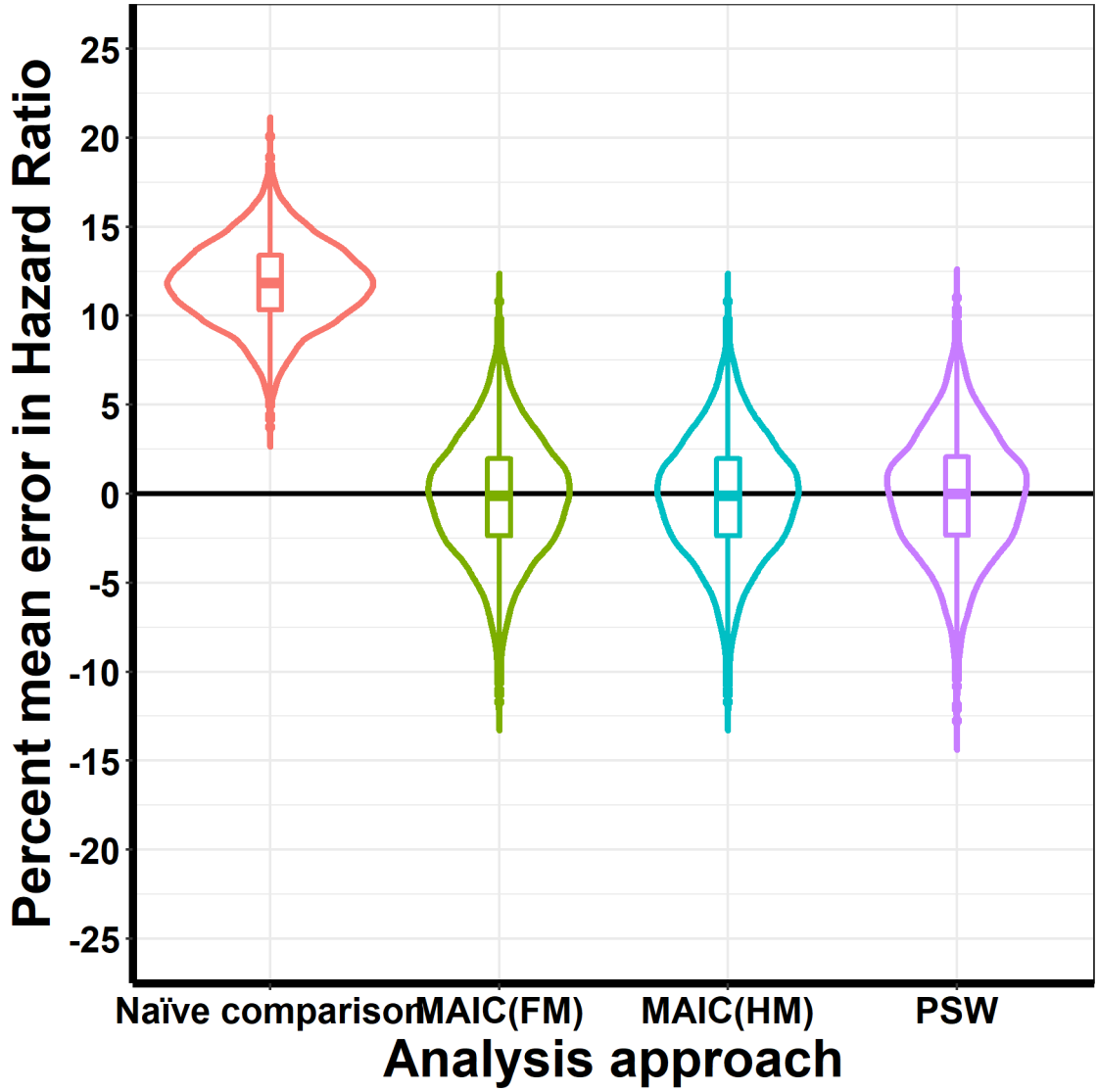
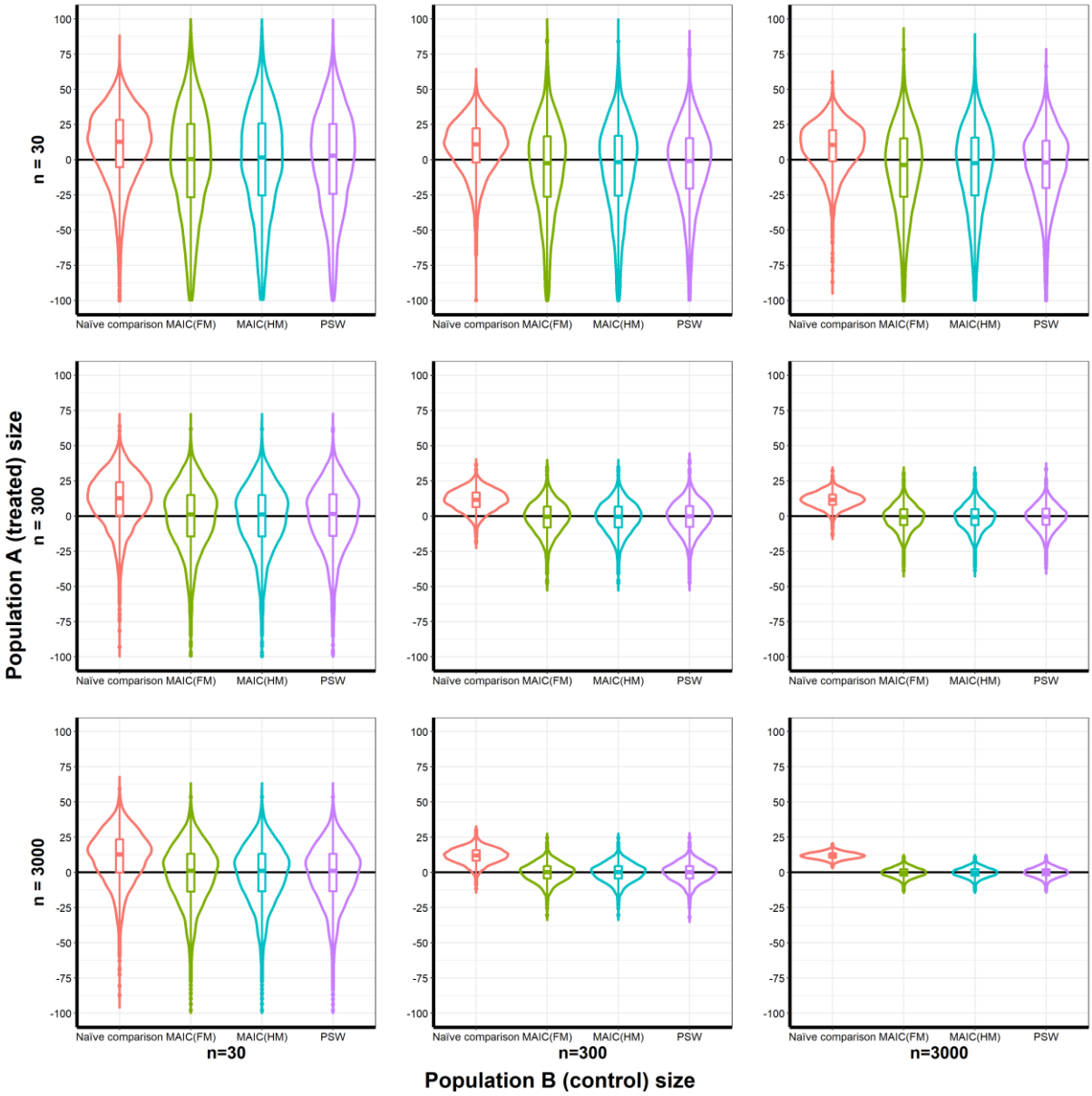


Figure 4: Violin plots of the percent mean error in hazard ratio when changing the number of patients available in Population A and Population B



Appendix

Table A1: Tabulated results of scenario analyses varying patient numbers

3

Method	Mean Percentage Error	Absolute Percentage Error	Mean Standard Error	Coverage probability	Percent of scenarios worse than a naïve comparison
Population A = 30, Population B = 30					
Naïve comparison	8.9% (<0.01)	23.1% (<0.01)	0.27	90%	-
$MAIC_{MM}$	-15.3% (0.012)	41.5% (0.011)	0.37	88%	62%
$MAIC_{HM}$	-12% (<0.01)	38.7% (<0.01)	0.36	88%	61%
PSW	-4.4% (<0.01)	31.4% (<0.01)	0.33	90%	60%
Population A = 30, Population B = 300					
Naïve comparison	8.8% (<0.01)	17.1% (<0.01)	0.19	90%	-
$MAIC_{MM}$	-12.5% (<0.01)	31.1% (<0.01)	0.28	86%	63%
$MAIC_{HM}$	-12% (<0.01)	30.8% (<0.01)	0.27	86%	62%
PSW	-6% (<0.01)	23.6% (<0.01)	0.24	89%	56%
Population A = 30, Population B = 3000					
Naïve comparison	8.9% (<0.01)	15.8% (<0.01)	0.18	89%	-
$MAIC_{MM}$	-11.9% (<0.01)	28.8% (<0.01)	0.25	86%	63%
$MAIC_{HM}$	-12.2% (<0.01)	29.4% (<0.01)	0.25	85%	62%
PSW	-6.1% (<0.01)	21.7% (<0.01)	0.22	90%	56%
Population A = 300, Population B = 30					
Naïve comparison	10.9% (<0.01)	18.1% (<0.01)	0.19	85%	-
$MAIC_{MM}$	-1.2% (<0.01)	18% (<0.01)	0.21	92%	41%
$MAIC_{HM}$	-1.2% (<0.01)	18.1% (<0.01)	0.21	92%	41%
PSW	-0.9% (<0.01)	18.1% (<0.01)	0.21	91%	41%
Population A = 300, Population B = 300					
Naïve comparison	11.4% (<0.01)	12% (<0.01)	0.08	68%	-

$MAIC_{MM}$	-1% (<0.01)	8.7% (<0.01)	0.11	95%	30%
$MAIC_{HM}$	-1% (<0.01)	8.7% (<0.01)	0.11	95%	30%
PSW	-0.7% (<0.01)	8.8% (<0.01)	0.11	94%	30%
Population A = 300, Population B = 3000					
Naïve comparison	11.4% (<0.01)	11.6% (<0.01)	0.06	47%	-
$MAIC_{MM}$	-0.9% (<0.01)	6.9% (<0.01)	0.09	94%	23%
$MAIC_{HM}$	-0.9% (<0.01)	6.9% (<0.01)	0.09	94%	23%
PSW	-0.7% (<0.01)	6.9% (<0.01)	0.09	94%	23%
Population A = 3000, Population B = 30					
Naïve comparison	10.7% (<0.01)	17.6% (<0.01)	0.18	83%	-
$MAIC_{MM}$	-1.4% (<0.01)	16.4% (<0.01)	0.18	90%	37%
$MAIC_{HM}$	-1.4% (<0.01)	16.4% (<0.01)	0.18	90%	37%
PSW	-1.4% (<0.01)	16.4% (<0.01)	0.18	90%	37%
Population A = 3000, Population B = 300					
Naïve comparison	11.8% (<0.01)	11.9% (<0.01)	0.06	45%	-
$MAIC_{MM}$	-0.1% (<0.01)	5.3% (<0.01)	0.07	94%	16%
$MAIC_{HM}$	-0.1% (<0.01)	5.3% (<0.01)	0.07	94%	16%
PSW	-0.1% (<0.01)	5.3% (<0.01)	0.07	94%	16%
Population A = 3000, Population B = 3000					
Naïve comparison	11.9% (<0.01)	11.9% (<0.01)	0.03	0%	-
$MAIC_{MM}$	-0.1% (<0.01)	2.7% (<0.01)	0.03	95%	2%
$MAIC_{HM}$	-0.1% (<0.01)	2.7% (<0.01)	0.03	95%	2%
PSW	0% (<0.01)	2.7% (<0.01)	0.03	95%	2%

Figure A1: Violin plots of mean error for scenario analyses changing the setup of the simulation study

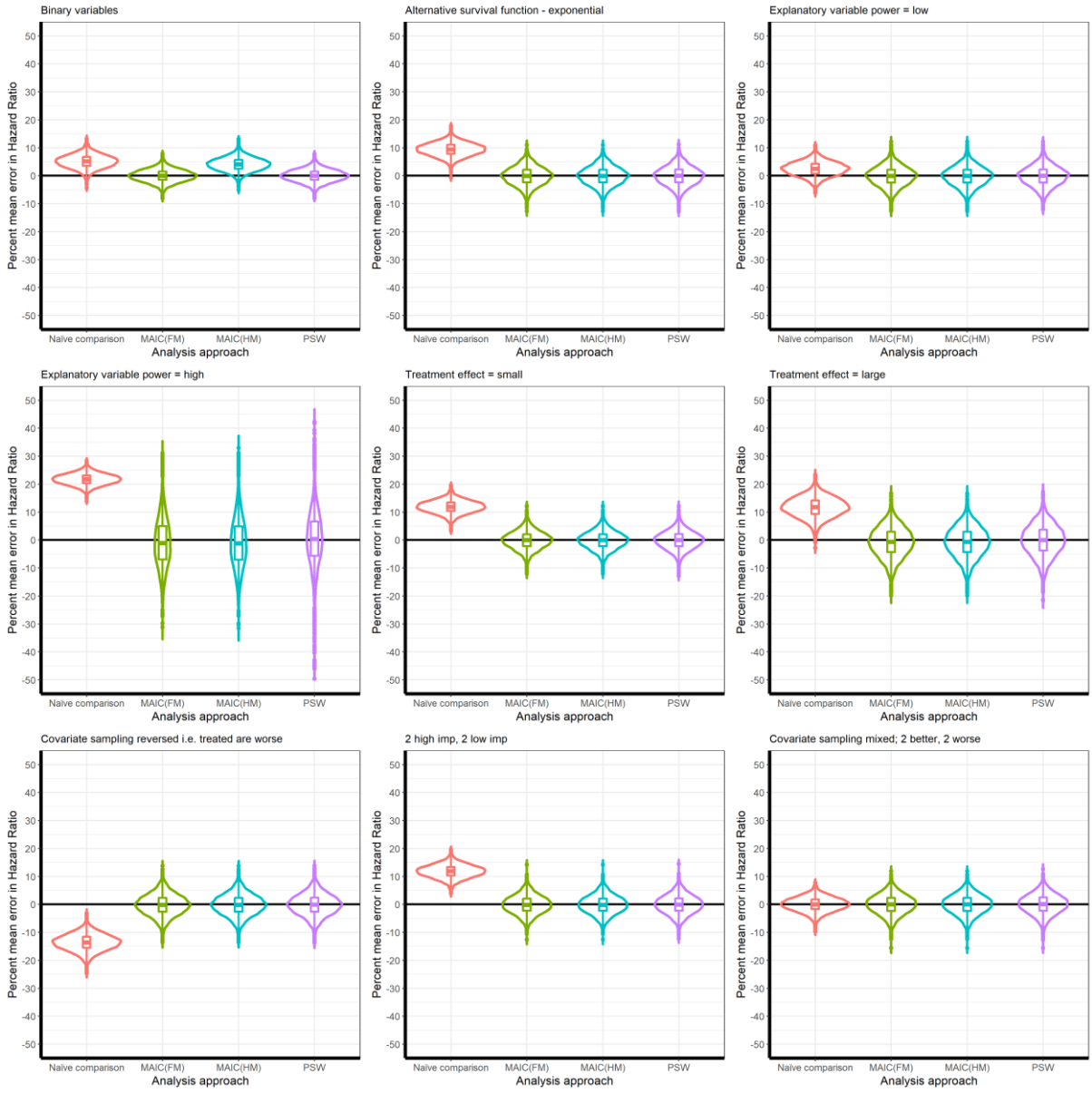


Figure A2: Violin plots of mean error for scenario analyses exploring the limits of unanchored Matching Adjusted Indirect Comparison

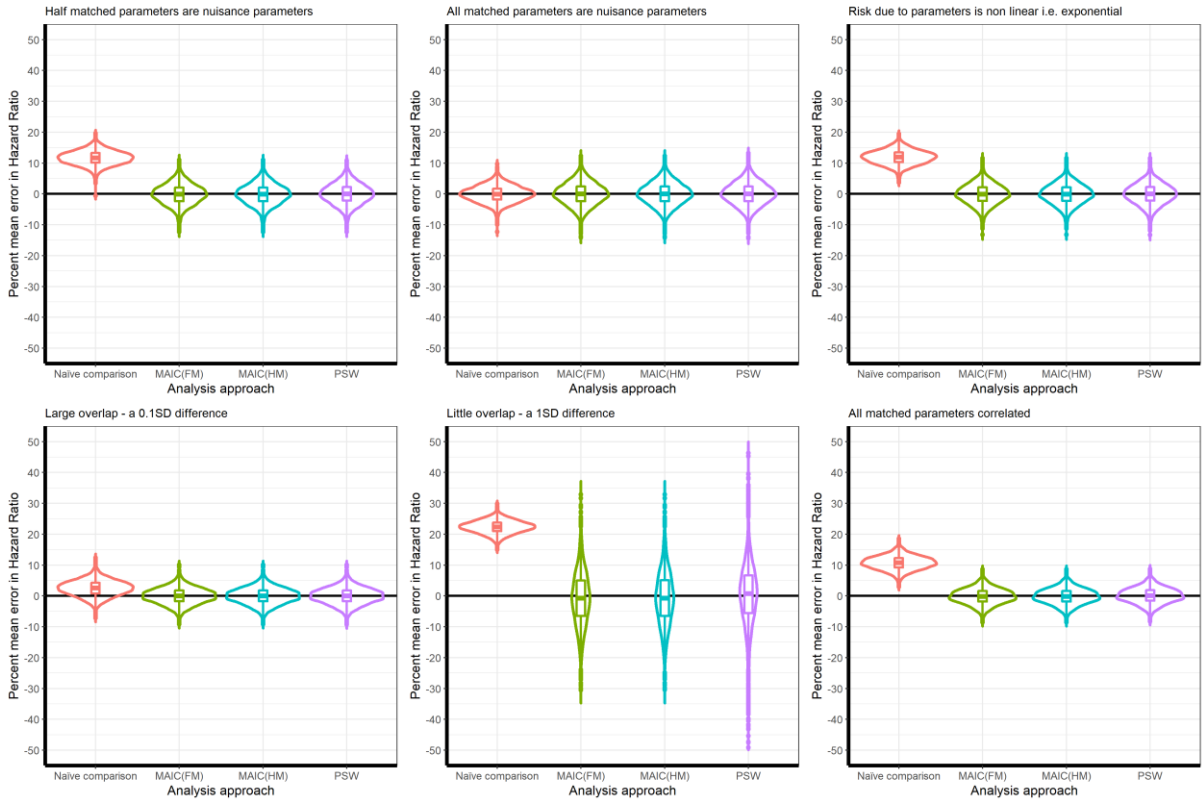


Figure A3: Violin plots of mean error for scenario analyses violating the assumptions underpinning unanchored Matching Adjusted Indirect Comparison